Factors Affecting the Utility of Emotional Stimuli in Research

K DICONNE

PhD 2022

Factors Affecting the Utility of Emotional Stimuli in Research

KATHRIN DICONNE

A thesis submitted in partial fulfilment of the requirements of Manchester Metropolitan University for the degree of Doctor of Philosophy

Department of Psychology

Manchester Metropolitan University

Faculty of Health, Psychology and Social Care

2022

To M. D. & W. P.

Intellectual Property and Publications

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

One publication has been produced from research that was undertaken as part of this thesis. This publication is listed below with full reference and details of its location within the thesis. In case of this publication where the candidate is the first author, the candidate was solely responsible for the production of the content with named authors providing support through review and modification.

Diconne, K., Kountouriotis, G. K., Paltoglou, A. E., Parker, A., & Hostler, T. J. (2022). Presenting KAPODI – The Searchable Database of Emotional Stimuli Sets. Emotion Review*,* 14*(1), 84–95.* https://doi.org/10.1177/17540739211072803

This paper is based on the data presented in *Chapter Two.* As the publication presents the created searchable KAPODI - database itself, all analyses conducted on the content of the database and referring to the state of the database at the time of writing were removed for the publication.

Abstract

Emotional stimuli such as images, words, or video clips are often used in studies researching emotion. These stimuli are provided to the research community in sets accompanied by normative rating data indicating the emotional value of each stimulus. With emotional stimuli sets continuously being published, an immense number of available sets are complicating the task for researchers looking for suitable stimuli. Therefore, a systematic review was conducted to find all existing emotional stimuli sets that are freely available or available upon request. The result was the creation of the KAPODI-database containing 364 sets and presenting a comprehensive list of set characteristics. A searchable online version allows researchers to find and compare individual sets, as well as to add new published sets. Previous research has shown that factors such as assessors' age, gender, or ethnicity, influence stimulus perception. Nevertheless, researchers often rely on the provided normative rating data without verifying its validity for their own participant sample. Additionally, findings regarding the effect of emotions onto memory are inconsistent, with sometimes enhancing, and sometimes detrimental effects. A possible reason for these contradictory results could be factors influencing stimulus validity that have yet not been investigated. Therefore, two additional studies were conducted. A first study sought to analyse these possible factors by investigating validity of stimuli in relation to assessed dimension, namely valence (negative to positive), arousal (calming to exciting), and dominance (no dominance to high dominance), as well as different dimension categories (e.g., low/medium/high valence, arousal, and dominance, respectively), and standard deviation (*SD*) categories (low/medium/high) both for images and words. In the second study, the factor of sensory processing sensitivity (SPS) that is known to positively correlate with the depth of processing of emotional content was investigated. In this latter experiment perception as well as episodic recognition of emotional image stimuli were assessed and analysed in relation to level of SPS. The two experimental studies suggest that solely valence seem reliable for both image and word stimuli, while arousal, dominance, dimension, and *SD* category are not reliable. Moreover, perception of emotional stimuli differs between individuals of *low* vs. *high* SPS regarding low-valence stimuli only, with *high* SPS individuals perceiving these stimuli as more negative. Finally, recognition of stimuli increased with increasing arousal, and decreased with increasing valence. Together, these results urge researchers to validate arousal and dominance ratings of selected stimuli for their participant sample prior to study conduction, as well as to consider the participants' sensitivity if the study uses negative (*low*-valenced) stimuli.

Acknowledgement

I would like to thank the following people who helped me throughout the completion of this present research project in the past four years:

First, I would like to thank my four supervisors Dr. Thomas Hostler, Dr. George Kountouriotis, Dr. Aspasia Paltoglou, as well as Dr. Andrew Parker, who provided me with excellent support throughout the entire time. I highly estimate their guidance, advice and feedback they have given me to complete this work. They have helped me to continue and improve my path within research.

Second, I would like to thank all those who have received me with open arms and hosted me during different time periods over the past years, especially during the final year of completion.

Moreover, I would like to thank all study participants – among which were also family, friends, as well as fellow PhD students – who have been willing to participate in research and who have thus contributed to the extension of scientific knowledge.

Finally, I would like to thank my family, especially my parents and sisters for all the love and support they have given me throughout the past four years, always willing to offer a sympathetic ear. I am very grateful for everything they have done to help me.

Table of Contents

List of Tables and Figures

List of Abbreviations

AI	-	Artificial Intelligence
ES	-	Emotional Stimuli
HSP	-	Highly Sensitive Person
HSPS	-	Highly Sensitive Person Scale
SD	-	Standard Deviation
SM	-	Supplementary Material
SPG	-	Sensitive Person Group
SPS	-	Sensory Processing Sensitivity

Chapter One – Introduction

Background

Emotions play a central role in human decision-making processes and hence affect behaviour. That is, in a situation of direct physical danger (e.g., fire), an individual may perceive fear and thus run away aiming to minimize the threat of the danger. The feeling of deep love that a parent perceives towards a newborn, assures caring behaviour such as providing protection and feeding. Nevertheless, trying to find a satisfying definition of the term *emotion* has turned out difficult, and until today, no generally and scientifically accepted definition has been made (Izard, 2007). Additionally, the question arises regarding how many emotions there are and how these can be differentiated.

Paul Ekman for example, who has done extensive research on facial emotion expression, concluded that there are six basic emotions given to the humans by evolution: *anger, fear, sadness, enjoyment, disgust* and *surprise* (1999). This means, that under normal circumstances these emotions can be expressed and understood throughout different cultures that had no prior contact, thus could not have learned from each other through experience. There is no doubt that Ekman's classification has been a valuable contribution to the psychology of emotions, but other researchers, such as Sabini and Silver criticized his list of "basic emotions" as being incomplete and added the two emotions *jealousy* and *parental love* (2005).

Other researchers rather emphasize the distinction of emotions from other similar categories, with the three main categories *affect, mood,* and *emotion* being formed (Ekkekakis, 2012): *Affect*, describes a neurophysiological state consciously accessible and most evident in mood and emotion (Russell & Feldman Barrett, 2009, p. 104) that is constantly experienced in varying intensity, for example energy and tiredness, pleasure and displeasure. *Mood* is more global, rather than specific, and mostly temporally remote from its cause (Morris, 1992), such as anxious or depressive mood. *Emotion* can be seen as triggered by a "complex set of interrelated sub-events concerned with a specific object" (Russell and Feldman Barrett, 1999, p. 806), and the cognitive appraisal that is involved is crucial. Examples are pride, jealousy and love. Further, Greenberg (2002) proposed a classification by differentiating between *primary* and *secondary* emotions. The primary emotion follows the situation directly; the secondary emotion implies an evaluation of the primary emotion.

Many differentiations have been made over the course of research on this subject, nevertheless, naming them all would go beyond the scope of this work. The meta-analysis

made by Kleinginna and Kleinginna (1981) where a classification into eleven categories has been proposed, can be consulted for more information and a broad representation. The authors concluded that "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (*a*) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (*b*) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (*c*) activate widespread physiological adjustments to the arousing conditions; and (*d*) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive" (p. 355). This conclusion is partly overlapping with the biological viewpoint, stating that an emotion is characterized by three aspects: (1) a specific functional state of the brain, (2) a typical subjective state, and (3) characteristic processes on somatic level (Schandry, 2011). Taken together, these definitions suggests that emotions are reactions on physiological, psychological, as well as behavioral levels.

With regard to the causes, effects, and the purpose of emotions, various theoretical positions exist among scientists. These are (*a*) physiological, (*b*) neurological, as well as (*c*) cognitive theories. The physiological theories suggest that emotions are caused following a response felt in the body: the James-Lange-theory for instance posits that in order to feel an emotion, a response from the peripheral expression of emotion back to the central nervous system is necessary (James & Lange , 1922); the Cannon-Bard theory states that physiological response to emotions can only form after stimulus perception and evaluation have taken place and that physiological arousal and emotional experience occur simultaneously, yet independently (Cannon, 1927). In contrast, the neurological theories suggest that emotional response is cause by activity within the brain, with unique neural circuits for distinct emotions (e.g., Moors, 2009), as well as specific neurons involved in the emotional response to faces (Rolls, 1990). Finally, the cognitive theories suggest that emotions are formed under the influence of thoughts and other mental activities. In this line, the Schachter-Singer theory for instance posits that a felt physiological state is cognitively interpreted and subsequently labeled as a distinct emotion (Singer & Schachter, 1962).

Within research, one of the main ways to investigate emotions is by using emotional stimuli (ES). These are for example images, words, or audio and video clips with a distinct emotional value that is assessed for each individual stimulus through a rating population. Against this background, the present work was conducted taking the position of cognitive theories of emotion, acknowledging that individual past experiences cognitively influence interpretation and hence perception of emotions elicited by emotional stimuli. While image and

video stimuli may for instance display content on a variety of subjects such as facial expressions, beverages and food, natural scenes, threat and crime, as well as social contexts, word stimuli range from individual words, to sentences or text passages. Audio stimuli include spoken individual words, sentences, pseudo language/gibberish, as well as sound and music. ES are often used to elicit emotions, support diagnostic processes, or train emotion expression and recognition. Word and speech stimuli for instance, may be used as an aid for diagnoses in a medical context (Haro et al., 2017; Nieuwenhuis-Mark et al., 2009), and video or image stimuli are applied within the frame of therapy for alcohol addiction (Pronk et al., 2015), or eating disorders/obesity (Miccoli et al., 2014; Miccoli et al., 2016). Furthermore, speech or music stimuli may be used for advertisements (Zander, 2006), learning contexts (Schön et al., 2008), or political and linguistic research (Cullen & Harte, 2018; Edelman et al., 1992), while images of facial expressions are useful for emotion recognition and emotion expression training for instance in children (Cardos et al., 2016), individuals with autism (Wingenbach et al., 2016), or schizophrenia (Gutiérrez-Maldonado et al., 2014). These stimuli can enhance intercultural understanding (Sacco et al., 2016), human-machine interaction (Battocchi et al., 2005), or also machine learning (Zafeiriou et al., 2016) within artificial intelligence (AI). Moreover, automatic recognition of anger in speech (Neiberg & Elenius, 2008) for example, is used in some systems of telephone services to indicate the customers' emotion to the staff. Within texts, emotion detection can be used as an indicator of the general population's mood, if applied to online messages, comments on websites, or blog entries (e.g., Ramalingam et al., 2018). Another example is using automatic emotion recognition in human movement (Crane & Gross, 2007) such as detecting aggression, which can be applied within video surveillance of prisons or public spaces for security reasons. The ability to automatically detect emotions is thus of great interest to many organizations, including private business and the government. The ground truth used for creating the algorithms training machines are ES. In other words: The range of applications for ES is wide.

Although all these above-mentioned examples provide only a small insight into the possible use of ES, they certainly reflect the important role of ES within emotion research, and hence the necessity of the availability of validated sets not only for the scientific community, but also for other contexts where eliciting and reacting to emotions plays a central role. Moreover, they urge for a profound understanding of set characteristics to ensure a suitable use of the stimuli.

The Large Number of Emotional Stimuli

Both, a growing interest in emotion research in humans, as well as a technological development pushing forward research within automatic emotion detection through AI and within human-machine interaction, have fuelled the creation of new stimuli sets. Numerous sets have been published and presented to the research community throughout past decades. While some sets were created specifically for the purpose of providing stimuli to the research community, others were originally created for use in a particular study. A large pool of available stimuli can facilitate research and the use of standardized ES can allow comparison of findings across laboratory settings. Nevertheless, it is difficult for researchers to compare a large number of sets across multiple characteristics when searching suitable stimuli for their own study construction. Additionally, some of the sets may be outdated, difficult to access, or unavailable. Moreover, stimuli sets may have been assessed in and/or validated for a specific language, ethnicity, age, or medical background of participants. Therefore, sets may not always be suitable for all research objectives or for particular populations. The selection of suitable ES for the research goals can thus become one of the most challenging aspects of designing a study involving emotions.

As a result, researchers may feel encouraged to use stimuli that have been used in many previous studies. A few well-known stimuli sets such as the *Affective Norms for English Words* (ANEW; Bradley & Lang, 1999a), the *International Affective Picture System* (IAPS; Lang et al., 1997), the *NimStim Set of Facial Expressions* (NimStim; Tottenham et al., 2009), or the film set by Gross and Levenson (1995), have established research credentials having been widely used in emotion-related studies. On the one hand this may help researchers to compare their own results to previous research that has used the same stimuli, on the other hand, the set's renown often overshadows smaller, recently published, or lesser-known sets and their repetitive use amplifies the gap between small, unknown, and well-known sets, reinforcing the overshadowing effect. Additionally, the use of stimuli from only a few well-known sets increases the possibility that individuals participating in numerous studies become familiar with stimuli. However, research has shown increased liking of stimuli due to repeated exposure (*mere-exposure effect*) (Palumbo et al., 2021; Zajonc, 1968), and thus an increased familiarity may affect stimulus validity.

A possible solution to the amplified use of only a small selection of stimuli sets could be the creation of an openly accessible database listing all ES sets published to this date. Next to facilitating access to as well as easing comparison across different sets, such an overview would narrow, even close the gap between well-known and less-known sets. An easier access

to the central set characteristic such as the type and number of stimuli, displayed content, included emotion(s), as well as aspect of the participant sample (e.g., age, gender, ethnicity) used for validation, or also applied assessment scales could help researchers, therapists or businesses in their search process when seeking specific emotional stimuli. This would not only considerably accelerate the research process, however, also allow a comprehensive consideration of all sets when selecting stimuli.

Reliability of Emotional Stimuli

Besides the difficulty of a time-consuming task when searching for suitable stimuli among all existing sets, another difficulty exists in relation to the use of ES, that is, the assured reliability of the stimuli. The specific emotional charge or value of stimuli is usually assessed regarding distinct emotions (e.g., happiness, sadness) or dimensions (e.g., valence, arousal) by numerous participants. The resulting rating data, also called *normative* data, is then provided to the research community along with the stimuli. Using the rating data as an indicator of the emotional value of stimuli allows fellow researchers to select stimuli for their specific research goal (e.g., emotion induction). In doing so, researchers rely on reliability and validity of the data from the norming sample. Nevertheless, various factors such as participants' age (Isaacowitz et al., 2007), gender (Lithari et al., 2010; Nater et al., 2006), ethnicity (DeBusk & Austin, 2011), or social and cultural background (Boiger et al., 2018; Matsumoto et al., 2008) are known to influence the perception of ES. Therefore, these factors regarding the rating population are usually mentioned in the study description or highlighted in relation to the normative rating data provided along with a stimulus set.

Yet, further factors may affect stimulus perception: In that regard research has for instance shown that rating results may differ depending on the implemented assessment scale (Bolton & Wilkinson, 1998; Brunier & Graydon, 1996; Hasson & Arnetz, 2005), as well as that stimulus presentation duration influences perceived pleasantness (Marin & Leder, 2016; Reber et al., 1998). Moreover, interpretation of symbol stimuli is influenced by context (Cahill, 1975; Wolff & Wogalter, 1998). That is, the symbol of a key for example, could indicate the location of a key-cutting store in a shopping mall, whereas in a computer it may indicate the need for a password. In a similar vein, some emotional stimuli may capture content that can be related back to a specific point in time such as a specific fashion/hair style, or type of technology in image stimuli, or outdated content such as archaic words in word and/or audio stimuli. These examples raise the concern that when stimuli are assessed today, the resulting rating data may not be equivalent to the normative rating data reported in the original study. In other words, the disparity in time between assessment of normative data and the use of stimuli today, may cause

changes in stimulus perception. Moreover, with increasing disparity in time, the risk of changes in emotion perception of stimuli also increases, hence reducing or even erasing stimulus reliability.

Among existing sets are stimuli that were created as long as 60 years ago (e.g., Barrington, 1963). Consequentially, if researchers are relying on normative data without verification of the validity of stimuli for their study sample, expected effects (e.g., specific emotion induction) may not result. This could lead to distorted study results and wrong conclusions drawn from these results. To give an example, a researcher aiming to investigate the facial expression in response to highly arousing image stimuli would for instance select existing stimuli based on the provided normative data indicating their high arousal value. If, however, – due to disparity in time – these stimuli are not perceived as highly arousing when used today, the facial expression of participants made in response to the presented stimuli does not reflect a response to highly arousing content.

Yet, it is common practice among researchers to use ES and rely on the normative data provided along with the set without verifying the validity of the stimuli for their study. That is, many researchers use available ES for their studies without questioning stimulus validity. However, this may have outreaching consequences: The wrong calibration of AI for example to automatically detect aggression in a public space for crime prevention purposes may cause erroneous detection but also overlook alarming signs. A teacher relying on emotion detection in class settings – as frequently applied in China – may for instance miss students' incomprehension of a subject. Finally, regarding research this may lead to the creation of inconclusive study results.

To ensure high standard research, as well as high accuracy in practice, it is therefore crucial to verify reliability of stimuli for today's use. Moreover, it seems central to understand the role of influencing factors (e.g., participants' gender, assessment scale) that have been shown to affect stimulus perception (see above). Therefore, it is important to identify distinct factors and to assess the magnitude of their influence onto emotion perception to help evaluating the reliability of stimuli in relation to these factors. Consequentially, this will help researchers to estimate the necessity of stimulus verification for their study sample depending on their own research aims in comparison to the original study construction.

Testing the Application of Emotional Stimuli on Memory

Of particular interest among emotion research has been the relation and dependence of cognition and emotion (see Barkus et al., 2010) and more specifically investigating the effect of emotion on memory. Previous research has shown that two brain areas, the hippocampus and

the amygdala, are key areas for regulation and production of emotion (e.g., Cahill & McGaugh, 1998; Deacon et al., 2002; Glascher & Adolphs, 2003; Gray, 1982; McGaugh, 2000). These areas are also associated with memory processes, consolidation of memory, and learning as well as spatial orientation (Scoville & Milner, 1957; Klüver & Bucy, 1937). In other words, the hippocampus and the amygdala, both, represent a linkage between emotion and memory as they associate memory content with emotional assessments when transferring temporary memories to other brain areas to form long-term memories (e.g., McGaugh et al., 2002; Nalloor et al., 2012).

Studies conducted to investigate the effect of emotions on memory mostly use ES conveying different emotions (e.g., sadness vs. happiness) or of varying emotion intensity (e.g., neutral, positive, highly positive) to compare the effect of distinct emotions/ emotion intensities. It is therefore a research area in which assurance of stimulus reliability is particularly important, as study results may otherwise lead to erroneous conclusions.

For example, Kensinger et al., (2004), used word stimuli to investigate the role of valence and arousal on memory. Results showed that distinct cognitive and neural processes contribute to emotional memory enhancement for arousing information (depending on an amygdalar– hippocampal network) compared to valenced, nonarousing information (supported by a prefrontal cortex– hippocampal network). Nevertheless, in their study, stimuli were selected based on the provided normative rating data, without being reassessed prior to study conduction. If, however, participants' perception of stimuli differs from the normative rating (in the provided example this would mean perceiving stimuli as arousing although originally categorized as nonarousing), the resulting measured effect (here, of valenced, nonarousing content) onto memory is at high risk of being misinterpreted.

In a similar vein, Libkuman et al., (2004) used image stimuli to investigate the role of valence and arousal in relation to memory for central and background details. The authors found that distraction after stimulus presentation decreases memory for negative stimuli compared to positive and is independent of arousal. Moreover, arousal increases memory for central details for positive and negative stimuli, and memory for background detail solely for positive stimuli. In this study, stimuli were reassessed prior to study conduction and the authors found differences regarding emotion perception between their assessment and the normative data. They justified the use of stimuli based on significant rating differences between low and high arousal as well as negative and positive stimuli. The authors conclude with the suggestion to also consider image detail when conducting research investigating the relation between emotion and memory. This study is an example highlighting the importance to verify stimulus validity

prior to study conduction as not all studies may in fact compare extreme rating categories (low to high valence /arousal) however, compare less extreme stimuli groups (e.g., low to medium valence/ arousal), for which in turn rating differences may not be significant anymore. Moreover, considering image detail may have a great impact on the procedure of future study conductions especially when using available ES, as stimuli are rarely assessed for contained detail.

Although a few meta-analyses regarding the relation between emotion and memory have been conducted, these are often very specific to a certain research area within this field. For example, Murphy and Issacowitz (2008) focussed on age and conducted a meta-analysis regarding memory and attention tasks comparing older and younger adults. The authors concluded that age significantly affected the effects for emotion salience, and that the measurement type appeared to influence the magnitude of effect. Similarly, Mather (2007) acknowledges the presence of contradictory research findings regarding the relation between arousal and memory binding within the field of emotion and memory. In conclusion the author therefore proposes an object-based framework, that is, the attention-grabbing nature of an object in visual stimuli is interfering with the working memory, to explain existing contradictory findings.

Moreover, on the one hand, research investigating the effect of stress (e.g., Kirschbaum et al., 1996; Schwabe & Wolf, 2010), anxiety (Harris, 1999; Harris & Cumming, 2003), sadness (Chepenik et al., 2007), happiness (Storbeck & Clore, 2005) and boredom (Goldberg & Todman, 2018), has shown an impairing effect onto memory. On the other hand, other existing research results display an improved memory for emotional content when learning is directly followed by stress (e.g., Cahill et al., 2003; Wolf, 2008). To summarize, contradictory research findings concerning the relation between memory and emotion exist and seem often highly dependent on study construction (e.g., the use of real-life events as material being retrieved, sample size, or participants' age) (Ucros, 1989).

As mentioned earlier, a possible reason for inconsistent research findings could be that researchers rely on stimuli without verifying validity for their participant sample. In fact, Quas and Lench (2007), conducted a study investigating the role of physiological arousal at encoding and retrieval of video content and found that memory was more accurate for individuals with increased arousal during encoding and less accurate for individuals with increase arousal at retrieval. These results show that physiological arousal in response to ES varies among participants and hence suggest a differing emotion perception of presented stimuli between individuals. This in turn suggests that next to known factors such as age, gender, or social

background (Boiger et al., 2018; Isaacowitz et al., 2007; Lithari et al., 2010; Matsumoto et al., 2008; Nater et al., 2006), the reason for a differing perception of emotional content could hence lay within personality differences such as trait sensitivity. High sensitivity for instance, describes the ability to process stimuli and information more strongly and deeply than others (Aron 1996c; Aron & Aron, 1997; Aron et al., 2010; Aron et al., 2012). That is, individuals with high sensitivity for instance adopt a strategy of pausing to analyse before acting, leading to increased responsiveness to subtle, environmental, and social stimuli such as loud noise or changes in temperature (Aron et al., 2012). In that regard, research has shown that individuals with high (vs. low) sensitivity perceived positive images as more arousing (Jagiellowicz et al., 2016). Consequentially, if emotion perception in high sensitive individuals differs significantly from perception in the remaining population, ES that have not been verified for this sample group may not have the intended effect. Moreover, if trait sensitivity affects emotion perception, it may also influence memory.

In fact, the lacking verification of personality differences within the participant samples could be a possible reason for inconclusive results regarding research conducted on emotion and memory, if not the *main* reason for existing inconclusive findings within emotion research. However, to this day, no research has investigated whether the reliability of emotional stimuli may be affected by trait sensitivity. With high sensitivity concerning approximately one fifth of the human population (Kagan, 1994), the risk of neglecting a large population percentage is substantial. Therefore, it seems necessary to verify stimulus validity for this population group. Additional investigations regarding the effect of emotion perception onto episodic recognition memory for high (vs. low) sensitive individuals will help to gain insight into possibly existing differences between these two groups regarding memory.

The Present Research

A large number of ES sets have been published to this date. However, a few sets such as the *IAPS* (Lang et al., 1997) or the *ANEW* (Bradley & Lang, 1999a), have been repeatedly used, emphasizing their appearance among published research while overshadowing smaller, less known sets. Additionally, various factors such as age, gender, or cultural background affect emotion perception of stimuli. Nevertheless, researchers usually select stimuli from existing sets without verifying the reliability of the normative data for their own participant sample – an approach that is of high risk of leading to distorted study results.

To ensure high standard research in the future, the central aim of the present research is to explore the question of stimulus reliability. More specifically, the present research aims to investigate factors that may affect validity and hence utility of ES within emotion research.

Prior to investigating possible factors, it was necessary to gain a comprehensive overview of all available stimuli sets and their key characteristics. Therefore, in a first step, a comprehensive review of existing stimuli sets was conducted leading to the creation of a database. In that regard, decisions made by the authors concerning the focus on distinct emotions (categorical approach) and/or the dimensional approach (emotions regarded for example in the dimensions valence, arousal, or dominance) were not questioned so that information could be maintained as provided in the original work. In a second step, a selection of ES was used to investigate reliability of emotional stimuli with regards to different factors that may determine stimulus reliability. The prevalence of stimuli found to be reliable when assessed in the present day was used as an indicator for stimulus reliability in relation to investigated factors. Finally, next to factors that have been scientifically investigated in the past and shown to affect stimulus perception (e.g., participant's age and gender), another factor influencing stimulus perception (and therefore affecting stimulus reliability) may be related to personality characteristics, more specifically, trait sensitivity. However, these are not typically assessed in research concerning ES perception and in consequence, the relation between trait sensitivity ES perception has not been sufficiently investigated to this day. Therefore, within the overall aim of the present work to explore validity and utility of ES, an experimental study was conducted in which participants' trait sensitivity as well as stimulus rating was assessed to explore their relationship. Due to the frequent use of ES within research focussing on the relation between emotion and memory, the conducted study moreover investigated the relation between emotion perception of individuals scoring high on sensitivity (vs. non-high sensitive), and episodic recognition for presented stimuli. As these last two steps were conducted partly as a type of study replication (reassessment of emotional stimuli), these were hence also conducted in front of the background of the cognitive theories. That is, by asking participants to indicate their perceived emotion on a virtual scale, the present studies are acknowledging that each participant may perceive the stimuli differently, and that assessment is influenced by the memory of past experiences triggered by the stimulus.

Together, the results of these three studies will *(a)* provide researchers with a comprehensive overview of existing ES sets, saving researchers time when searching for specific stimuli and allowing a more elaborate comparison across sets based on specific set characteristics; *(b)* assess the validity of normative data and thus stimulus reliability in relation to specific factors, hence indicate to researchers in which case reassessment of stimuli will be necessary for their own participant sample; and *(c)* investigate the aspect of trait sensitivity in relation to stimulus perception and recognition memory.

Principal Research Questions

Each of the steps (e.g., systematic review, investigating stimulus reliability in relation to different factors, investigating the effect of emotion on recognition memory) will be presented in an individual chapter. For each step (a) principal research question(s) were formulated. These questions are:

Q1: How many emotional stimuli sets are available to the research community? and *What are the key characteristics of each set?*
Q2: Do ratings of emotional image and word stimuli remain generally reliable numerous years post publication, and what factors related to the stimuli influence their reliability?
Q3: What is the relationship between trait sensitivity, emotion perception of stimuli, and recognition memory?

Outline of Chapters

The entire research project will be presented and discussed in five chapters: The introductory chapter (*Chapter I*), briefly outlines the theoretical background of emotion research along with the rationale for all three conducted studies. *Chapter II* will describe the work of a comprehensive systematic review of existing emotional stimuli sets, leading to the creation of the searchable online database (KAPODI database) in which stimuli sets can be searched and filtered according to set characteristics. In *Chapter III* the experimental study investigating different factors that may determine stimulus reliability through the example of image and word stimuli will be described. *Chapter IV* presents a study investigating the relationship between person sensitivity and perception of image stimuli along the dimensions of valence and arousal, as well as the effect onto recognition memory. Finally, research findings as well as implementations for future research will be discussed in *Chapter V*.

Chapter Two – A Systematic Review of Emotional Stimuli Sets

One of the most challenging aspects of designing a study involving emotions can be the selection of suitable ES for the research goals. Along with a rapidly growing interest in emotion research throughout the past few decades [e.g., emotion regulation (Gross, 2015) or emotional development (Pollak et al., 2019)], the need for ES has also increased. This is reflected in the publication of numerous ES sets. Examples of ES include images, words, music, speech, or video-clips that can be used in studies aiming to elicit specific emotions.

Following the original aim to investigate possible factors influencing stimulus reliability, it is crucial to gain an overview of existing ES sets prior to conducting any experimental study. The first study within the framework of the current research project on ES therefore aimed to systematically review previously published sets. To understand similarities and differences between sets it was moreover important to extract key set characteristics.

The three main research questions were formulated as follows:

Q1.1: How many emotional stimuli sets are available to the research community?
Q1.2: What is the prevalence with regards to different types of stimuli?
and
Q1.3: Which are the key characteristics of each set?

Background

The Origins of Normative Data

Various factors may play an important role in influencing perception and emotional experience in relation to stimuli. Examples are the participants' age (Isaacowitz et al., 2007), gender (Lithari et al., 2010; Nater et al., 2006), ethnicity (DeBusk & Austin, 2011), hormone levels (see Little, 2013 for a review), or social and cultural background (Boiger et al., 2018; Matsumoto et al., 2008). Therefore, as well as the actual ES themselves, validation data can be just as important to researchers when planning a study. This assessed rating data provided along with stimuli is called *normative* data.

Seeking to quantify subjective perception of ES, two main approaches have developed over time, namely the *dimensional* approach and the *categorical* approach. The former focuses on the three dimensions valence, arousal and dominance, rooted in research by Osgood and colleagues (Osgood, 1952; Osgood, Suci, & Tannenbaum, 1957); the latter is based on specific emotions, most frequently *the big six*, namely happiness, sadness, anger, fear, disgust, and surprise (Ekman et al., 1969). However, extensions have been proposed for both approaches, arguing that they were yet not exhaustive. Stevenson and colleagues for example suggest that

for the assessment of certain words it is advisable that valence, arousal and dominance are further distinguished from sexual valence, sexual arousal, and sexual dominance (Stevenson et al., 2011). The emotion categories have been extended for some studies by adding for instance moral disgust, joy, amusement, and tenderness (Ge et al., 2019), sarcasm/irony (Esposito et al., 2009), pride, contempt, embarrassment (Wingenbach et al., 2016), or also guilt, interest, and affection (Gilman et al., 2017).

As well as the preference for a specific approach to emotional categories, the *measurement* of emotional reaction to stimuli also differs across sets. Certain rating scales are preferred over others, depending on the research aim or participant sample. Lang and colleagues developed a picture-oriented assessment method, the Self-Assessment Manikin (SAM), (Bradley & Lang, 1994; Lang, 1980): in a dimensional approach, valence, arousal and dominance are assessed by displaying three sets of five figures arranged along a continuum. These SAM-scales have been widely used. Sometimes partially by assessing only one (e.g., Katsimerou et al., 2016) or two (e.g., Ferré et al., 2012) of the three dimensions, or extended by creating figurines for further dimensions for individual research questions such as significance (high vs. low), source (internal vs. external) (e.g., Imbir, 2015, 2016), or food craving (Miccoli et al., 2016). However most frequently, they are modified by inserting inter-pictorial steps without figures, creating 9-point scales with five SAM-figures (e.g., Fairfield et al., 2017; Goodman et al., 2016; Soares et al., 2012). Researchers have also created additional SAM-figures to represent each of the 9 points within one scale (Provost et al., 2015). Further scales that are frequently utilised in the dimensional approach are the Likert-scale (Likert, 1932) and the visual analogue scale (VAS) (Hayes & Patterson, 1921). Latter are indeed often the preferred tool in the categorical approach, applied for each specific assessed emotion (Ge et al., 2019; Stadthagen-González et al., 2018; Wierzba et al., 2015).

With a value that represents the emotional charge of a stimulus, researchers can select stimuli according to their research aims. For example, for a study aiming to induce sad emotions through music, validated music stimuli can be accessed, and their accompanying normative data used as an indicator for their degree of sadness. In a study that aims to compare the perception of facial emotion expression across different age groups, stimuli that have been validated for one specific age group could then easily be accessed, sparing the researcher from assessing the same age group again. Furthermore, studies focussing on memory in connection to emotion could benefit from stimuli that have been validated previously in eliciting specific emotions. Finally, there are stimuli validated for specific uses such as in studies with noisy backgrounds, for instance fMRI assessments (e.g., Lepping et al., 2016). Pre-validated sets can

thus save researchers time as they do not need to create and validate their own stimuli beforehand. Finally, providing the normative rating along with the stimuli enables researchers to replicate studies or to manipulate (or control for) external factors such as country of the survey, year of the study, ethnic background of participants etc., but also internal factors such as luminance, colour, display duration, or video/audio speed of presented stimuli.

However, relying on normative rating data provided along with the stimuli can be a double-edged sword: while researchers are assured that chosen stimuli will have the intended effect on most participants, the normative ratings may have been affected by a plethora of external factors. Hence, a set validated for one population or context, may not have the same effect, thus not be valid, in another population or context. Research investigating the emotional perception of images in countries suffering from violence showed that Israeli adults rated images differently than young adults in the United States (Okon-Singer et al., 2011). Similarly, in a study investigating the interpretation of symbols, Cahill (1975) was able to show that context eases interpretability compared to symbols in isolation. Further, the year the survey was conducted may play an important role: that is, an image of the World Trade Center in New York City, presented to participants prior to the 11$^{th\ of}$ September, 2001, may have elicited different emotions, compared to after that date, and images or video stimuli may include cues such as hair style or fashion that can easily be associated with a specific decade and thus seem outdated when seen today. Unfortunately, not all available stimuli sets record or highlight these factors.

Therefore, it is important to consider the details of ES set construction such as characteristics of the rating population, date of created stimuli, or country of research, etc., when selecting stimuli and/or relying on normative data. Especially regarding research conducted in relation to automatic emotion detection through AI, a cautious selection of stimuli based on normative data is particularly important, as a 'wrong calibration' of stimuli used to train the machines may have outreaching consequences (e.g., automatic detection of aggression).

Research Rationale

For the above reasons, a central database giving an overview of available ES sets and documenting their central characteristics is needed. Additionally, with every year and every new ES set that is published, the need for an updated and exhaustive database is increased. Although previous research reviewing emotional stimuli sets has been conducted by Krumhuber et al. (2017) specifically for dynamic sets of facial expressions, as well as by Grühn and Sharifian (2016), these attempts are not comprehensive, as they focus only on specific types of ES, or have not systematically reviewed existing literature. Some researchers give a short

overview of existing stimuli sets for context when introducing their own new set, however, these mostly include sets of similar content such as exclusively *words* (e.g., Riegel et al., 2015; Scott et al., 2019) or *faces* (e.g., Prada et al., 2018; Tu et al., 2018).

With no existing comprehensive review to date, the objective of this study was to systematically review existing and freely available sets of ES and provide an overview by documenting the central characteristics of each set. To achieve this, a database was created in which stimuli sets were listed and coded with respect to specific criteria such as type and number of stimuli, included emotion(s), number of raters (where applicable), and applied rating scale(s). Moreover, a searchable online version of the database was created. This online version may serve as a tool allowing specific set characteristics to be selected, leading to the display of filtered results. To keep the content updated, newly published sets can be added continuously by other reserchers. The hope is that the resulting searchable KAthrin POs. DIconne (KAPODI) database of emotional stimuli will be a useful resource for researchers planning studies including ES, and possibly be directly beneficial to other contexts such as in a therapeutic setting, the creation of avatars and cartoon characters, or also human-machine interaction.

Method

The methodological procedure of the current systematic literature review consisted of three main building stages: first, a systematic literature review was conducted aiming to detect all existing ES sets; second, information of all included sets was coded; third, set characteristics were summarised for visualization of the results and a searchable online version of the database was created. This systematic literature review was hereby conducted largely independently. That is, while key-word selection was discussed and agreed upon by the study supervisors, reading, coding, as well as transferring of the data into the online-version of the resulting database was completed by the author of the present work. Due to the scope of the review, this was a time-consuming task taking multiple months. Hence, an updated review was necessary to verify if additional sets had been published in the meantime (see *Results, Stage One: Systematic Literature Review*).

Stage One: Systematic Literature Review

In order to capture the greatest possible number of papers proposing ES sets, an appropriate keyword selection had to be made and inclusion- as well as exclusion criteria, determined. In two consecutive steps publications not meeting inclusion criteria were excluded.

Information Sources

The keyword search was conducted in April 2019, on *PsychInfo*, *Medline (EBSCOhost)*, and *Web of Science*. The time frame for publication date was set to 1950-2019 for PsychInfo

and Medline, as well as 1970-2019 for Web of Science, respectively, as the early beginnings of emotion research and proposal of stimuli sets can be pinpointed approximately to the 1950s (Osgood, 1952). The reduced time frame for Web of Science was restricted by the database entry options. Due to a long time-spanning coding process after the first search, the same search was conducted a second time in June 2020 aiming to detect all studies presenting ES published between January 2019 and June 2020. Literature search included emotion- and stimuli-related keywords and were kept limited to the six basic emotions (Ekman et al., 1969). The applied exact keywords were: *"(emotional) OR (emotion) OR (affect) OR (affective) OR (fear) OR (disgust) OR (happiness) OR (anger) OR (angry) OR (sad) OR (sadness) OR (surprise) [IN all text] AND (stimulus OR stimuli OR picture$ OR word$ OR video$ OR audio OR film$ OR sentence$) [IN all text] AND (set OR database OR list OR library OR norms) [IN title]"*. The keyword *database* could only be searched for [IN title], as many search-engine databases commonly include this term below a paper's abstract, leading to tens of thousands keyword search results when searched for [IN all text] or [IN abstract].

Eligibility Criteria

Papers were selected according to the criteria below:

To be included, papers had to *(I)* be peer-reviewed; *(II)* be published in English, French, or German; *(III)* be published between 1950 and 2020; *(IV)* include ES that are either *(a)* images; *(b)* video; *(c)* audio; or *(d)* words; and *(V)* include sets accessible to the research community. Excluded were all sets containing ES such as heat, pressure, or odour, sets of ES created for animal studies, as well as sets providing solely physiological data stimuli (e.g., to train AI).

As discussed earlier, various factors can influence the perception of ES; therefore, in this systematic literature review, validation of presented stimuli was not considered a prerequisite. Publication without validation was sometimes the case for studies where models were asked to express certain emotions, (e.g., Minear & Park, 2004; O'Toole et al., 2005; Yingliang et al., 2006), or also for word lists created by the researchers themselves (Barrington, 1963).

Study Selection

Literature search results were uploaded to Rayyan Software (Ouzzani et al., 2016), an internet-based software program that facilitates the study selection process. All search results were manually and independently screened against inclusion criteria, based on title and abstract. Uncertainties concerning inclusion were resolved through discussion and consensus of two to four researchers. When necessary, additional information was sought directly from study

authors. Concerning inquiries regarding availability, authors were contacted twice via e-mail within approximately two to three months. If they did not respond after the second enquiry, the set was considered unavailable.

Stage Two: Data Collection Process and Data Items

Each publication that met inclusion criteria was then read two times independently and assigned to one of six subfolders: *(1) audio*, *(2) faces*, *(3) images*, *(4) video, (5) words*, and *(6) mixed*, depending on the type of most stimuli included in the presented set. Characteristics (e.g., year of the publication, type of stimuli, number of stimuli, resolution, number of raters, or applied rating scales) were coded. A detailed outline of these characteristics can be found in Table 1. Whenever information was not provided in the paper, this was noted as not available (n/a). In case of information inconsistencies within the publication, or to resolve any uncertainties, study authors were contacted.

Table 1

Coded Characteristics for Each Subfolder

Coded characteristics	Subfolder					
	Audio	*Faces*	*Images*	*Video*	*Words*	*Mixed*
Title of publication	yes	yes	yes	yes	yes	yes
Authors	yes	yes	yes	yes	yes	yes
Year of publication	yes	yes	yes	yes	yes	yes
APA citation	yes	yes	yes	yes	yes	yes
University affiliation	yes	yes	yes	yes	yes	yes
Stimuli set name	yes	yes	yes	yes	yes	yes
Type of stimuli	yes	yes	yes	yes	yes	yes
Resolution	yes	yes	yes	yes	-	yes
Content	yes	yes	yes	yes	yes	yes
Expression authenticity	yes	yes	-	-	-	yes
Ethnicity	-	yes	yes	yes	-	yes
Number of stimuli	yes	yes	yes	yes	yes	yes
Stimuli length	yes	yes	-	yes	yes	yes
Number of models	yes	yes	-	yes	-	yes
Sex of models	yes	yes	-	-	-	yes
Age of models	yes	yes	-	-	-	yes
Specific number of emotions	yes	yes	yes	yes	yes	yes
Colour / hue	-	yes	yes	yes	-	yes
Language	yes	-	-	yes	yes	yes
Categorical approach	yes	yes	yes	yes	yes	yes
Dimensional approach	yes	yes	yes	yes	yes	yes
Categorisation	yes	yes	yes	yes	yes	yes
Rating scale(s)	yes	yes	yes	yes	yes	yes
Additional Information	yes	yes	yes	yes	yes	yes

Included specific emotions	yes	yes	yes	yes	yes	yes
Number of included specific emotions	yes	yes	yes	yes	yes	yes
Validation for subgroup	yes	yes	yes	yes	yes	yes
Rating by	yes	yes	yes	yes	yes	yes
Student/non-student raters	yes	yes	yes	yes	yes	yes
Country of study	yes	yes	yes	yes	yes	yes
Continent of study	yes	yes	yes	yes	yes	yes
Number of raters	yes	yes	yes	yes	yes	yes
Number of raters per stimulus	yes	yes	yes	yes	yes	yes
Source access	yes	yes	yes	yes	yes	yes

Note. University affiliation = author-affiliation as indicated in the publication; content = short description of the database content; expression authenticity = e.g., natural vs. acted or posed, applicable to only emotional expressions (i.e. vocalisations or faces); categorical approach = stimuli assessed after the categorical approach of emotion; dimensional approach = stimuli assessed after the dimensional approach of emotion; categorisation = assessed dimensions; additional information=information that seems important that could not be coded otherwise; validation for subgroup = sets that have specifically been validated for a subgroup/ information for certain research fields are proposed by authors; rating by = characteristics of stimuli raters; student/non-student raters = coded whether raters were university/college students.

Stage Three: Summarization and Visualization of Results

All extracted information was coded in an Excel sheet. An online version of the database was created. It serves as a search tool in which specific criteria such as type of stimuli, models' age, rating scales, included emotions, etc. can be selected, leading to the display of solely sets containing these characteristics. The searchable database can be found online.

Results

Stage One: Systematic Literature Review

The first keyword search yielded 1,877 pieces of published work (443 in PsychInfo, 393 in Medline, and 1,041 in Web of Science). Duplicates (n = 616) were removed, 1,261 search results remained for manual scanning. This manual scanning was conducted in two subsequent steps: first, a coarse selection based on title and abstract, then thorough reading of the full publication. Based on title and abstract, n = 951 results were excluded due to unrelated content (e.g., studies on animals, publications originating from chemistry or physics), leaving 310 publications for thorough reading. In this second step, another 73 publications were excluded because *(a)* their content was not relevant to the systematic review (n = 56), or *(b)* the described set was not available to the research community/authors did not respond to e-mail requests

concerning availability of the set (n = 17). A more detailed overview of the individual steps can be found in the PRISMA flow diagram (Moher et al., 2009) in Figure 1.

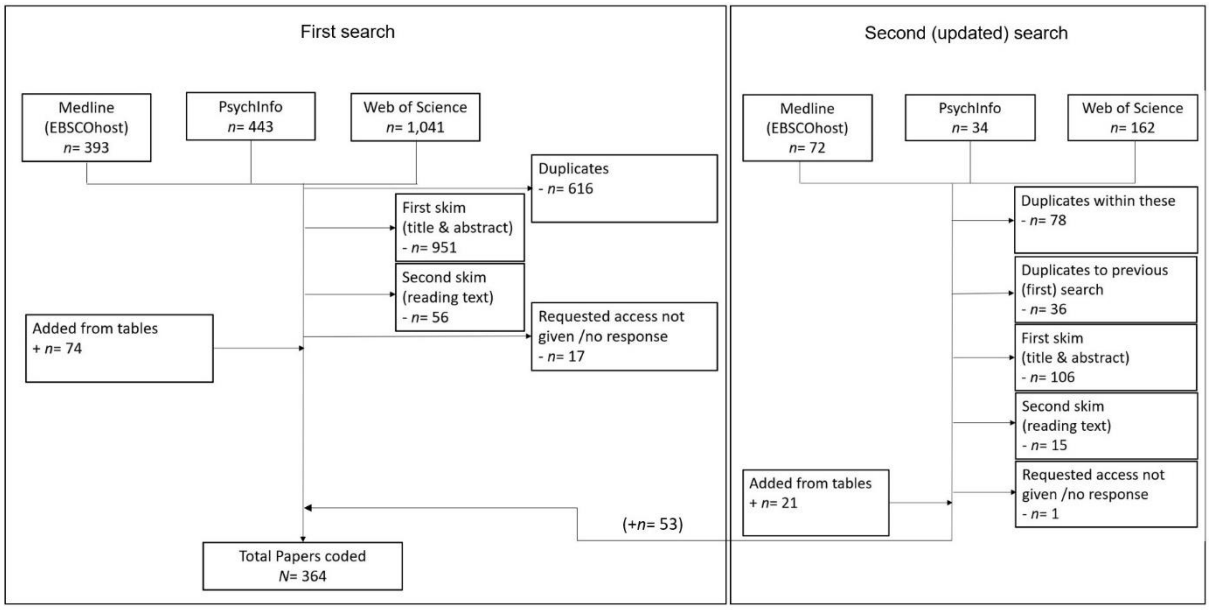


Figure 1. PRISMA flow diagram of the research procedure. *Note*. Research procedure conducted applying key words in three databases (PsychInfo, Web of Science, and Medline (EBSCOhost)). First search conducted in April 2019, and the second (updated) search conducted in June 2020.

A few publications mentioned further stimuli sets that were not detected by the initial key word search. Therefore, the availability of all sets presented in tables within the publications that were already part of the database (extended table-search) was further verified. All additional sets conforming to the inclusion criteria as mentioned above were included: Another n = 74 publications were added. At this point, a total of 311 publications had been included and coded. The same search and selection process were conducted for the second updated search covering all studies published between January 2019 and June 2020.

This updated keyword search yielded another N = 268 results (34 in PsychInfo, 162 in Medline, and 72 in Web of Science). Duplicates within these three search databases (n = 78), as well as duplicates of results from the first search (n = 36) were removed; 154 papers remained for manual scanning. A further n = 106 publications were removed based on title and abstract; a further n = 15 papers were removed after thorough reading. One paper was excluded, as authors did not respond to the request regarding set availability. A total of N = 53 publications were added through this second updated search: an initial n = 32 publications, and an additional n = 21 publications from the extended table-search.

With 311 publications from the first search and 53 publications from the updated search, at the point of creation the database contains a total of N = 364 publications. Each publication

presents at least one set of ES and/or new assessed rating data. All publications and their extracted main criteria are listed in an Excel spreadsheet available as *Supplementary Material (Study 1)*, and an online version of the database is also available. Note that the supplementary material contains information only up to 2020, while the online version of the database will keep being updated.

In the following section, analyses of main extracted aspects are presented:

Number of Publications

Publication dates span from 1963 to 2020. Separation by decades clearly indicates a steep increase in publications throughout the last two decades (see Table 2). Note, that the final keyword search in this study was conducted in June 2020, therefore the reported number for the last decade (2011 - 2020) may be underestimated if additional stimuli sets are published before the end of this year.

Table 2

Number of Publications and Percentage per Decade From 1961-2020 in Total and per Subfolder

Subfolder		1961-1970	1971-1980	1981-1990	1991-2000	2001-2010	2011-2020
	n	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Audio	35	-	-	-	3 (8.57)	9 (25.71)	23 (65.71)
Faces	117	-	-	-	5 (4.27)	32 (27.35)	80 (68.38)
Images	35	-	-	-	-	3 (8.75)	32 (91.43)
Mixed	45	-	-	-	-	10 (22.22)	35 (77.78)
Video	43	-	-	-	2 (4.65)	6 (13.95)	35 (81.40)
Words	89	1 (1.12)	2 (2.25)	3 (3.37)	6 (6.74)	19 (21.35)	58 (65.17)
∑	364	1 (0.27)	2 (0.55)	3 (0.82)	16 (4.40)	79 (21.70)	263 (72.25)

Note. $N = 364$ publications. Percentages calculated horizontally.

Stimuli Subfolders

For easier comparison across stimuli, each publication was allocated to one of six subfolders describing the type of stimuli: *(1)* audio ($N = 35$), *(2)* faces ($N = 117$), *(3)* images ($N = 35$), *(4)* video ($N = 43$), *(5)* words ($N = 89$), and *(6)* mixed ($N = 45$). Note, that the number of *publications* is named, as some publications present more than one set (e.g., *FEEDBver.1* and *FEEDB*, Szwoch, 2014; *ATAL* and *ETAL*, Torkamani-Azar et al., 2019). Furthermore, publications were allocated to the best fitting subfolder which means that included stimuli types are not exclusive: the folder *faces* for instance, contains sets of images as well as video stimuli, which are not separately named in the *images* or *video* folder. It is also important to mention that some sets (partially or in their entirety) have been translated into another language, validated in another country, or validated for a different age group, and thus do not contain *new* stimuli. However, they present new normative rating data. Such examples are the translation of

the ANEW set (Bradley & Lang, 1999a) into Italian (Montefinese et al., 2014) or Portuguese (Soares et al., 2012); the validation of the IAPS (Lang et al., 1997) for a Brazilian population (Ribeiro et al., 2005), a population of countries suffering from violence (Okon-Singer et al., 2011), or the creation of a subset relevant for Borderline Personality Disorder (Sloan et al., 2010); the creation of an audio version of an existing word set for cross-modal validation (Kanske & Kotz, 2011) or clinical subsamples (Kanske & Kotz, 2012); the validation for different age groups as can be found with the *Besançon Affective Picture Set* (BAPS) (*BAPS-Ado*, Szymanska et al., 2015; *BAPS-Adult*, Szymanska et al., 2019) or also the modification of stimuli for example by morphing existing face images and creating dynamic stimuli as can be found with the *FACES* set (Ebner et al., 2010) modified into *DynamicFACES* (Holland et al., 2019), and the *KDEF* set (Bartlett et al., 1999) modified into *KDEF-dyn* (Calvo et al., 2018). A short description of results per subfolder follows:

Audio

Audio-stimuli contain spoken individual words, sentences, pseudo language/gibberish, as well as music. The focus of emotion varies between intonation and semantic content of stimuli. This means that stimuli can be selected with a focus on perception of emotional tone, or emotional content, or a combination of both for example by using stimuli with emotional semantic content expressed in various emotional tones. Across all types of stimuli, *audio*-stimuli resulted in the fewest number of publications (equal to *image*-stimuli).

Faces

A specific focus within video and image stimuli are facial expressions. This subfolder contains the most publications in relation to the other folders and accounts for almost one third (32.14%) of all sets. A growing interest in automatic emotion detection within AI and progress in human-machine interaction are pushing researchers to continuously adjust and improve algorithms that strongly depend on standardised and/or validated sets of face-stimuli. Proposed sets contain both grey-scale stimuli and colour-stimuli, in 2D as well as 3D. Furthermore, still-image stimuli, as well as video and dynamic stimuli are included. Dynamic stimuli can be constructed artificially based on individual images. Similarly, this folder also includes sets with morphed stimuli which are stimuli that are created by superimposing multiple images or video frames. In some cases, researchers provide video streams, as well as individual frames (image-stimuli) of the recorded videos. Standardisation such as fixed pupil position across all models from one set, or removal of potentially distracting information (e.g., jewellery, hair, makeup, clothes) is coded. Both posed, as well as natural/spontaneous emotion expressions are included. Sets also vary in the degree of homogeneity regarding models' gender, age, and ethnicity.

Images

Image-stimuli in this database cover a variety of subjects. Examples are beverage and food images, natural scenes, threat and crime stimuli, as well as line drawings of social contexts. Image sets were often created for research in specific subgroups such as alcohol addiction, eating disorders, or phobia.

Video

Video-stimuli are one of the most effective type of stimuli for emotion elicitation (Gross & Levenson, 1995; Westermann et al., 1996) and thus are often used in studies aiming to induce a specific emotion in participants. Stimuli vary from video-only, to audio-visual stimuli that are accompanied by speech or music. Included are sets that present video clips extracted from movies and TV-shows, clips recorded specifically for the study, as well as motion-capture data where only point-lights representing body-part position in space are visible. Especially when stimuli are extracted from movies or TV-shows, quality and camera angle as well as microphone sources can vary widely within one set.

Words

Word-stimuli are stimuli that range from written individual words, to sentences or text passages. Similar to *audio*-stimuli, sets include real words in various languages, as well as gibberish speech. Multiple studies report extracting a selection or using all words from an already existing set such as the ANEW (e.g., Nieuwenhuis-Mark et al., 2009; Schmidtke et al., 2014; Sutton & Altarriba, 2016).

Mixed

The *mixed* subfolder was created to list all stimuli sets that cannot clearly be allocated to any of the other subfolders. Often, these sets contain a combination of types of stimuli or additionally provide physiological data such as for example of respiration, heart rate, skin conductance, or temperature. Physiological data recordings can be extremely valuable to research focussing on emotion recognition through AI, or also the understanding of physiological processes during the experience of different emotions. This sort of data may for instance be useful for the investigation of emotion regulation or understanding of disabilities in relation to emotion such as anxiety, apathy, or psychopathology.

Stage Two: Data Collection Process and Data Items

Number of Stimuli

Depending on the aim of research, sets were either created for a specific study (e.g., Belhumeur et al., 1997) and then made available or, in most cases, created specifically with the aim to be presented to the research community as a stimulus set. Further, sets vary between

presenting stimuli specifically for emotion elicitation (e.g., Lepping et al., 2016; Maffei & Angrilli, 2019; Samson et al., 2016), and stimuli that rather 'represent' specific emotions (e.g., Likforman-Sulem et al., 2017; Song et al., 2019).

Influenced by the research aim, the number of stimuli included in sets widely varies (Table 3): *audio*-stimuli sets contain 5 to over 14,900 stimuli (mean: appr. 1,404.89; *SD*: 2,864.04), and *face*-stimuli sets contain 42 to 1,503,495 stimuli (mean: > 37,674.15; *SD*: 201,855.25). *Image*-stimuli sets contain 15 to 2,941 stimuli (mean: 535.29; *SD*: 601.11), *video*-stimuli sets contain 14 to 9,800 stimuli (mean: 964.88; *SD*: 2,189.52), *word*-stimuli sets contain 24 to 23,475 stimuli (mean: > 1,708.79; *SD*: 3,386.95), and *mixed*-stimuli sets contain 10 to 22,326 stimuli (mean: > 1,975.77; *SD*: 4,066.04). Four publications could not be included as information was not available or, as in one case, stimuli were reported in available recorded minutes (e.g., *37h 13m*) rather than a number (Nazareth et al., 2019).

Table 3

Mean, Median and SD for Number of Stimuli per Publication for Each of the Six Subfolders

Subfolder	*n*		
	M	median	*SD*
Audio	> 1,405	240	2,864
Faces	> 37,674	535	201,855
Images	535	276	601
Video	965	147	2,190
Words	> 1,709	718	3,387
Mixed	> 1,976	450	4,066

Note. N = 360 publications. > = over (exact number cannot be calculated); M = mean; *SD* = standard deviation.

Naming and comparing the exact mean number of stimuli across sets therefore remains impossible. Furthermore, authors use different ways of reporting the number of stimuli, impeding comparison between studies: while some authors may report a certain number of available stimuli (e.g., videos) and may mention that individual frames are also available separately, others may report number of videos and frames available together, inflating the number of available stimuli reported. An example is the publication by Lubis et al. (2018), in which the number of recorded sessions (n = 60) is reported and manually refined transcriptions of the conversations are provided, however not counted as stimuli. The *PersianESD* (Keshtiari et al., 2015) in contrast, reports a total number of N = 558 stimuli, which are divided into n = 90 sentences in text format plus n = 468 vocal recordings thereof.

Assessment Approach

As previously discussed, quantification of the subjective perception of ES has typically taken either a dimensional or categorical approach. When assessing stimuli through at least one approach, categorical (55.22 %) and dimensional approach (56.59 %) have been used to an equal amount. Seventy-seven publications (20.88 %) assessed stimuli using both approaches. Separated by subfolder, the decision for either the categorical approach or the dimensional approach was taken approximately equally as often in publications presenting audio (71.43 % and 68.57 %), video (69.77 % and 55.81 %), and mixed stimuli (60 % and 55.56 %). The gap between both approaches in the remaining subfolders was more pronounced: face stimuli were more often assessed using the categorical approach (80.34 %), while image and word stimuli were mainly assessed using the dimensional approach (images: 100 %; words: 82.02 %). A total number of 35 publications (9.62 %) did not assess their stimuli using either of the two approaches. However, the number remained below 16 % for each of the subfolders (see Table 4). For these cases, other evaluation methods were for example coding of action units (AUs) in faces (e.g., Cosker et al., 2011; Mavadati et al., 2013; Savran et al., 2008), free recall of words (e.g., Nieuwenhuis-Mark et al., 2009), or rating of technical correctness, expressivity and beauty of dance movements (Christensen et al., 2019).

Table 4

Number of Publications per Subfolder Assessing Stimuli After Categorical and Dimensional Approach

Subfolder	*n*					
	N	Categorical	Dimensional	Both	None	%
Audio	35	25	24	15	1	
Faces	117	94	25	20	18	100
Images	35	10	35	9	0	75
Video	43	30	24	12	1	50
Words	89	15	74	7	9	25
Mixed	45	27	25	13	6	0
∑	364	201	206	76	35	

Note. ∑*N* = 364 publications. *Both* = stimuli assessed with both, categorical and dimensional approach; *none* = stimuli not assessed through categorical or dimensional approach; note: read colours of heatmap in table horizontally.

Dimensional Approach

Rooted in research by Osgood and colleagues (Osgood, 1952; Osgood et al., 1957), the dimensional approach focuses on the three dimensions valence, arousal and dominance. For all 206 publications assessing emotional stimuli using the dimensional approach, valence was always of highest interest (97.09 %), followed by arousal (79.61 %) and dominance (24.76 %)

(see Table 5). This order of preference was also maintained when regarding the subfolders individually (valence: 88 % to 100 %; arousal: 64 % to 95.83 %; dominance: 12 % to 40 %). The only exception can be found for video stimuli where valence and arousal were assessed equally as often (95.83 %).

Table 5

Number of Publications per Subfolder Assessing Stimuli on Valence, Arousal, and Dominance

Subfolder		*n* (%)			
	N	Valence	Arousal	Dominance	%
Audio	24	24 (100)	19 (79.17)	5 (20.83)	
Faces	25	22 (88.00)	16 (64.00)	3 (12.00)	100
Images	35	33 (94.29)	32 (91.43)	7 (20.00)	75
Video	24	23 (95.83)	23 (95.83)	8 (33.33)	50
Words	74	73 (98.65)	52 (70.27)	18 (24.32)	25
Mixed	25	25 (100)	22 (88.00)	10 (40.00)	0
∑	206	200 (97.09)	164 (79.61)	51 (24.76)	

Note. ∑*N* = 206 publications. Read colours of heatmap in table horizontally.

Categorical Approach

Based on specific emotions, the categorical approach most frequently relies on *the big six*, namely happiness, sadness, anger, fear, disgust, and surprise (Ekman et al., 1969). As explained earlier, extensions have been proposed (e.g., Esposito et al., 2009; Ge et al., 2019; Gilman et al., 2017; Stevenson et al., 2011; Wingenbach et al., 2016). Throughout all publications including emotions, between 1 and 93 distinct emotions have been assessed (Table 6). While images were assessed on a maximum of 6 different emotions, faces were assessed on up to 93 different emotions (Schmidtmann et al., 2020). However, this study reporting such a great number of emotions was an exception, as the mean number of assessed emotions (mean: > 7.09; *SD*: 9.47) remained relatively similar to that of audio stimuli (mean: 6.12; *SD*: 4.39), mixed stimuli (mean: 6.43; *SD*: 4.69), and video stimuli (mean: 8.23; *SD*: 7.67). For a comparison, word stimuli were assessed on a mean of 4.56 emotions (*SD*: 2.26), while image stimuli were assessed on a mean of 2.90 emotions (*SD*: 1.91).

Table 6

Mean, Median, and SD for Number of Emotions Assessed per Publication for Each of the Six Subfolders

Subfolder	*n*		
	M	median	*SD*
Audio	6.12	5.50	4.39
Faces	> 7.09	6.00	9.47
Images	2.90	2.50	1.91
Video	8.23	6.00	7.67
Words	4.56	5.00	2.26
Mixed	6.43	6.00	4.69

Note. N = 226 publications. > = over (one study did not mention exact number); M = mean; SD = standard deviation.

Assessment Scales

An aspect that is often relevant for research including ES is the measurement scale used for stimulus validation. Although not all publications present stimuli sets validated through participant assessment (e.g., validation via algorithms), those that did used various different scales. In three cases, information regarding the scale was not accessible. Therefore, the following calculations are based on N = 361 publications:

Overall, the Likert scale was the most applied scale (55.68 %), followed by forced-choice answer option (36.84 %), SAM-scale (21.05 %), and visual analogue scale (9.70 %). Other forms of assessment for example though free answers or assessment tools such as FeelTrace (Cowie et al., 2000), G-trace (Cowie et al., 2013), joysticks, or 2D spaces, were used in 20.78 %. Exact numbers of scale applications along with percentages can be found in Table 7. These percentages can exceed 100, as many studies assessed stimuli ratings on multiple scales.

Generally, apart from face-stimuli that were mostly assessed through the forced-choice method (52.17 %), assessment using the Likert scale was applied more often than any other scale within the remaining 5 subfolders.

Table 7

Number of Publications and Percentage per Subfolder and in Total Applying Forced Choice, Likert-Scales, SAM-Scales, VAS-Scales, and Other Scales

Subfolder		*n* (%)				
	N	Forced choice	Likert	SAM	VAS	Other
Audio	35	19 (54.29)	21 (60.00)	8 (22.86)	5 (14.29)	7 (20.00)
Faces	115	60 (52.17)	46 (40.00)	2 (1.74)	9 (7.83)	21 (18.26)
Images	35	9 (25.71)	22 (62.86)	15 (42.86)	8 (22.86)	1 (2.86)
Mixed	44	20 (45.45)	25 (56.82)	15 (34.09)	6 (13.64)	13 (29.55)
Video	43	16 (37.21)	26 (60.47)	8 (18.60)	2 (4.65)	16 (37.21)
Words	89	9 (10.11)	61 (68.54)	28 (31.46)	5 (5.62)	17 (19.10)
∑	361	133 (36.84)	201 (55.68)	76 (21.05)	35 (9.70)	75 (20.78)

Note. ∑*N* = 361 publications. VAS = Visual Analogue Scale; other = includes tools such as FeelTrace, G-trace, joysticks, or 2D spaces; percentages calculated horizontally and can exceed 100, given that many studies used multiple scales.

Of all 76 publications applying SAM-scales, 9-point scales were most often used (81.58 %), followed by 5-point scales (17.11 %), and 3- as well as 11-point scales (both 1.32 %). In one study information concerning the applied SAM-scale was not available (1.32 %).

The overall length of Likert scales ranged from 2-point to 21-point scales: throughout the 201 publications applying the Likert-scale, 5-point scales were mostly used (36.82 %), followed by 7- (34.33 %), 9- (26.37 %), 3- (5.97 %), 6- (5.47 %), 10- (4.48 %), and 11-point scales (3.48 %). Four- and 8-point scales were each applied in 2.49 % of the publications, and 2-, 15- and 21-point scales each below 1 % of the publications. Detailed information can be found in Table 8.

Table 8

Number of Applied SAM- and Likert-Scales in Total and per Subfolder

Subfolder	*n*-pt. SAM-scale					*n*-pt. Likert-scale											
	n/a	3	5	9	11	2	3	4	5	6	7	8	9	10	11	15	21
Audio	1	-	-	7	-	-			10	2	8	1	6	1	1	-	1
Faces	-	-	1	1	-	-	6	1	16	5	14	-	6	5	2	2	-
Images	-	-	2	13	-	-	-	-	9	-	7	-	8	-	-	-	-
Mixed	-	-	3	11	1	1	1	2	11	1	5	-	5	1	1	-	-
Video	-	1	2	6	-	-	2	1	12	1	4	1	13	1	-	-	-
Words	-	-	5	24	-	-	3	1	16	2	31	3	15	1	3	-	-
∑	1	1	13	62	1	1	12	5	74	11	69	5	53	9	7	2	1

Note. N = 201 publications. N/a = information not provided in publication; *n*-pt. = *n*-point scale.

Stimuli containing faces are frequently also annotated regarding facial Action Units (AUs) following the *FACS* coding (e.g., Ekman, 1982; Ekman et al., 2002; Ekman & Friesen,

1978). AUs can occur in more than 7000 complex combinations (Cohn et al., 2007) however, they can be consulted as objective measure of (emotion) expressions and are reliably distinguishable from each other (Ekman, 1982). They further enable tracking and detection of changes in emotion expression over time. In video stimuli for instance, annotation of micro expressions becomes possible. These are expressions that are subtle and last under approximately ½ second (Matsumoto et al., 2000).

Included Emotions

To investigate the frequency of included distinct emotions, a differentiation between the basic emotions (Ekman et al., 1969) and 'other' proposed emotions was implemented (Table 9). However, it is important to highlight that analyses were conducted on *included* distinct emotions, which does not necessarily equal *assessed* distinct emotions. This differentiation is important, as some sets may include stimuli for example audio recordings of an expressed distinct emotion, however, these stimuli were not separately assessed for emotion (e.g., *LANG-audition database*, Kanske & Kotz, 2012; or the *VENEC*, Laukka et al., 2010), or on the contrary, researchers may have originally proposed specific emotions for the validation process, however, found that certain emotions were not represented by the stimuli and were hence not kept as an emotion included in the set. Note, therefore that the number of publications including specific emotions is not congruent with the number of publications assessing stimuli using the categorical approach (see Table 6 vs. Table 9).

Of 228 publications including distinct emotions, 200 publications (87.72 %) included at least one of the *basic six*. A total of 121 publications (53.07 %) included other emotions such as boredom (e.g., Burkhardt et al., 2005), contempt (e.g., Wingenbach et al., 2016; Yan et al., 2013), doubt (Xue et al., 2006), guilt (Li et al., 2017), pain (e.g., Frowd et al., 2009), relief (Yoshie & Sauter, 2019), thoughtfulness (e.g., Schmidtmann et al., 2020), threat and shock (Bertels et al., 2009), uncertainty (Gunes & Piccardi, 2006), or untrustworthiness (Keefe et al., 2014). The use of the basic emotions is not exclusive of the use of other emotions and vice versa. This means that researchers may include or have stimuli assessed on at least one of the basic emotions, while also including/assessing further emotions (e.g., Lassalle et al., 2019; van der Schalk et al., 2011; Volkova et al., 2014; Zammuner, 2011; Zhalehpour et al., 2017).

Table 9

Number of Publications per Subfolder Including Distinct Emotions

Subfolder		*n* (%)	
	N	1-6 of Basic Six	Other
Audio	27	24 (88.89)	13 (48.15)
Faces	106	97 (91.51)	50 (47.17)
Images	10	7 (70.00)	3 (30.00)
Video	32	30 (93.75)	24 (75.00)
Words	18	14 (77.78)	8 (44.44)
Mixed	35	28 (80.00)	23 (65.71)
∑	228	200 (87.72)	121 (53.07)

Note. ∑*N* = 228 publications. 1-6 of basic six = publication includes at least one of the six basic emotions (happiness, sadness, anger, fear, disgust, surprise) after Ekman et al., 1969; other = includes any other emotions as stated by authors in publication such as: boredom, doubt, pain, relief, thinking, threat, triumph, uncertainty, untrustworthiness, etc.

Investigating the frequency of inclusion for each of the six basic emotions revealed that happiness was included most frequently, followed by sadness, anger, fear, disgust, and surprise (see Table 10). Comparison by subfolder showed that the most common emotions of interest were happiness and sadness (both 91.67 %) for audio stimuli, happiness (90.72 %) for face-stimuli, fear (85.71 %) for image stimuli, anger (90 %) for video stimuli, happiness, sadness, and anger (all 100 %) for word stimuli, and sadness (92.86 %) for the mixed stimuli. On average, each publication included 4.64 of the basic emotions, while 90 publications (45 %) included all six (see Table 11). Comparison by subfolder showed that publications presenting audio and word stimuli usually included five basic emotions (33.33 %, and 78.57 % respectively), while image stimuli referred to solely one of the basic emotions in most cases (42.86 %).

Note that the number of percentages can quickly rise, when fewer publications are included per subset (e.g., there are only 7 publications with image stimuli including basic emotions).

Table 10

Number of Publications per Subfolder Including Each of the Basic Six Emotions

Subfolder		*n* (%)					
	N	Happiness	Sadness	Anger	Fear	Disgust	Surprise
Audio	24	22 (91.67)	22 (91.67)	20 (83.33)	18 (75.00)	16 (66.67)	9 (37.50)
Faces	97	88 (90.72)	76 (78.35)	77 (79.38)	78 (80.41)	77 (79.38)	76 (78.35)
Images	7	2 (28.57)	2 (28.57)	3 (42.86)	6 (85.71)	4 (57.14)	2 (28.57)
Video	30	23 (76.67)	25 (83.33)	27 (90.00)	23 (76.67)	18 (60.00)	15 (50.00)
Words	14	14 (100)	14 (100)	14 (100)	13 (92.86)	13 (92.86)	2 (14.29)
Mixed	28	25 (89.29)	26 (92.86)	23 (82.14)	21 (75.00)	16 (57.14)	14 (50.00)
∑	200	174 (87.00)	165 (82.50)	164 (82.00)	159 (79.50)	144 (72.00)	118 (59.00)

Note. ∑*N* = 200 publications. Basic six = six basic emotions (happiness, sadness, anger, fear, disgust, surprise) after Ekman et al., 1969.

Table 11

Number of Publications per Subfolder Including Specific Number of the Basic Six Emotions

Subfolder		Number of emotions included from the basic six					
		1	2	3	4	5	6
	N	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Audio	24	-	4 (16.67)	3 (12.5)	2 (8.33)	8 (33.33)	7 (29.17)
Faces	97	9 (9.28)	3 (3.09)	9 (9.28)	7 (7.22)	11 (11.34)	58 (59.79)
Images	7	3 (42.86)	2 (28.57)	-	-	-	2 (28.57)
Video	30	3 (10.00)	3 (10.00)	2 (6.67)	5 (16.67)	5 (16.67)	12 (40.00)
Words	14	-	-	1 (7.14)	-	11 (78.57)	2 (14.29)
Mixed	28	-	2 (7.14)	7 (25.00)	3 (10.71)	7 (25.00)	9 (32.14)
∑	200	15 (7.50)	14 (7.00)	22 (11.00)	17 (8.50)	42 (21.00)	90 (45.00)

Note. ∑*N* = 200 publications. Basic six = six basic emotions (happiness, sadness, anger, fear, disgust, surprise) after Ekman et al., 1969; percentages are calculated horizontally.

Raters

Research involving humans is often highly dependent on voluntary participation. However, recruitment among the general population is difficult. Therefore, researchers often rely on student samples from colleges and universities, as studies can easily be advertised, and participation rewarded with course credit points. To give an overview of included sets that relied on student participation for stimuli rating/assessment, this information was also extracted. Participant information was available for 344 publications (94.51 %). Publications presenting stimuli sets that were not evaluated by humans were not considered. One hundred and sixty-eight publications (48.84 %) report evaluation/assessment through a student population, partly or in entirety. Assessment through crowdsourcing or a platform such as Amazon Mechanical Turk (MTurk), is reported in 13 publications (3.78 %).

Table 12

Number of Publications with Stimulus Assessment Through Student Sample and/or Crowdsourcing per Subfolder

Subfolder		*n* (%)	
	N	Student Sample	Crowdsourcing
Audio	35	18 (51.53)	1 (2.86)
Faces	101	34 (33.66)	1 (0.99)
Images	35	20 (57.14)	2 (5.71)
Video	42	21 (50.00)	4 (9.52)
Words	88	59 (67.05)	1 (1.14)
Mixed	43	16 (37.21)	4 (9.30)
∑	344	168 (48.84)	13 (3.78)

Note. $\sum N = 344$ publications.

Additionally, reliability of a stimulus rating increases with increasing number of raters. The total number of raters throughout the study is often reported. However, these are not easily comparable across studies, as some studies report rating procedures in which each stimulus has been rated by each participant in one 'flow' (e.g., valence arousal, and dominance rating for each stimulus) while stimuli in other studies have been rated in a series of sub-experiments (e.g., for different dimensions), each by a few participants from the entire participant population. Hence, comparison across sets based on the total number of raters should always be consulted with caution. Therefore, in the database, the total number of participants reflecting *individual* participants, as well as the number of raters per stimulus were coded (note that in the latter, a stimulus may also have been rated by the same participant on multiple occasions, increasing the number of ratings per stimulus). Not all publications report numbers in such detail. However, analysis of all sets including assessed stimuli and reporting this information, showed that each stimulus in the audio subfolder was rated between 10 and 945 times, face-stimuli were rated between 1 and 1428 times, image stimuli between 16 and 264 times, video stimuli 1 to 180 times, word stimuli 10 to 960 times, and stimuli from the mixed folder were rated 1 to 70 times.

Due to the difficulty of comparing reported numbers across stimuli sets, descriptive statistics cannot be calculated and reported here. Researchers shall therefore consider the original publication for additional information when comparing the total number of raters or number of ratings per stimulus across sets.

Country of Research

To compare possible differences in the focus of research regarding types of stimuli, country of research was also investigated and a broader classification of research location per

continent listed (Table 13). In five publications, researchers were affiliated to universities from different continents, and information concerning where the research was conducted was not accessible. Therefore, percentage calculations are based on the remaining 359 publications. Cases where the country of research was not clear, however, in which researchers' affiliation was on the same continent as for instance The Netherlands and the United Kingdom (Valstar & Pantic, 2010), were not excluded from analysis.

Overall, research leading to the presentation of ES sets was mainly conducted in Europe (56.82 %), followed by North America (24.23 %), Asia (10.58 %), Australia and Oceania (2.79 %), and South America (1.95 %). A few studies (3.62 %) combine research conducted in multiple continents (e.g., Baveye et al., 2013; McCurrie et al., 2018) and were therefore allocated to the *Multiple* column (see Table 12). Separated by subfolders, however, an accentuated focus on types of ES per continent became visible: Publications from research conducted in Asia, as well as Australia and Oceania, North America, and South America, mainly present face-stimuli sets, whereas most publications of research conducted in Europe present word stimuli. Information concerning the exact country of research can be found in the *Supplementary Material (SM)* (*Study 1*, column *"country of study"*) for each individual subfolder.

Table 13

Number and Percentage of the Publication's Research Location per Continent and Subfolder

Subfolder		*n* (%)					
	N	Asia	Australia and Oceania	Europe	North America	South America	Multiple
Audio	34	2 (5.88)	1 (2.94)	17 (50.00)	12 (35.29)	-	2 (5.88)
Faces	115	25 (21.74)	4 (3.48)	46 (40.00)	33 (28.70)	5 (4.35)	2 (1.74)
Images	35	2 (5.71)	1 (2.86)	19 (54.29)	7 (20.00)	2 (5.71)	4 (11.43)
Mixed	45	2 (4.44)	1 (2.22)	33 (73.33)	8 (17.78)	-	1 (2.22)
Video	42	5 (11.90)	2 (4.76)	24 (57.14)	7 (16.67)	-	4 (9.52)
Words	88	2 (2.27)	1 (1.14)	65 (73.86)	20 (22.73)	-	-
∑	*359*	38 (10.58)	10 (2.79)	204 (56.82)	87 (24.23)	7 (1.95)	13 (3.62)

Note. ∑*N* = 359. Multiple = research conducted on more than one continent; percentages are calculated horizontally.

Independent from subfolders, the majority of research conducted in Asia has been conducted in China; in Australia and Oceania in Australia; in Europe in Germany; in North America in the United States of America; and in South America in Brazil.

In conclusion, the analysis of the individual subfolders revealed differences concerning number of publications, as well as continent of research, depending on the type of stimuli. Most studies presenting a set were conducted in Europe as well as North America. Moreover, the

faces subfolder was the only folder displaying a suddenly increasing interest in emotional face-stimuli during the past decade in Asia. Here, the number of publications (n = 25) closely follows that of Europe (n = 46) and the USA (n = 33).

Stage Three: Summarization and Visualization of Results

A spreadsheet containing all extracted and coded information is available in the *SM* (*Study 1*). The created searchable online version of the database is publicly available. At the time of writing, both contain information regarding 364 available emotional stimuli sets.

Navigation of the Database

In addition to listing all ES sets with their key characteristics as presented in the Excel sheet version, the online version of the KAPODI database includes a search tool allowing the selection of stimuli sets according to specific criteria. Based on the stimuli type (e.g., audio, faces, images, video, words, mixed), each set is listed in one of the six subfolders (see Figure 2, [a]). Further, the selection between gallery or grid view presents the stimuli sets in a list (grid view; Figure 2, [b]), or as individual cards that can be selected for more detailed information (gallery view; Figure 3, [b]). Within each subfolder (e.g., audio stimuli), (see Figure 3, [a]) the filter tab [c] allows the selection of stimuli sets based on the specific filter(s) (e.g., language, [d]) and a refined search (e.g., English and Japanese, [e]). All extracted key characteristics mentioned in Table 1 can be set as a filter. Moreover, users may search sets by entering key words in the search bar. Only sets matching the search criteria are displayed to the viewer.

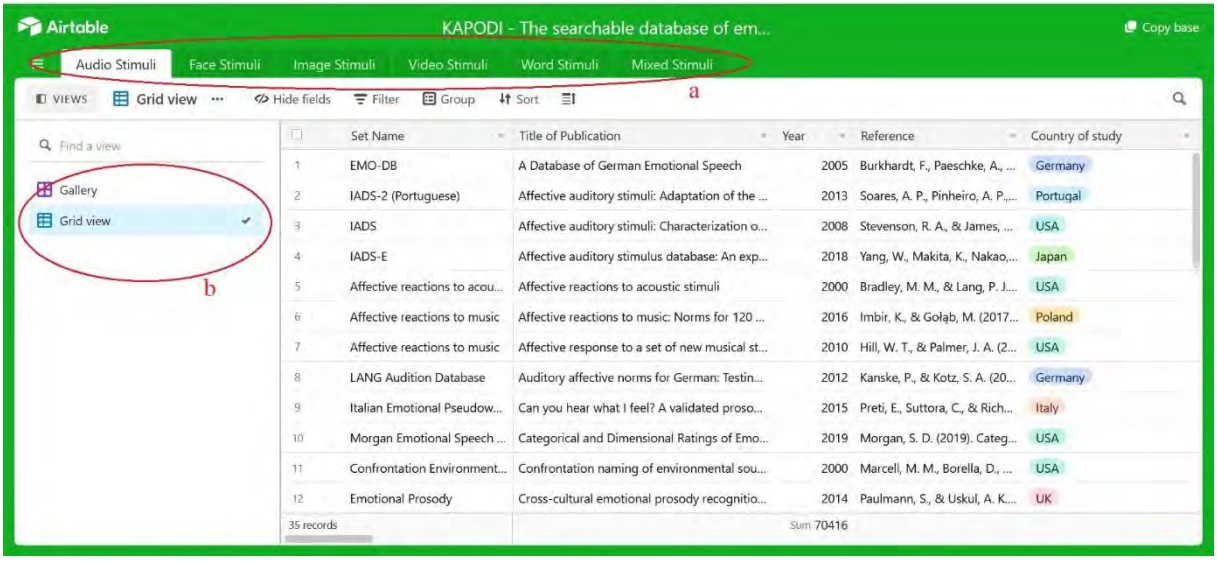


Figure 2. Exemplary view of the KAPODI searchable database I. *Note.* Stimuli sets are separated by type of stimuli (a) and can be viewed in a gallery or grid view (b).

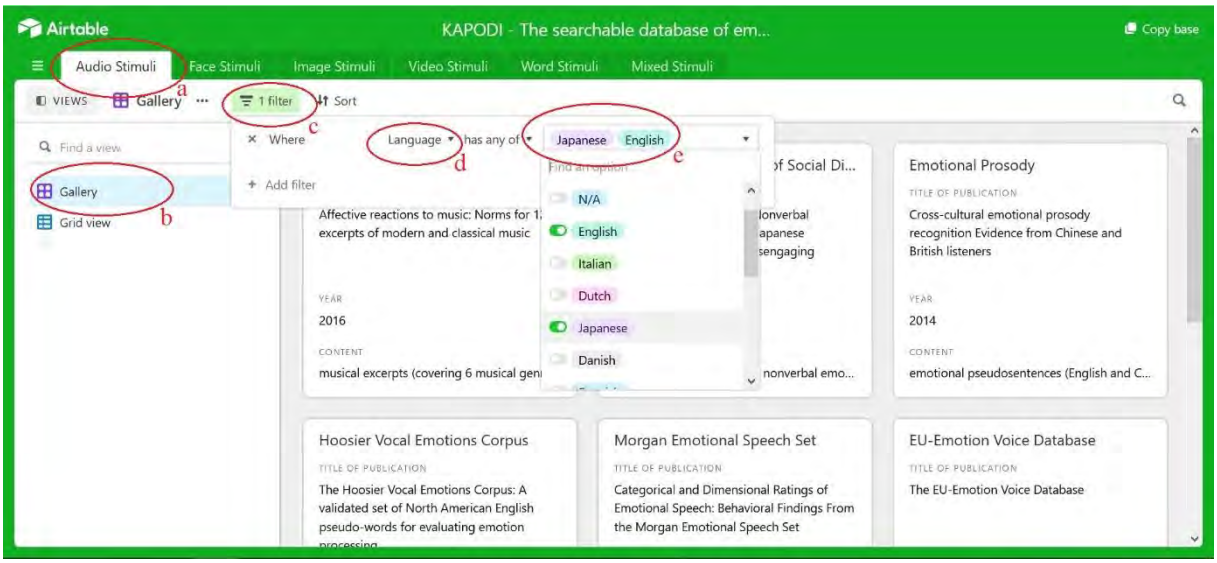


Figure 3. Exemplary view of the KAPODI searchable database II. *Note.* The audio subfolder is selected (a) and view set to gallery view (b); the filter (c) allows the selection of key set characteristics (d) (e.g., language) with a refined selection among all available options (e) of the set filter.

Discussion

The searchable KAPODI database was constructed in response to an increasing number of publications providing ES sets to the research community. The numerous publications reflect a growing demand for tailored stimuli as well as a growing interest in the field of emotion research. In the present work, the first comprehensive systematic review of emotional stimuli sets was conducted, and key set characteristics were coded to allow for comparison between sets and aid researchers in choosing appropriate sets for their research. The resulting KAPODI database contains the largest list of available stimuli sets to this date and is therefore a useful contribution to research on emotion and beyond. In the following section, the database will be discussed regarding its use and its limitations. Finally, the creation and publication of emotional stimuli sets will be discussed with recommendations regarding how these should be reported.

Assessment Approach and Emotion Theory

With emotions being the key characteristic of stimuli, the assessment approach, as well as the theoretical framework used for emotion selection, were aspects of great interest in the present study. While categorical and dimensional approach were applied approximately equally as often overall, it was interesting to see that certain types of stimuli focus more on one approach than another (e.g., dimensional for word stimuli; categorical for face-stimuli). There are various factors that may influence the choice of assessment approach (e.g., time or financial restrictions, as well as reasonability of tasks for participants). The findings that audio and face-stimuli are mainly assessed through the categorical approach, may reflect that these stimuli contain human emotion *expressions* that can easily be categorised. Word stimuli that are often context-

dependent, and image stimuli that contain broader non-human content (e.g., landscapes, food, or animals) do not express a specific emotion as such. Hence, with no *directly* expressed emotion, researchers may instinctively be more inclined in assessing how these stimuli are *perceived* (vs. an emotion they may express), which automatically directs the assessment towards a dimensional approach.

Generally, normative rating data regarding *perception* may be more valuable for stimuli created for emotion induction, while stimuli intended for emotion recognition (e.g., in therapy or for automatic emotion detection), may need normative rating data regarding *expression* of emotions in stimuli. However, one fifth of the studies using both approaches (dimensional and categorical), and coming mainly from the past decade, certainly reflect a growing interest within emotion research in seeking answers concerning the congruence, compatibility, and demarcation of these two approaches. Future research could benefit from existing sets that have been assessed using one approach, to assess them using another approach. The results could be analysed to find the assessment approach leading to more valuable normative data, potentially influencing validation procedures for stimuli sets created in the future.

Two hundred twenty-eight publications include distinct emotions. Among these, emotion selection based on the *basic emotion theory* (Ekman, 1992), was chosen more often (n = 200) than an extension thereof (e.g., by adding or selecting other emotions; n = 121). However, it would be misleading to conclude that the basic emotion theory is the preferred foundation when selecting emotions for studies, as it is impossible to disentangle one from the other. That is, ever since the official establishment of the basic six emotions, every published study that includes any of these emotions (e.g., *happiness*), will automatically be classified as a study including a basic emotion.

Using the Database

Researching the availability of different ES sets is a time-consuming task and it often leads researchers to resort to well-known and widely used stimuli sets such as the *IAPS* (Lang et al., 2008) or *ANEW* (Bradley & Lang, 1999a), despite a wide range of other stimuli sets being available. This creates a dilemma, comparable to an iceberg: Sets well known within the research community – analogous to the visible tip of the iceberg – are widely used, therefore attract more attention, and are hence also selected more frequently again by researchers looking for stimuli for their own studies. A rapidly growing number of created tailored ES – the hidden part of the iceberg – remain unseen and unused, often overshadowed by the widely used sets.

At the time of publication, the KAPODI database comprises 364 publications from 1963 to 2020 that cover various types of stimuli such as audio, image, video, and word stimuli, or a

combination thereof (e.g., audio-visual). Six subfolders were created within the database for easier comparison of similar stimuli and to facilitate stimuli search for researchers in the future. Over 25 key characteristics have been coded for find appropriate stimuli (e.g., the characteristic rating scale with the criteria SAM-scale, Likert scale, visual analogue scale, forced-choice, other). The database allows researchers to see whether a study has created a new set or has validated stimuli from a pre-existing set in a new population (e.g., different age group, different ethnicity, or different country). Set characteristics of interest can easily be compared, facilitating choice of set, and accelerating the research process.

Information coded about the included emotions may for instance allow researchers to select stimuli of a distinct emotion, or select stimuli rated as neutral, for comparison. Researchers may search for stimuli sets based on applied rating scales (e.g., SAM-scale, Likert scale, visual analogue scale) or length of used scales, which may be important information for choosing new stimuli for replication studies. Researchers can also select sets that include a minimum number of stimuli, a minimum number of included models (e.g., within face-stimuli), or choose stimuli sets that include a specific type of content (e.g., images of food, or fear-inducing images). Moreover, it is possible to search for stimuli of a specific language or sets that include an additional type of data (e.g., physiological recordings).

The coded number of raters per stimulus allows researchers to search for stimuli assessed by a minimum number of raters, which is relevant for discerning the reliability of the ratings. Additional information regarding the rating population allows selection of stimuli based on the type of assessors (e.g., sets normed with student populations or via crowdsourcing). For certain types of stimuli, further information is coded to allow researchers to find exactly what they are looking for. For example, a researcher may wish to find audio stimuli which include natural (non-acted) expressions of happiness. The selected filters would be as follows: within the audio stimuli subfolder, expression authenticity is natural (Filter 1), and all included emotions, has any of, happiness (Filter 2). The only dataset currently included in the database and meeting these search criteria is the OxVoc (Oxford Vocal Sounds Database; Parsons et al., 2014).

Finally, further information is provided regarding the context of creation for each ES (e.g., researchers' affiliation and country of study), as this may provide valuable information for researchers interested in the development of different types of stimuli from a geographical perspective.

In summary, the database provides the researcher with more flexibility in selecting an appropriate stimulus set and provides a systematic basis for going beyond classic ES sets (e.g.,

ANEW). This central database facilitates access to, and eases comparison between stimuli and/or sets for a wide range of applications and for researchers from a wide range of disciplines.

Strengths and Limitations

Despite its benefits for research, the database described is not without limitations. These, along with a few examples, are outlined in the following section.

Similar to the Pictures of Facial Affect presented by Ekman (POFA, Ekman, 1976), some researchers created stimuli for use in their study without consideration of the stimuli being used in further experiments by different researchers. Without doubt, each of the sets included in the KAPODI database has distinct strengths and was created for a specific aim, filling the gap in the availability of standardized stimuli. Nevertheless, depending on the initial research aim, different key characteristics were regarded important by the authors and therefore reported along with the stimuli, while others were not. This is a difficulty that is reflected through unclear or incomplete available information when comparing all included sets.

The root of this difficulty could be due to incompatible theories of emotion. For instance, there is no agreement on a unitary definition of emotion (Izard, 2007). Therefore, the current database employed categories that were thought to best represent the various facets of emotional information and allow comparison between sets without subscribing to a specific overall theory of emotion categorization. In summary, criteria from different sets were coded according to the best understanding of set content and study procedure, while also keeping in mind the usability of the created database (e.g., set search by key words) for interested researchers.

To give an example, some scientists accept only the *basic six* (Ekman et al., 1969) as emotions and reject any others. However, in the present work, the different terms named *emotions* by fellow researchers, were not judged but rather listed as proposed in the source article. As an illustration, in two studies, *smile* was mentioned as an expression (McDuff et al., 2019), or images have been classified/labelled according to smile (Samaria & Harter, 1994). These two studies were treated as exceptions and smile was listed as an emotion, so that these sets are also detected when searching through the database.

Comparably, other terms may have differed in some publications, though these terms were not modified while coding: despite most publications naming happiness as one of the basic emotions, in a few cases, happiness was replaced by *joy* (e.g., Costantini et al., 2014) or *amusement* (e.g., Yan et al., 2013). Though this seemed to depend on translation from other languages (e.g., French or German: Bertels et al., 2014; Hewig et al., 2005), it was necessary to find a consensus and create categories to facilitate the search within the database without changing the meaning of terms used in the original study. In this example regarding

happiness/joy/amusement, the original authors' decision was accepted during the coding process. This means that in a few cases, joy may be listed as one of the basic emotions, while in other cases it was listed as an emotion that differs from happiness (e.g., Soleymani et al., 2012). The same applies to amusement.

Another example of the variability in usage of terminology is that of the three dimensions *valence*, *arousal*, and *dominance*, on which stimuli are assessed in many publications. In a study conducted by Marcell et al. (2000), *pleasantness* was assessed. This was coded as the equivalent of valence. In another publication, *potency* was assessed (Kleinsmith et al., 2011) with authors mentioning that it is also referred to as dominance. However, given that the authors decided to use the term potency rather than dominance, suggests that the term was chosen for a specific reason, and was therefore not coded as dominance. Additionally, Schmidtke et al. (2014) suggest differentiating between dominance and potency, as the latter 'mainly differs in its independence from the raters' perspective' (p. 1110).

Recommendations for the Creation and Publication of Stimuli Sets

One of the main difficulties during the coding process was missing or incoherently reported information. Despite the best effort in contacting authors, not all inconsistencies could be solved. During this procedure, qualitative differences between publications became apparent. An example that arose in multiple publications, is the reporting of age and gender distribution of participants: While authors may initially mention the exact distribution (e.g., number of female and male participants; mean age), this number was not always adjusted after participant exclusion. Although in the present study this was noted while coding, it means that some data reported in the database can solely be regarded as an estimate due to lacking precision in the report of the source publication.

Another example and important factor especially concerning image stimuli, is the stimulus resolution: While resolution of provided stimuli may be reported in the publication, in some cases stimuli were not presented to raters in this same resolution during the validation procedure. Stimuli from the *Dartmouth Database of Children's Faces* (Dalrymple et al., 2013) for instance, are provided in a resolution of 900 x 900 pixels (300 dpi), while they were validated in a version cropped to 300 x 300 pixels (100 dpi). By implication, this means that the provided stimuli were not *really* validated, or that the provided normative rating data may not match with the provided stimuli.

Similarly, the fact that almost half of the included publications mention validation and/or assessment of stimuli through a student population should not be disregarded. In fact,

this proportion can be considered an underestimation due to lacking or unspecified information in a few cases. As age and education influence perception, there may be differences between the general and a student population (Henry, 2008). Consequently, stimuli ratings may not be valid for the general population if assessed through students.

To offer a large applicability of good quality ES, researchers creating and presenting stimuli in the future should generally consider three aspects: high-quality stimuli, good validation procedure, and clear reporting. That is, researchers should *1)* aim to create high-quality stimuli (e.g., high resolution, or high number of frames per second for video recordings); *2)* validate the created stimuli by a large sample size of a well-justified selection of assessors (e.g., assessors with the same main characteristics as potential target groups) which is especially important for stimuli created for specific target groups such as individuals with alcohol or food addiction; and *3)* include and clearly communicate detailed technical information (e.g., colour spectrum or luminance) regarding each stimulus.

The issue of clear reporting is particularly important and will therefore be expanded on. Especially among video stimuli, information regarding sex and age of models is frequently missing. Similarly, stimuli sets within the *mixed* folder frequently did not contain information regarding the included ethnicities in their video recordings. Furthermore, some sets did not include information regarding language or colour of stimuli. Though individual sets may have been created for a specific research aim and therefore may have suited a specific survey design, missing information may limit the appropriate use of the stimulus set. Moreover, the absence of detailed information – especially regarding the stimuli and validation characteristics – complicate the interpretation of study effects.

More recently published stimuli sets are often good examples of comprehensive reporting and high-quality stimuli, reflecting an increasing understanding of the need for relevant information to be included, but also of improving technological abilities (e.g., *CAFE*-set, LoBue & Thrasher, 2015; *EU-Emotion Voice Database*, Lassalle et al., 2019; *Food-Cal*, Shankland et al., 2019).

A central aim regarding future research conducted in relation to ES sets should be to improve the uniformity in reporting the characteristics of the set. Hence, it is suggested that researchers developing stimuli sets in the future should include information regarding all the key characteristics established through this current systematic review. For guidance, researchers may use the KAPODI submission form (see the link in the *Final Discussion* section of this chapter) as a checklist when reporting information of their stimuli set. It is further suggested that established terminology is used (e.g., the dimensions valence, arousal,

dominance), unless authors specifically wish to differentiate and justify their own language and terms.

Finally, researchers should ensure that their ES sets are made available freely and openly to other researchers, which will substantially contribute to transparency and reproducibility of research procedures (see Munafò et al., 2017).

With the central aim of supporting the efficiency of scientific research and knowledge accumulation, only stimuli sets that are publicly available/ freely available upon request were included. The information regarding availability of the set was taken from the original source. Some publications include an internet link, directly leading to the freely accessible set or to a compliance form for researchers. Others provide an e-mail address through which sets and data can be requested to the author(s) directly. Nevertheless, links and/or e-mail addresses may have changed, or sets may no longer be available even if indicated so in the publication. It can therefore not be guaranteed that sets are truly available at present, even if stated so in the original source. Rather than relying on requests for access via email, in the future, researchers should upload their created sets to a website or repository granting availability and easing access to stimuli to colleagues. An automatic validation system (for instance through a form requiring assurance regarding the academic purpose of accessing the set) could restrict access to researchers only. This further ensures access to the stimuli to remain the same, even if the set creator has changed or left their institution.

Final Discussion

Using a systematic review methodology, the current study aimed to identify as many available ES sets as possible. The resulting searchable database, which can be found online, currently contains 364 different stimuli sets. It is available to the research community and all included stimuli sets are freely available or available upon request. By making all extracted and listed set key characteristics available in an Excel sheet as well as through the website Airtable.com, permanent availability is moreover ensured. Researchers who wish to add their new stimuli set to the searchable KAPODI database can fill out the corresponding set form. The submitted set will be verified for suitability by one for the authors. Once approved, that is, (*a*) the set complies with the requirements of being freely available to the research community, as well as including stimuli and their accompanying normative rating data, and (*b*) that the creation procedure of the stimulus set has been published in a scientific journal, the submitted information will be uploaded to the database. The long-term aim is to maintain the searchable online version of the database updated, continuously extending the database content.

Chapter Three – Investigating the Prevalence of Reliable Emotional Stimuli in a Typical Psychology Study Sample of Adults

The created KAPODI database contains the largest collection of freely available stimuli sets to this date and therefore represents a useful resource for researchers aiming to find suitable stimuli for their study. Nevertheless, various factors such as participants' age (Isaacowitz et al., 2007), gender (Lithari et al., 2010; Nater et al., 2006), ethnicity (DeBusk & Austin, 2011), or social and cultural background (Boiger et al., 2018; Matsumoto et al., 2008) have been shown to influence emotion perception. In consequence this means that if emotion perception is affected by certain factors, and studies including ES are conducted without controlling for these factors, researchers may risk distorting their study results by applying unreliable stimuli.

Therefore, the second study within the framework of the current research project on ES aimed to investigate the reliability of a range of emotional stimuli in a typical psychology study sample, and moreover investigate stimulus related factors that may influence the stimulus' reliability. The created the KAPODI database from the previous work (see *Chapter Two*) was hereby used as a resource of stimuli sets for the selection of stimuli in the present study.

The two main research questions were formulated as follows:

Q2.1: Is emotional stimulus reliability determined by factors associated with assessment of the stimuli [such as dimension *(e.g., valence, arousal, dominance),* dimension category *and* SD category *(e.g., high, medium, low),* stimulus type *(e.g., images and words), or* gender *(e.g., female, male)]?*

and

Q2.2: What is the prevalence of reliable emotional stimuli?

In this context, emotional image and word stimuli published in emotion research were reassessed, to *(1)* compare assessed data to normative rating data, *(2)* identify stimuli that are rated reliably across different populations to investigate the prevalence, that is proportion of reliable stimuli among investigated stimuli, and *(3)* investigate factors associated with the stimuli that may determine reliability. For this latter aspect, several factors such as stimulus type (e.g., images and words), assessed dimensions (e.g., valence, arousal, dominance), as well as normative rating and standard deviation categories (e.g., high, medium, low) were investigated with respect to gender.

The results aim to help researchers identify suitable stimuli more effectively with respect to all provided information of set characteristic when using the KAPODI database.

Introduction

As described in the previous chapter, ES play a central role within the field of emotion research and numerous sets have been created throughout the past decades. Individual stimuli as well as entire sets have been used in research, therapeutic contexts, for machine learning, and other areas.

Creating ES and assessing their emotional value prior to conducting an experiment can be a very time-consuming task. Therefore, the availability of normatively rated stimuli within the research community is of great advantage. Since the beginning of research on subjective perception of ES, various sets have been presented and proposed to the research community. Among the large number of freely available sets, a few are very well-known and have been widely cited and used for studies. Examples are the *International Affective Picture System* (IAPS; Lang et al., 1997), the *NimStim Set of Facial Expressions* (NimStim; Tottenham et al., 2009), *Affective Norms for English Words* (ANEW; Bradley & Lang, 1999a), *International Affective Digitized Sound* (IADS; Bradley & Lang, 1999b), the *Oxford Vocal Sounds Database* (OxVoc; Parson et al., 2014), EMDB (Carvalho et al., 2012), or also the *Karolinska Direction Emotional Faces* (KDEF; Bartlett et al., 1999) and the *FACES* database (Ebner, Riediger & Lindenberger, 2010). The normative rating data accompanying the set provides information regarding emotional characteristics of each stimulus in relation to assessed dimensions (e.g., distinct emotions such as happiness, or dimensions such as valence and arousal).

Examples of Factors That Influence Perception of Emotional Stimuli

When selecting suitable stimuli for their own studies, researchers may thus rely on reliability and validity of the data from the norming sample. Hence, they do not verify the effect for their own participant sample prior to study conduction. This is, however, inevitably associated with risks, as study results have shown that various factors may influence stimuli perception. These factors may be assessor-related, stimulus-related, or related to study construction itself:

Assessor-Related Factors

Assessor-related factors are factors that are inherent to the participant who is rating the stimulus to provide the normative data. Many studies mirror the desire of researchers to investigate generalizability of ES across different populations by reassessing sets in part or in their entirety by specific participant (sub)-groups. In that regard, a variety of studies have displayed the relation of assessor-related factors (e.g., gender, ethnicity, cultural background, age, language) and the perception of emotion:

Gender

The most frequently investigated aspect that researchers highlight when reporting normative rating data is the possible difference between *genders* (see e.g., Garrido et al., 2017; Weierich et al., 2019). Research has demonstrated that females respond with greater event related potentials (Lithari et al., 2010) as well as hightened physiological reactions (e.g., finger temperature, skind conductance; Nater et al., 2006) to unpleasant and high arousing stimuli compared to males. Moreover, gender differences in the perception of word stimuli were shown for taboo words, sexual terms as well as words denoting weapons (Janschewitz, 2008; Warriner et al., 2013). Therefore today, in most cases, assessed data are reported combined as well as separately for male and female participants. Examples for such gender-specific reporting can be found regarding audio stimuli (e.g., Soares et al., 2013), face stimuli (e.g., Garrido et al., 2017), image stimuli (e.g., Magalhães et al., 2018), video stimuli (e.g., Petridis et al., 2013), and word stimuli (e.g., Sianipar et al., 2016).

Ethnicity and Cultural Background

Research investigating the influence of *ethnicity* on emotion perception (e.g., DeBusk & Austin, 2011; Meissner & Brigham, 2001; Shapiro & Penrod, 1986) has displayed that emotion expressions are more correctly identified within one's own ethnic group (own-ethnicity bias). Therefore, especially face-stimuli sets have been developed with focus on a variety of ethnicities (e.g., Zhang et al., 2013; Gur et al., 2002), or solely one ethnicity (e.g., Deng et al., 2017; Vaiman et al., 2017). A comparison of image perception across ethnic groups revealed higher arousal ratings by the Brazilian population compared to ratings collected in a U.S. sample (e.g., Ribeiro et al., 2005). A similar study conducted by Okon-Singer et al., (2011) also displayed that image stimuli were rated as less positive and less negative by young Israeli compared to U.S. participants, and Israeli women perceived images as more arousing compared to American women. Nevertheless, latter findings may be due to an effect of *cultural background* (e.g., exposure to violence) rather than solely ethnicity. In fact, research suggests that people from different cultures may appraise the same event in very different ways depending on their own culture's system of meaning (Scherer & Brosh, 2009). An effect of cultural background onto the experience of emotion (e.g., anger and shame), can moreover be found in the work conducted by Boiger and colleagues (2018): In their study using situation vignettes (e.g., text stimuli), authors found that western cultures (e.g., U.S. and Belgian participats) tend to experience an anger type resulting in blaming close others, while Japanese experience an anger type resulting in blaming distant others. Experienced shame was more easily relativized in U.S. and Belgian individuals in relation to public exposure compared to

when personal flaws were pointed out, which was the opposite for Japanese participants who were feeling less shame in relation to personal flaws, however more shame when looking bad in public. Similarly, Davis and colleages (2012) showed that Chinese participants report less intense negative emotion when viewing negative image stimuli, compared to American participants. Moreover, a comparison between cultures sharing the same language (e.g., Portuguese and Guinea-Bissauan) was conducted by Cosme and colleagues (2021). In their study, authors used nonverbal vocalizations (e.g., audio stimuli) and displayed an ease of emotion recognition of sounds from the own (vs. foreign) culture.

Age

Investigations regarding the relation between *age* and emotion recognition in lexical stimuli and stimuli of facial emotion expressions has shown that older adults are less accurate than younger adults in emotion recognition for both types of stimuli (Issacowitz et al., 2007). More specifically, in the conducted study authors found that differences in recognition accuracy between both age groups were present for five of the basic six emotions (all except fear) for lexical stimuli, while recognition accuracy between age groups differed for anger, disgust, fear and happiness for stimuli of facial expressions. A few examples of sets that were assessed separately for specific age groups highlight the need for a distinction based on the participant's age: Comparisons of stimuli perception among different age groups are for instance available for the *Berlin Affective Word Lis*t (BAWL; Võ et al., 2006) (e.g., Võ, et al., 2006; Sylvester et al., 2016), or also the *Besançon Affective Picture Set* (BAPS) assessed by adolescents in the *BAPS-Ado* (Szymanska et al., 2015) and by adults in the *BAPS-Adult* (Szymanska et al., 2019).

Language

Another assessor-related factor that may influence emotion perception is *language*: Perunovic and colleagues (2007) for instance investigated emotion-related language shifts in multilingualism, displaying a changing pattern of emotion experience dependening on the language that was spoken (e.g., English-Chinese speakers reporting a Western emotional pattern after speaking English, however Asian emotional patterns after speaking Chinese). In their review Chen and colleagues (2012; p. 370) summarize: "Hearing or speaking a particular language can influence a speaker's emotional response, and a speaker's affective state may also influence his or her choice of language.". Nevertheless, translations of word stimuli from the *ANEW* (Bradley & Lang, 1999a) into Italian or Portuguese (see Montefinese et al., 2014; Soares et al., 2012) have shown high correlations of stimulus perception between the original and translated language version, suggesting a same perception across languages in relation to word stimuli.

Cognitive Ability

Finally, *cognitive ability* may influence emotion perception: For instance, evidence suggests that for individuals diagnosed with autism, responsiveness to others' emotions increases with cognitive functioning (Dissanayake et al. 1996). Furthermore, significantly higher cognitive empathy has been found in psychology students compared to business students (Litten et al., 2020). As the systematic review (see *Chapter Two*) has shown, validation of stimuli is often conducted at universities with university student participants. This means that for many stimuli sets normative data are based on a participant sample of young psychology students (see *Table 12*). Whether psychology students hence perceive emotional stimuli differently than the general population, however, still needs additional investigation.

Factors Related to Study Construction and Stimulus-Related Factors

In addition to factors related to the assessor, research has also shown that factors inherent to study construction affect (emotional) perception of stimuli and hence their rating data. An example is the use of different assessment scales: Hasson and Arnetz (2005) found that participants respond with an end aversion bias (avoiding the extreme ends of the scale) on multi-item Likert scales, compared to one-item VAS scales measuring the same construct. A similar result had been shown by Brunier and Graydon (1996) in a previous study, where the variance shared between the two scales (single-item VAS and multi-item Likert) was only 64 %. Bolton and Wilkinson (1998) who compared the use of VAS, Likert-type scale, and verbal rating scale to report about the levels of pain, conclude the Likert-type scale being most responsive of the measures. Another factor that has been shown to influence perception of emotional stimuli is contextualization. That is, assessment data of stimuli may vary depending on the presentation context. Cahill (1975) as well as Wolff and Wogalter, (1998) conducted studies displaying that symbol stimuli are more accurately identified in context (vs. no-context). Moreover, stimuli displayed on a large screen (vs. medium and small screen) have been found to result in greater skin conductance indicating higher arousal (Codispoti & De Cesarei, 2007; Reeves et al., 1999). Additionally, stimuli that are presented to participants for a longer duration are liked more and disliked less than stimuli presented for a shorter duration (Reber et al., 1998). For example, Marin and Leder (2016) showed increasing valence and arousal ratings of stimuli presented for a duration of 5 seconds compared to 1 second. Particularly these two research examples suggest that effects caused by differences in study construction may vary across types of stimuli: while auditory stimuli as well as video clips are usually displayed over a set continuous timespan, in comparison, the display duration of image or word stimuli is manipulable to a greater degree (e.g., from a fraction of a second to multiple seconds or longer).

Moreover, a few examples raise the concern that the content captured *in* a stimulus itself may cause a change in emotion perception due to changes in fashion (e.g., hair, fashion, or makeup style), technical development (e.g., cars, machines, or computers and cell phones), historical changes (e.g., construction, or demolition of famous buildings such as the World Trade Center in New York), as well as topics of which general acceptance has changed over time (e.g., tattoos or homosexual relationships), depending on the disparity in time between assessment of normative data and use of stimuli in a study. Not limited to solely visual stimuli such as image or video clips, this effect may also apply to audio, word, and text stimuli. Examples are the use of archaic words that can render stimuli outdated, as well as the creation of new words that may cause an increasingly differentiated understanding of former existing words (e.g., *hangry* first used by the millennial generation and meaning becoming angry because of feeling hungry, or *the cloud* which has become a metaphor for the internet enabling anyone with an online connection to access data). The word *Corona* may serve as a recent example, known as a Mexican beer brand, that may however, elicit different emotions since the outbreak of the Coronavirus-disease pandemic in 2019.

Although examples exist for different types of stimuli (e.g., images, words, video clips), to date, it remains unclear whether these changes in emotion perception are equally pronounced or differ between stimuli types. It could be argued that between word and image stimuli, the former remain relatively abstract, while the latter displays distinct content. That is, while the perceiver has to transform a word into a pictorial representation which is distinct and inherent to every perceiver, an image depicts tangible content and therefore leaves less room to own imagination. Following this line of argument, individuals viewing an image may tend to compare its distinct content (e.g., image of house built in classical architecture) with all previously experienced encounters of similar content (e.g., all previously seen images or real-life experiences of houses) and therefore make a *comparative* rating. This, however, may render the stimulus more prone to changes in perception in relation to the current internal (assessor-related) and external (e.g., familiar environment; norms of society) state, and hence less reliable. The nonspecific character of word stimuli in contrast (e.g., the word *house*), will trigger more general visual representations within individuals depending on the own background. That is, this same word (*house*) will trigger a different image in the inner eye for individuals who for instance grew up in a big city in Europe (e.g., Paris), compared to a small village in Mexico (e.g., Las Coloradas). Nevertheless, this inner-eye-representation is constantly adjusted or "updated" based on experiences (e.g., the word *computer* is more likely to trigger the image of a modern computer as it is used today, than an image of a computer as it was in the 1990s). In

consequence, unless affected by major changes in the language within society (see examples given above), these stimuli will remain relatively stable regarding their rating data, and thus reliable.

The Importance of Reliable Emotional Stimuli

Differentiation between, as well as re-assessment of stimuli for specific (sub-)groups (e.g., gender, age, ethnicity) has allowed researchers to select stimuli more precisely based on their own participant sample characteristics. A few studies have highlighted the validation or reassessment of a set through a specific participant sample, emphasizing its usability for a particular subgroup. Examples are the *Military Affective Picture Set* (MAPS) (Goodman et al., 2016) with stimuli assessed for 5 subgroups (female civilians, non-combat-exposed female military, male civilian, non-combat-exposed male military, and combat exposed male military), an IAPS-subset assessed by individuals with borderline personality disorder (Eddie & Bates, 2017), or also the *The Cambridge Mindreading* (CAM) *Face-Voice Battery* (Golan et al., 2006a), assessed by individuals with Asperger syndrome.

Nevertheless, considering the many available ES sets (see *Chapter Two*), validation or reassessment of ES perception of *all* published sets for *all* possibly existing (sub-)groups remains impossible. While, on the one hand, comparison of ES perception among specific (sub-)groups may be of interest for certain research questions, other experimental studies using ES for instance for emotion induction, are often conducted with a 'typical' psychology study sample aiming to represent the general (healthy) population (e.g., Van Dyck et al., 2014; Vuilleumier et al., 2001; Williams et al., 2005). Such a study could the investigation of emotion on recall memory as conducted by Kamp et al., (2015) who selected word stimuli from the *ANEW* database (Bradley & Lang, 1999a) to compare the effect of emotion valence and categorized the stimuli as positive, negative, and neutral, based on the normative rating data.

Similar examples can be found in the study conducted by Kensinger et al., (2007) who used image stimuli to investigate the effect of emotion on memory, or also Noulhiane et al., (2007) who investigated the effect of emotion onto the perception of time by using sound stimuli selected from the *IADS* (Bradley & Lang, 1999b). This procedure of selecting stimuli based on normative data is commonly conducted, however, researchers rarely verify the extent to which ratings collected in the original work (e.g., by Bradly and Lang in 1999) reflect emotion perception of their participants (e.g., in 2015 for the study conducted by Kamp and colleagues, mentioned above). To investigate the effect of different dimension categories (e.g., low, medium, high), researchers typically select stimuli that are matched for a dimension category (e.g., high) of another dimension as the one under investigation. For instance, selecting

stimuli matched on arousal when assessing differences between stimuli that vary in valence. This procedure allows researchers to control for the influence of one dimension (e.g., here: arousal). Study examples where stimuli were selected matching on one dimension can be found for instance for arousal (e.g., Balzus et al., 2021; Bireta et al., 2021; Kensinger et al., 2006; Tse et al., 2009), or valence (e.g., Vermeulen et al., 2019).

Reliability of normative data may, however, not always be granted due to assessor-related or study construction/ stimulus-related factors affecting stimulus perception (see above). Hence, if researchers rely on unreliable stimuli, the study results and conclusions may be undermined. This, moreover, threatens generalizability of the study results. Consequentially, the use of reliable stimuli is vital to the validity of research. As such, researchers need to ensure the reliability of stimuli for their own participant sample. The verification of the normative data suggesting the effect of each stimulus can for instance be done through a re-assessment of stimuli prior to study conduction.

Factor-Related Reliability – Investigating the Interplay of Individual Factors That may Determine Reliability

Numerous examples support the assumption that various factors may influence the perception of emotional stimuli. Especially, as some of the available original stimuli were created as long as 60 years ago (Barrington, 1963), the question arises, whether and to what extent researchers can *indeed* rely on the normative rating data. The answer to this question may impose the need for a re-evaluation of stimuli regarding certain factors. It is conceivable that if the simple manipulation of contextual information (Cahill, 1975; Prada et al., 2016), as well as the content inherent to a stimulus (e.g., politically, morally, economically, aesthetically, etc.) influences stimulus interpretation, the normative rating of (certain) stimuli may not remain reliable when reassessed today by a typical psychology study sample. Given the possible questionability regarding the reliability and validity of extant normative rating data, reported results from previously studies may need re-evaluation, and normative rating data may need updating.

While the influence of factors such as participant's age, gender, or ethnicity have been widely accepted by the research community, the influence of other factors such as assessed dimension (e.g., valence, arousal), dimension category (e.g., positive/negative/neutral valence; low/medium/high arousal), or stimulus type have only received little to no attention. This means that despite the availability of this information (as can be seen in the KAPODI database) empirical data regarding the relation between these factors and the perception of ES is scarce. To this date, focus has never been put on stimulus reliability in dependence of individual factors

as mentioned above, or an interplay of multiple factors (e.g., image compared to word stimuli assessed by female compared to male participants; comparison of dimension categories for female and male participants); a *factor-related reliability*. In fact, this could be one possible reason for inconclusive findings reported within emotional research (for a comparison, see Hostler et al., 2018).

If a factor-related reliability exists, a comparison across dimension categories may for instance suggest that a specific combination of factors (e.g., high-arousal stimuli) are especially reliable, while other combinations (e.g., low-arousal stimuli) are unreliable. Moreover, this may be true for one type of stimuli (e.g., images), but not for another (e.g., words). To give an example, a researcher seeking to investigate the effect of valence onto skin conductance may select image-stimuli that are matched on low arousal to control for the effect of arousal (see *The Importance of Reliable Emotional Stimuli*, above). If, however, low-arousing image-stimuli are unreliable (which means that perception of these stimuli assessed today significantly differs from the normative data because a number of participants may perceive the stimuli as medium arousing), the researcher fails to control for arousal and hence may draw false conclusions from the measured skin conductance that was – unknowingly – affected by arousal.

Therefore, to determine the parameters of reliability of emotional stimuli, broad investigations incorporating multiple factors at once are necessary. This will simultaneously allow for a more specific analysis regarding the interplay of factors, as well as display distinct combinations of factors that determine data reliability. Rather than denouncing stimuli as reliable or unreliable in general, the factor-related reliability will help indicating (un)reliability of stimuli in relation to a specific study construction. Therefore, the present study aims to refine the determination of stimulus reliability by investigating the effect of various factors that may influence stimulus reliability (e.g., *dimension (valence, arousal, dominance)*, *dimension category* and *SD category (high, medium, low)*, *stimulus type (images and words)*, and *gender (female, male)*.

Besides containing information regarding many of assessor-related factors (e.g., the rating population's age, gender, or ethnicity), the KAPODI database (see *Chapter Two*) contains information such as applied type of rating scale and rating scale length, or also year of set publication. The results of the present study may hence serve as a valuable indicator for scientists planning to use ES and will allow researchers choose stimuli more effectively from the database for their study.

The Current Study

To allow generalizability of results, existing stimuli will be reassessed by a typical psychology study sample of adults, including both female and male participants from the age of 18 years onwards. The ratings will be assessed on the commonly used dimensions valence, arousal, and dominance and will furthermore be separated into three dimension categories (low, medium, high) for all three dimensions. Additionally, an approach/avoidance tendency will be assessed allowing a comparison with valence. Among existing types of stimuli (e.g., images, audio clips, video clips, or words) (image)-frames are the basis of video clips, and words the basis for phrases, text passages as well as language-based audio clips (sound excluded). Therefore, word and image stimuli were selected for the current study. Moreover, in comparison, it would not have been possible to ensure the proper functioning and display of video stimuli during the study completion, as this study was conducted online. The inclusion of both, word as well as image stimuli, will allow investigation and comparison across stimuli types. The prevalence, that is the percentage of reliable stimuli among all included stimuli, will serve as an indicator for the factor-related reliability.

To calculate the number of participants required in these studies, a power analysis was conducted based on comparable previously reported effect sizes taken from Hostler et al., (2018). To detect the smallest observed effect of e.g., dimension category that is, for example positive vs. neutral cues (d = 0.32), a sample size of N = 80 is required.

Method

Participants

A total of 142 participants (81 female, 59 male, 1 non-binary; 18-72 years old, mean age: 31.83 years) completed the survey. Of these, 17 participants were excluded: 16 because they did not pass at least 50 % of the attention checks, and one participant who took more than 9 hours to complete the survey. All analyses are based on the remaining 125 participants. Of these, 107 were recruited via the webpage Prolific and 18 were recruited through advertisements around a university in North West England as well as word-of-mouth. The participation criteria within Prolific were set through filters; these were: a minimum age of 18 years, no medical history, normal/corrected-to-normal vision, as well as fluency of the English language. There were 74 female and 50 male participants as well as 1 of non-binary gender; mean age was 31.88 years (*SD* = 11.90). Female participants had a mean age of 32.26 years, and male participants a mean age of 31.10 years. One participant of non-binary gender was 43 years old. Forty-nine participants were from the United Kingdom, *n* = 27 from South Africa, *n* = 25 from the United States of America, and *n* = 24 from Canada. All participants were native English-speaking, at

least 18 years old and without a history of mental disorders. They had normal or corrected-to normal vision.

Materials

The complete set of stimuli (n = 100) used in this study, comprised 50 images and 50 words. All stimuli were selected from previously published sets listed in the KAPODI database (Diconne et al., 2022) which offers comparison of available ES sets across various set characteristics (see *Chapter Two*). The 100 stimuli were combined as follows: 25 words from the *Indiana Sexual and Affective Word Set* (ISAWS; Stevenson et al., 2011); 25 words from the *taboo, emotionally valenced and emotionally neutral word list* (Janschewitz, 2008); 25 images from the *DIsgust-RelaTed-Images* (DIRTI; Haberkamp et al., 2017); 21 images from the *International Affective Picture System* (IAPS; Lang et al., 1997); and 4 images from the *Galician Beverage Picture Set* (GBPS; López-Caneda & Carbia, 2018). The sets are referred to as *ISAWS*, *Janschewitz*, *DIRTI*, *IAPS*, and *GBPS* from here on; all included stimuli are listed in the *SM (Study 2, File B)*.

Normative rating data reported in the original work, had been collected on 9-point scales for all five included sets. A balanced choice of stimuli was selected based on normative ratings. Stimuli from the two word sets were selected based on normative values of valence and arousal: twelve words were selected based on their valence rating (n = 3: highest rating; n = 3: lowest rating; n = 3: highest standard deviation (*SD*); n = 3: lowest *SD*), and thirteen words were selected based on their arousal rating (n = 3: highest rating; n = 3: lowest rating; n = 3: highest *SD*; n = 3: lowest *SD*; n = 1: with medium *SD*).

The 25 images taken from the DIRTI, were selected with a similar strategy. Finally, the 21 images selected from the IAPS, were selected based on their classification into high arousal/low arousal pleasant photographs and high arousal/ low arousal unpleasant photographs. The remaining 4 images taken from the GBPS were selected as neutral filler images; they all display beverages with (n = 2) or without (n = 2) human individuals in the frame.

Procedure

Four blocks were created, each comprising 25 stimuli that were either solely words or solely images. Each block contained stimuli from only one set, except for one block combining the 21 images from the IAPS and the 4 images from the GBPS.

In the original studies, the stimuli from the DIRTI and the taboo word list had been rated on Likert-type scales (Likert, 1932), the other three sets of stimuli had been rated using the *Self-Assessment Manikin* (SAM-)scale (Bradley & Lang, 1994; Lang, Bradley, & Cuthbert, 1999).

In order to keep factors affecting rating results minimal, the same scale and instruction as in the original work were used introducing each block. For an overview of the study flow, see Figure 4.

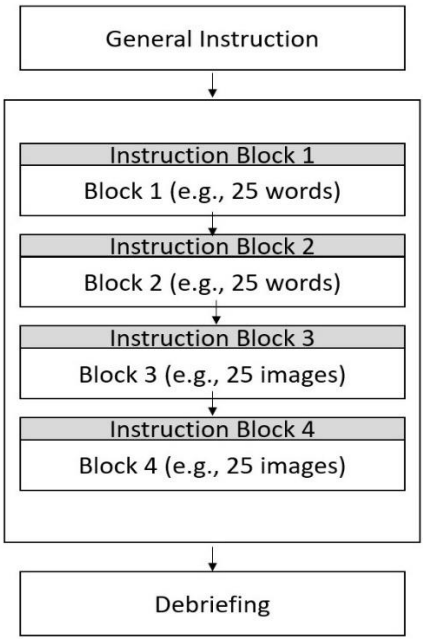


Figure 4. Flow of the study procedure. Block order and order of items within each block randomized across participants.

Prior to commencing the questionnaire, participants were informed about the aim of the survey. Confidentiality of personal data was assured, and participants were free to quit participation at any time by closing the web-browser. Names of the researchers including e-mail contacts as well as contact information in case of distress caused by the content of the survey were provided. The criteria of the consent form were presented in separate boxes that participants agreed with, by ticking them. Only upon full approval could participants begin the survey.

Following a general instruction, the two different types of rating scales (SAM-scale; Likert scale) were explained. During the rating procedure, each of the four blocks were introduced with their specific rating instruction accompanied by an example. Participants were asked to carefully read these instructions, as they slightly varied across blocks. Participants were introduced to the anchor terms used for the three dimensions (valence, arousal, dominance) in dependence on the individual applied scales (see Table 14 for comparison between introductions). Within each block the rating scales remained identical across stimuli.

Table 14

Introductions to The Anchor Terms Used in The Block-Specific Instructions

General Instruction	*Set Name*	*Block-specific Instruction*
"In the following section you will be presented with 4 blocks, each containing 25 items. *There will be either words or images displayed individually on your device screen.* *Please rate each word and each image using the scales below it.* *Introducing each block, will be a short description of the scales and instructions on how to use them.* *There is no right or wrong answer. However, it is important that you select the answers according to* ***your very personal perception*** *of each word or image."*	Janschewitz	*"In the following block, you will be presented with words. Please indicate which option best describes the word for you. If you don't know the word, select "I do not know this word".* *On the scales (from 1 - 9), please select the number that best represents how you perceive the word, and how comfortable you feel, using it.* *Below are examples of the scales:* *How positive or negative is the word?* *Give a 1-9 rating:* *1: "strongly negative" – 9: "strongly positive"* *(5 represents "not negative or positive")* *How exciting is the word? Consider how much the word grabs your attention.* *Give a 1-9 rating:* *1: "not at all arousing" – 9: "very arousing"* *(5 represents "medium arousing")* *How dominant is the word?* *Give a 1-9 rating:* *1: "not at all dominant" – 9: "very dominant"* *(5 represents "medium dominant")* *How comfortable do you feel, using this word?* *"not at all comfortable" – "very comfortable""*
	DIRTI	*"In this following section you will see photographs. There will be 9 buttons in each scale. Please answer each of the questions below the displayed photograph, by selecting the button that best represents your feeling."* *(anchors were for* *valence: "very negative" – "very positive"* *arousal: "none" – "very strong"* *dominance: "not dominant" – "very dominant"* *"I would…* *…avoid it – …approach it)"*
	ISAWS	*"In this block, you will be presented with words. For this section you will use the SAM-scale. 3 sets of 5 figures will be presented to you. Each is arranged along a continuum from 1 - 9. You will be using these figures to rate how you perceive the displayed word.*

		SAM shows three different kinds of feelings: "extremely negative" vs. "extremely positive", "not at all arousing/low energy" vs. "extremely arousing/high energy", and "not at all dominant" vs. "extremely dominant". *Select the figure that best represents how you perceive the word. If your estimation lays between two figures, select the space in-between them.* *Below are examples of the scales.* *If you don't know the word, select "I do not know this word."'*
	GBPS & IAPS	*"In this block, you will be presented with images. For this section you will use the SAM-scale. 3 sets of 5 figures will be presented to you. You will be using these figures to rate how you perceive the displayed image. SAM shows these three different kinds of feelings:* *"unhappy" vs. "happy", "calm" vs. "excited", and "controlled" vs. "in-control".* *Select the figure that best represents your feeling. If your feeling lays between two figures, select the space in-between them.* *Below you will find a more detailed explanation for each scale.* *"unhappy" vs. "happy":* *One extreme of the unhappy vs. happy scale is when you felt completely unhappy, annoyed, unsatisfied, melancholic, despaired, bored; at the other end of the scale, you felt completely happy, pleased, satisfied, contented, hopeful.* *"calm" vs. "excited":* *At one extreme of the scale you felt completely relaxed, calm, sluggish, dull, sleepy, unaroused; at the other end of the scale you felt completely stimulated, excited, frenzied, jittery, wide-awake, aroused.* *"controlled" vs. "in-control":* *At one end of the scale you have feelings characterized as completely controlled, influenced, cared-for, awed, submissive, guided; at the other extreme of the scale, you felt completely controlling, influential, in control, important, dominant, autonomous."*

Note. Visual examples of the scales with anchors followed the set-specific instruction. Janschewitz = Janschewitz, 2008; DIRTI = Haberkamp et al., 2017; ISAWS = Stevenson et al., 2011; GBPS = López-Caneda & Carbia, 2018; IAPS = Lang et al., 1997.

Stimuli were individually presented on the centre top of the device screen, with the 9-point rating scales below. Each stimulus was assessed on four dimensions: (*1*) valence (1 = "*very/extremely/strongly negative; unhappy*" to 9 = "*very/extremely/strongly positive; happy*", (*2*) arousal (1 = "*not at all arousing/low energy/none; calm*" to 9 = "*very arousing/extremely arousing/high energy/very strong; excited*")*;* (*3*) dominance (1 = "*not at all dominant/ not dominant; controlled*" to 9 = "*extremely/very dominant; in control*")*;* (*4*) approach/avoidance tendency (images: 1 = "*I would avoid it*" to 9 = "*I would approach it*"; words: "*How comfortable do you feel using this word?*" from 1 = "*not comfortable at all*" to 9 = "*very comfortable*").

To prevent missing values, scales were set to forced response. To verify that participants were reading instructions carefully four quality check questions (one for each block) were included prompting participants to select a specific response, instead of freely rating the stimulus ("*Please select '7' for valence arousal and dominance here*", and for approach/avoidance tendency: "*On this scale from 1-9, please select the equivalent position to '7' here*"). The indicated number to be selected varied across the four quality check questions. Only the data of participants who passed at least two (50 %) of the quality checks were included.

Finally, participants were debriefed and given the choice to receive information about study results as well as to enter a prize draw for a £ 20 shopping voucher. Students from Manchester Metropolitan University received participation points; participants from Prolific were compensated with an average of £ 4.30 for their participation. The study received ethical approval from the Manchester Metropolitan University faculty ethics committee, and data were collected between December 2019 and June 2020.

Study Design

To answer the above-mentioned research questions, n = 100 stimuli (50 images; 50 words) were assessed on valence, arousal, dominance, as well as approach-avoidance tendency on 9-point scales (either SAM-scales, or Likert-tape scales) by up to 125 participants. Each participant rated each stimulus on all four dimensions in a one-point data collection study. The effects of the individual factors for two genders (females, males), two types of stimuli (images, words), and three dimensions (valence, arousal, dominance) were investigated in a 2 × 2 × 3 (gender × stimulus type × dimension) factorial design; as well as extended by dimension and *SD* category (low medium, high) in a 2 × 2 × 3 × 3 (gender × stimulus type × dimension × *SD* category) factorial design. While these factors were the independent variables, the assessed stimulus rating represented the dependent variable.

Results

Sample Characteristics

In the present study, 124 to 125 participants between the age of 18 and 72 years rated each of the 100 stimuli. Due to technical difficulties, 11 stimuli were rated by solely 124 participants instead of all 125 participants. Completion of the survey took participants between 15 and 128 minutes (mean: 38.8 minutes; *SD* = 18 minutes). Raw rating data per participant can be found in the *SM (Study 2, File A)*.

For comparison, number, age, and cultural background of participants of the original studies as well as the present study are listed in Table 15, below.

Table 15

Age, Gender, Ethnicity and Number of Participants Assessing Emotional Stimuli in Original Study and Present Study

Original Study		*Present Study*
Janschewitz	· *N* = 78 participants · *N* raters per stimulus: n/a · native-English-speaking college students, (USA) · age: n/a	· *N* = 125 participants · *N* = 124 - 125 raters per stimulus · (UK: *n* = 49; SA: *n* = 27; USA *n* = 25; CAN *n* = 24) · age: 18 - 72 years
DIRTI	· *N* = 200 participants · *N* = 200 raters per stimulus · (DE) · 18-75 years old	
ISAWS	· *N* = 1099 participants · *N* = 62 raters per stimulus · native English-speaking undergraduate students, (USA) · age: 18 - 50 years	
GBPS	· *N* = 201 participants · *N* raters per stimulus: n/a · college students, (ES) · age: 17 - 28 years	
IAPS	· *N* participants: n/a · *N* = 100 raters per stimulus · college students, (USA) · age: n/a	

Note. CAN = Canada; DE = Germany; ES = Spain; SA = South Africa; UK = United Kingdom; USA = United States of America; n/a = information not provided. Janschewitz = Janschewitz, 2008; DIRTI = Haberkamp et al., 2017; ISAWS = Stevenson et al., 2011; GBPS = López-Caneda & Carbia, 2018; IAPS = Lang et al., 1997.

Analytical Approach

The overall analytical approach was designed to explore patterns of reliability of specific groups of stimulus characteristics (e.g., dimension: valence, arousal, dominance;

assessors gender: female, male). This was conducted by correlating the collected idiographic rating to the available normative rating data. The results aimed to indicate patterns regarding which stimuli groups displayed similarity (or diversity) of stimulus ratings.

First, means and standard deviations were calculated from participant ratings for each stimulus on each dimension (valence, arousal, dominance, approach/avoidance). As previous research has displayed gender differences regarding emotion perception of stimuli (e.g., Kemp et al., 2004; Kuypers, 2017; Memon et al., 2019), means and standard deviations were calculated for all (see *SM: Study 2, File B),* female (*File C*), and male participants (*File D*) separately. A correlation between female and male ratings was calculated regarding each dimension (valence arousal, dominance, approach/avoidance) separately for image and word stimuli. This means that a bivariate correlation was calculated for example between female valence ratings and male valence ratings for image stimuli. An independent sample t-test was conducted to calculate a gender mean difference score for each stimulus and to investigate the prevalence of stimuli with significant gender differences (*File E*).

Second, to investigate the role of factors in relation to stimulus reliability in a typical psychology study sample, three blocks were formed changing in focus regarding the factors *dimension* (e.g., valence, arousal, dominance, approach-avoidance), *dimension categories* and *standard deviation categories* (e.g., high, medium, low), *stimulus type* (e.g., images and words), as well as *gender* (e.g., female, male).

First, a bivariate correlation between valence ratings obtained in the present study and the normative valence ratings of the same gender group was conducted for each participant separately and with regards to stimulus set (e.g., Janschewitz, DIRTI, ISAWS, GBPS, IAPS) as well as stimulus type (images, words). That is, for example valence rating data of each of the ISAWS stimuli from a female participant in the present study were correlated with normative valence rating data from females in the original ISAWS study; the same applied to male rating data respectively. This was repeated for the other two dimensions arousal and dominance. The calculated individual correlation coefficients can be found in the *SM (Study 2, File F)*. Note that approach/avoidance was not included, as stimuli sets were not assessed on this dimension in the original studies.

In other words: 7 + 7 + 2 = 16 correlation coefficients (valence for Janschewitz, DIRTI, ISAWS, GBPS, IAPS, only images, only words, + arousal for Janschewitz, DIRTI, ISAWS, GBPS, IAPS, only images, only words, + dominance for ISAWS, and IAPS) were calculated for each participant. This information was used for comparison between idiographic and normative data regarding the factors *stimulus set*, *dimension*, *gender*, and *stimulus type*. The

reliability of ES in dependence of the included factors is represented by the percentage of participants displaying a significant correlation ($p < .05$ as well as $p < .01$) between idiographic and normative rating data – with increasing percentages indicating increasing reliability.

Second, correlations as described above were repeated, however, instead of calculating correlations from data for a particular stimulus set, they were calculated for stimuli across three dimension categories: that is, based on the mean normative rating data, stimuli (from all sets) were allocated to one of the three categories *low*, *medium*, or *high* regarding valence, arousal, and dominance. For an equal width of the three categories, cut-off points for the 9-point rating scale were 1 – 3.66 for *low*, 3.67 – 6.33 for *medium*, and 6.34 – 9 for *high*. This way, (3 + 3 + 3) x 2 = 18 correlation coefficients [(low/medium/high valence + low/medium/high arousal + low/medium/high dominance) x 2 (images, words)] were calculated for each participant (stimulus allocation can be found in *File G;* the calculated individual correlation coefficients in *File J*). This information was used for comparison between idiographic and normative data based on *dimension*, *dimension category*, *gender*, and *stimulus type*.

Third, correlations as described above were repeated, however, instead of calculating correlations from the means of normative rating data categories (e.g., low, medium, high), three categories were created from the *SD* of stimulus ratings. This was done to explore differences regarding stimulus reliability in dependence of rating dispersion. That is, stimuli ratings with a small *SD* reflect a relatively higher agreement regarding stimulus perception compared to stimuli ratings with a high *SD*. It will be expected that stimuli with a low *SD* in the normative rating data also display a low *SD* when assessed today and hence are more reliable stimuli with regards to that categorization.

As stimulus *SD* ranged from 0 (e.g., valence rating for *basket*) to 3.6 (e.g., male arousal ratings for *kike*), selected cut-off points for the three categories were: *low* = 0 – 1, *medium* = > 1 – 2, and *high* = > 2. The exact stimulus allocation can be found in *File H*. This way, (3 + 3 + 3) x 2 = 18 correlation coefficients [(low/medium/high *SD* for valence + low/medium/high *SD* for arousal + low/medium/high *SD* for dominance) x 2 (images, words)] were calculated for each participant. *SD* for 4 images (GBPS) were not available, hence calculations include 96 stimuli. This information was used for comparison between idiographic and normative data based on *SD category*, *dimension*, *gender*, and *stimulus type*.

Finally, familiarity of word stimuli that was assessed through the answer option "*I do not know this word*", was calculated in relation to the number of raters and is available in the *SM (Study 2, File K)*.

In the following two sections, results will be presented in the same order as described above. This means that results regarding the analysis of gender differences will be presented first, followed by the comparison of idiographic to normative rating data. The results of the comparison between idiographic and normative data will be presented in individual blocks in relation to the focus on the specific investigated factors: *(1)* results of analyses conducted with focus on the stimuli sets as well as the factors dimension, gender, and stimulus type; *(2)* results of the analyses conducted with focus on the factors dimension, gender, stimulus type, and dimension category; and *(3)* results of the analyses conducted with focus on the factors dimension, gender, stimulus type, and *SD* category.

Gender Differences

Mean correlations between female and male ratings were high for all four dimensions for image stimuli (valence: r = .99; arousal: r = .91; dominance: r = .92; and approach/avoidance: r = .99), as well as word stimuli (valence: r = .98; arousal: r = .94; dominance: r = .92; and approach/avoidance: r = .96). The only non-binary participant completed the survey and was separated from female and male participant analyses. This participant had the same strength of correlation with both, female and male mean ratings on all dimensions: the correlation was the same to both genders for valence (r = .88), slightly higher with the female compared to male mean ratings for arousal (r = .71 vs. r = .72) and approach-avoidance (r = .77 vs. r = .76), and slightly higher with male compared to female mean ratings for dominance (r = .32 vs. r = .31). Differences in strength of correlations are negligible. Necessity of inclusion of non-binary gender participants for future research will be discussed in *Chapter 5*.

A significant gender difference was found in 4 % (image stimuli assessed on dominance) to 30 % (image stimuli assessed on valence) of the stimuli, depending on assessed dimension and stimulus type (see Table 16). Significant differences were more often found for image compared to word stimuli on the dimensions valence, arousal, and approach/avoidance. On the dimension dominance, not only were significant gender differences overall less frequent than in the other three dimensions, but also were differences more frequently for word compared to image stimuli.

Table 16

Percentages of Stimuli with Significant Gender Differences Separated by Type of Stimuli and Assessed Dimension

Stimulus type	Dimension				
	valence	*arousal*	*dominance*	*approach/ avoidance*	*mean all*
images	30 %	28 %	4 %	20 %	20.5 %
words	20 %	16 %	10 %	12 %	14.5 %

Note. $P < .05$.

Despite an overall high correlation between female and male ratings of ES, the present findings support previous research results displaying significant gender differences. Moreover, these results suggest that the factor gender may play a more important role regarding emotion perception of image stimuli compared to word stimuli.

Given the extensive number of stimuli displaying differences in perception between both genders, all following analyses were conducted for female and male participants separately.

Comparison of Idiographic to Normative Rating Data

Focus on the Factors: Stimulus Set, Dimension, Gender, and Stimulus Type

Results regarding the factor *dimension*, display the highest mean correlation for valence, followed by arousal and dominance (see Table 17 /Figure 5).

Investigations of the factor *gender* displayed similarly strong mean correlations for female and male participants. However, these were in a slightly different order when comparing the individual sets: on the dimension valence female participants showed the highest correlation for DIRTI ($r = .86$), followed by Janschewitz ($r = .80$), IAPS ($r = .80$), ISAWS ($r = .75$), and GBPS ($r = .02$). Correlation scores for the dimension arousal were highest for ISAWS, with a medium correlation of $r = .53$, followed by IAPS ($r = .5$), Janschewitz ($r = .4$), DIRTI ($r = -.24$), and a low correlation for GBPS ($r = .07$). Only two sets, the IAPS and ISAWS had normative data for the dimension dominance. Correlations were equally low for both sets with $r = .33$. For male participants mean correlation on the dimension valence was strong and highest for the DIRTI ($r = .83$), followed by IAPS ($r = .82$), ISAWS ($r = .81$), Janschewitz ($r = .77$), and a low negative correlation for the GBPS ($r = -.02$). Correlation scores for the dimension arousal were highest and of medium strength for ISAWS ($r = .58$), followed by IAPS ($r = .47$) and Janschwitz ($r = .42$), as well as DIRTI ($r = -.34$), and GBPS ($r = .10$) with a low correlation, that was additionally negative for the DIRTI. Correlations for the dimension dominance were equally low for both ISAWS (r =.32) and IAPS (r =.30). Note that the analyses conducted

regarding the individual stimuli sets are presented to provide additional information regarding the means for word and image stimuli provided in the table. This is particularly relevant, as there were for instance only 4 images included from the GBPS.

Regarding the factor *stimulus type*, solely on the dimension valence, both, image as well as word stimuli, displayed a significant correlation ($p < .01$) between idiographic and normative rating for *all* ($n = 125$) participants independent of gender. On the dimension arousal, there were more participants with a significant correlation for word stimuli compared to image stimuli, and on the dimension dominance there were more participants with significant correlations for image stimuli than word stimuli. This tendency was the same for both genders.

Table 17

Mean Participants' Correlation Coefficient per Dimension for Stimuli Sets and Types of Stimuli – Separated by Gender, and for two Alpha Levels

	Stimulus Set					Stimulus Type	
	Janschewitz	*DIRTI*	*ISAWS*	*GBPS*	*IAPS*	images	words
Females							
valence	0.795	0.860	0.752	0.017	0.790	0.800	0.772
* $p < .05$	100 %	100 %	98.65 %	0 %	98.65 %	100 %	100 %
** $p < .01$	95.95 %	100 %	95.95 %	0 %	95.95 %	100 %	100 %
arousal	0.393	-0.244	0.525	0.067	0.500	0.353	0.422
* $p < .05$	56.76 %	71.83 %	74.32 %	1.37 %	68.92 %	64.86 %	78.38 %
** $p < .01$	43.24 %	60.56 %	64.86 %	0 %	48.65 %	43.24 %	68.92 %
dominance	-	-	0.326	-	0.328	0.328	0.326
* $p < .05$	-	-	39.19 %	-	51.35 %	51.35 %	39.19 %
** $p < .01$			22.97 %		33.78 %	33.78 %	22.97 %
Males							
valence	0.769	0.827	0.806	-0.024	0.815	0.805	0.777
* $p < .05$	98 %	100 %	100 %,	2.17 %	96 %	100 %	100 %
** $p < .01$	96 %	100 %	96 %	2.17 %	90 %	100 %	100 %
arousal	0.418	-0.341	0.584	0.099	0.466	0.471	0.494
* $p < .05$	58 %	60 %	82 %	0 %	62 %	76 %	88 %
** $p < .01$	40 %	50 %	66 %	0 %	46 %	66 %	80 %
dominance	-	-	0.322	-	0.301	0.301	0.322
* $p < .05$			36.73 %		57.14 %	57.14 %	36.73 %
** $p < .01$	-	-	24.49%	-	44.90 %	44.90 %	24.49 %

Note. Females: $n = 71$ to 74, males: $n = 48$ to 50; correlations were calculated using same-gender normative rating data means; correlation scores rounded to third decimal; % = percentage of participants with a significant correlation to normative rating data. Janschewitz = Janschewitz, 2008; DIRTI = Haberkamp et al., 2017; ISAWS = Stevenson et al., 2011; GBPS = López-Caneda & Carbia, 2018; IAPS = Lang et al., 1997. Note that only four stimuli were included from the GBPS.

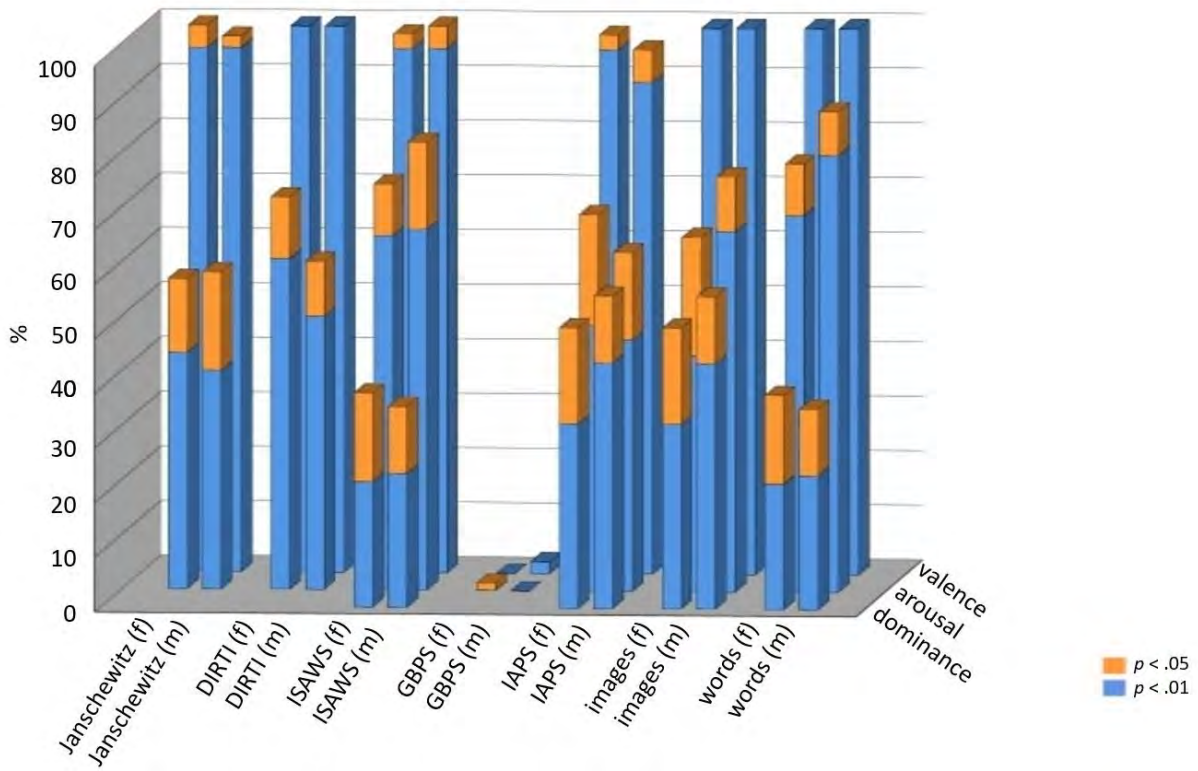


Figure 5. Percentages of participants with significant correlations ($p < .05$; $p < .01$) between idiographic and normative ratings. Janschewitz = Janschewitz, 2008; DIRTI = Haberkamp et al., 2017; ISAWS = Stevenson et al., 2011; GBPS = López-Caneda & Carbia, 2018; IAPS = Lang et al., 1997; f = females; m = males.

Focus on the Factors: Dimension, Dimension Category, Gender, and Stimulus Type

The distribution of stimuli across the three rating categories and separated by gender and dimension based on the normative rating can be found in Table 18 below.

Table 18

Number of Stimuli Across the Dimension Categories and Separated by Gender and Dimension

Dimension Category	Valence		Arousal		Dominance	
	female	*male*	*female*	*male*	*female*	*male*
low	34	32	48	48	9	5
images	14	13	27	29	4	3
words	20	19	21	19	5	2
medium	39	32	40	32	24	29
images	15	11	17	14	9	13
words	24	21	23	18	15	16
high	27	36	12	20	13	12
images	21	25	6	7	8	5
words	6	11	6	13	5	7

Note. $N = 100$ stimuli.

Results regarding the factor *dimension*, display the highest mean correlation between normative and idiographic rating data for valence, followed by arousal and dominance (see Table 19 /Figure 6). Overall, correlation scores remain low to medium ($r < .5$) irrespective of assessed dimension, rating category, gender, and stimulus type. Separated by gender, stimulus type and dimension category, for image stimuli and on the dimension valence female participants showed the highest correlation for stimuli falling into the *medium* rating category ($r = .34$), followed by *low* ($r = .23$), and *high* ($r = .19$). This same order was maintained for arousal: *medium* ($r = .24$), *low* ($r = -.10$), and *high* ($r = -.03$). Correlation scores on the dimension dominance were similarly low for the rating categories *medium* ($r = .05$) and *high* ($r = .02$); there were not enough stimuli falling into the *low* category on this dimension. For word stimuli assessed for valence, the highest correlation was found regarding stimuli falling into the *low* category ($r = .42$), followed by *high* ($r = .33$), and *medium* ($r = .30$). For arousal the order was *high* ($r = .48$), *low* ($r = .13$), *medium* ($r = .05$). There were not enough stimuli falling into the *high* or *low* category for dominance; the correlation regarding stimuli falling into the dimension category *medium* remained low with $r = .07$.

Correlation scores for male participants for image stimuli were similar to that of females (*medium*: r =.34; *low*: $r = .24$; *high*: $r = .16$); the same dimension category order was maintained for the arousal (*medium*: r =.15; *low*: $r = -.01$; *high*: $r = -.01$). There were not enough image stimuli falling into the dimension categories *low* and *high* for dominance; the correlation for *medium* was low with $r = .17$. Correlations for word stimuli assessed on valence were highest for the dimension category *low* (r =.29), followed by *medium* ($r = .02$) and *high* ($r = 0$). There were not enough word stimuli falling into the category *low* for dominance; correlations were similarly low for *medium* ($r = .06$) and *high* ($r = .08$).

With respect to the percentage of participants displaying a significant correlation ($p < .05$) of their idiographic rating to the normative rating, solely word stimuli assessed by female participants on valence and falling into the dimension category *low* reached above 50% with 67.57 %. Comparison between dimensions display the greatest number of participants with a significant correlation to normative rating for valence, followed by arousal and dominance. Separation into dimension categories showed that among the stimuli falling into the category *low*, a greater percentage of participants displayed significant correlations with normative rating data, compared to the dimension categories *medium* and *high*. This was especially true for valence and arousal. Nevertheless, depending on selected factors exceptions were visible (e.g., word stimuli assessed for arousal).

Table 19

Mean Participants' Correlation Coefficient per Dimension for Dimension Categories and Stimuli Types – Separated by Gender, and for two Alpha Levels

	Dimension Category (images)			Dimension Category (words)		
	low	*medium*	*high*	*low*	*medium*	*high*
Females						
valence	0.233	0.342	0.192	0.423	0.295	0.331
* $p < .05$	23.53 %	35.14 %	14.86 %	67.57 %	44.59 %	9.52 %
** $p < .01$	5.88 %	8.12 %	4.05 %	50 %	20.27 %	1.59 %
arousal	-0.095	0.240	-0.031	0.130	0.048	0.482
* $p < .05$	40.54 %	25.68 %	2.82 %	34.25 %	6.76 %	16.67 %
** $p < .01$	21.62 %	8.11 %	0 %	12.33 %	1.35 %	0 %
dominance	-	0.045	0.024	-	0.067	-
* $p < .05$		0 %	5.56 %		2.70 %	
** $p < .01$	-	0 %	0 %	-	0 %	-
Males						
valence	0.241	0.341	0.156	0.290	0.156	0.000
* $p < .05$	26.53 %	22 %	12 %,	34 %	14.29 %	2.08 %
** $p < .01$	2.04 %	6 %	2 %	8 %	2.04 %	0 %
arousal	-0.032	0.147	-0.012	0.170	0.019	0.120
* $p < .05$	36 %	6 %	2 %	38 %	2 %	14 %
** $p < .01$	26 %	0 %	0 %	16 %	0 %	4 %
dominance	-	0.173	-	-	0.060	0.078
* $p < .05$		10.20 %			12.24 %	2.08 %
** $p < .01$	-	4.08 %	-	-	0 %	0 %

Note. Females: n = 71 to 74, males: n = 48 to 50; correlations were calculated using same-gender normative rating data means; correlation scores rounded to third decimal; % = percentage of participants with a significant correlation to normative rating data, rounded to second decimal. Rating category *low* = 1 to 3.66, *medium* = 3.67 to 6.33, *high* = 6.34 to 9 (on a 9-pt. rating scale); blank cases indicate that there were 5 or less stimuli falling into that category, therefore a correlation score could not be calculated.

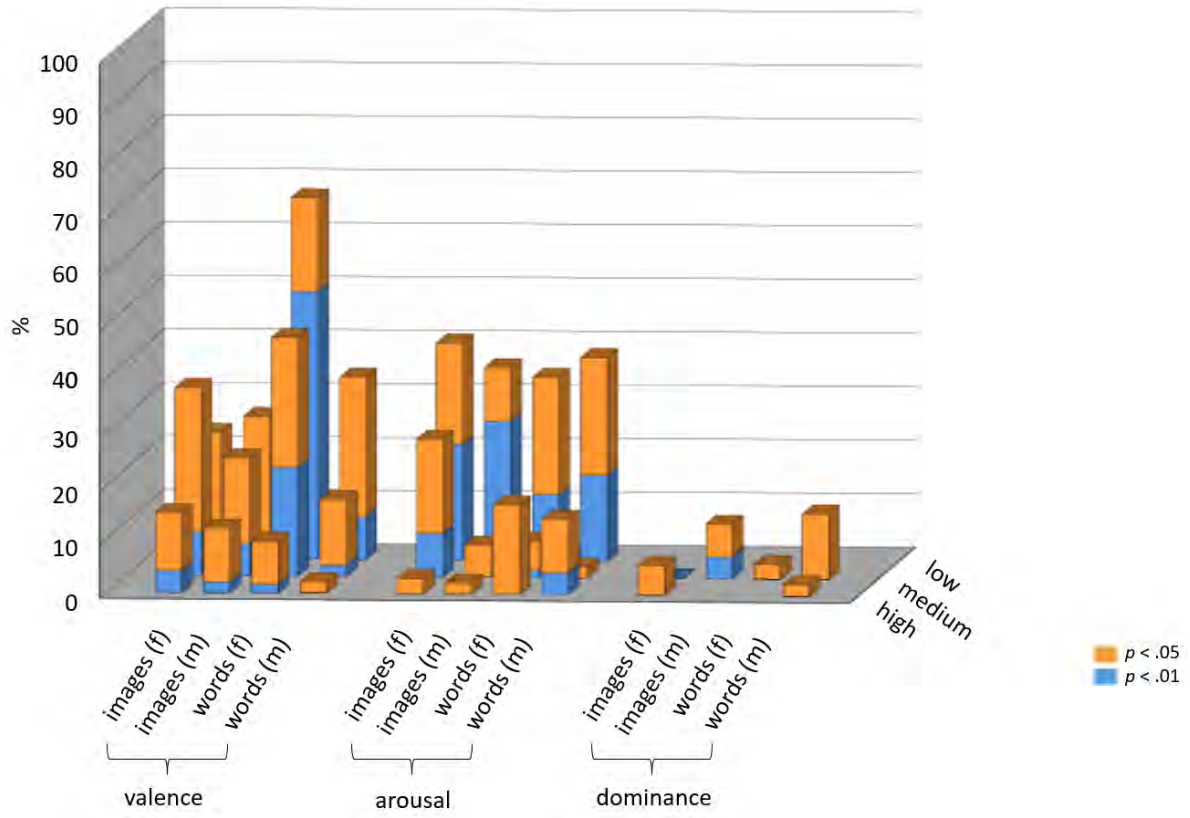


Figure 6. Percentages of participants with significant correlations between idiographic and normative ratings separated by rating category (low, medium, high), assessed dimension (valence, arousal, dominance), stimulus type (images, words), and gender (female, male).

Focus on the Factors: Standard Deviation Category, Dimension, Gender, and Stimulus Type

The distribution of stimuli across the three *SD* categories and separated by gender and dimension based on the normative rating can be found in Table 20 below.

Table 20

Number of Stimuli Across the SD Categories and Separated by Gender and Dimension

SD Category	Valence		Arousal		Dominance	
	female	*male*	*female*	*male*	*female*	*male*
low	17	15	12	14	0	0
images	7	2	4	6	0	0
words	10	13	8	8	0	0
medium	62	66	16	28	10	16
images	37	43	14	12	6	12
words	25	23	6	16	4	4
high	17	15	64	54	36	30
images	2	1	28	28	15	9
words	15	14	36	26	21	21

Note. $N = 96$ stimuli: valence = 96, arousal = 96, dominance = 46.

Results regarding the factor *dimension*, display the highest mean correlation between normative and idiographic rating data for valence, followed by arousal and dominance (see Table 21 /Figure 7). Overall, correlation scores varied between low (e.g., r = .02 for male arousal rating of word stimuli falling into the *low* normative *SD* category) and high (e.g., r = .80 for female valence rating of image stimuli falling into the *low* normative *SD* category), and were similar for both genders on the dimensions valence and arousal.

Regarding valence, female participants showed a higher correlation for image stimuli falling into the *low* (r = .80) than the *medium SD* category (r = .63). A higher correlation for image stimuli falling into the *high* (r = .36) than the *medium SD* category (r = .24) was visible for arousal. On the dimension dominance the correlation was slightly stronger for stimuli falling into the *medium SD* category (r = .30) than the *high SD* category (r = .13). For word stimuli correlation was highest for stimuli falling into the *medium SD* category (valence and arousal). Nevertheless, comparisons cannot be made across all categories equally as there were not enough stimuli falling into certain *SD* categories, thus not allowing calculation of a correlation (e.g., dominance ratings of image stimuli of *low SD*, or dominance ratings of word stimuli with *low* and *medium SD*).

For male participants the correlation for word stimuli on the dimension valence was highest for stimuli falling into the *SD* category *medium* (r = .60), followed by *high* (r = .54), and *low* (r = .48). The order was the same for arousal ratings (*medium*: r = .50; *high*: r = .29; *low*: r = .02).

The percentage of participants with a significant correlation ($p < .05$) between their idiographic rating and the normative rating, was highest regarding stimuli falling into the *medium SD* category on the dimension valence for both genders and both types of stimuli (e.g., 98 - 100%). Comparison across dimensions showed that valence displayed the highest number of participants with significant correlations to normative rating data, followed by arousal and dominance. This order was identical for both genders. Exact percentages can be found in Table 21.

Table 21

Mean Participants' Correlation Coefficient per Standard Deviation (SD) Category and Stimuli Types – Separated by Gender, and for two Alpha Levels

	SD Category (images)			*SD* Category (words)		
	low	*medium*	*high*	*low*	*medium*	*high*
Females						
valence	0.803	0.634	-	0.071	0.691	0.411
* $p < .05$	87.84 %	100 %		7.27 %	100 %	54.05 %
** $p < .01$	47.3 %	100 %	-	0 %	100 %	29.73 %
arousal	-	0.242	0.356	0.044	0.486	0.282
* $p < .05$		32.43 %	70.27 %	3.57 %	24.66 %	62.16 %
** $p < .01$	-	20.27 %	63.51 %	0 %	1.37 %	44.59 %
dominance	-	0.296	0.129	-	-	0.189
* $p < .05$		20.27 %	9.46 %			25.68 %
** $p < .01$	-	4.05 %	2.70 %	-	-	6.76
Males						
valence	-	0.630	-	0.482	0.596	0.541
* $p < .05$		100 %		62 %	98 %	78 %
** $p < .01$	-	100 %	-	18 %	92 %	50 %
arousal	0.221	0.320	0.459	0.019	0.500	0.289
* $p < .05$	4.76%	32 %	82 %	4.65 %	74 %	54 %
** $p < .01$	0 %	18 %	74 %	0 %	54 %	34 %
dominance	-	-0.044	0.353	-	-	0.234
* $p < .05$		4.08 %	38.78 %			32.65 %
** $p < .01$	-	2.04 %	8.16 %	-	-	14.29 %

Note. Females: $n = 55$ to 74, males: $n = 42$ to 50; correlations were calculated using same-gender normative rating data means; correlation scores rounded to third decimal; % = percentage of participants with a significant correlation to normative rating data, rounded to second decimal. *SD* category *low* = 0 to 1, *medium* = > 1 to 2, *high* = > 2; blank cases indicate that there were 5 or less stimuli falling into that category, therefore a correlation score could not be calculated.

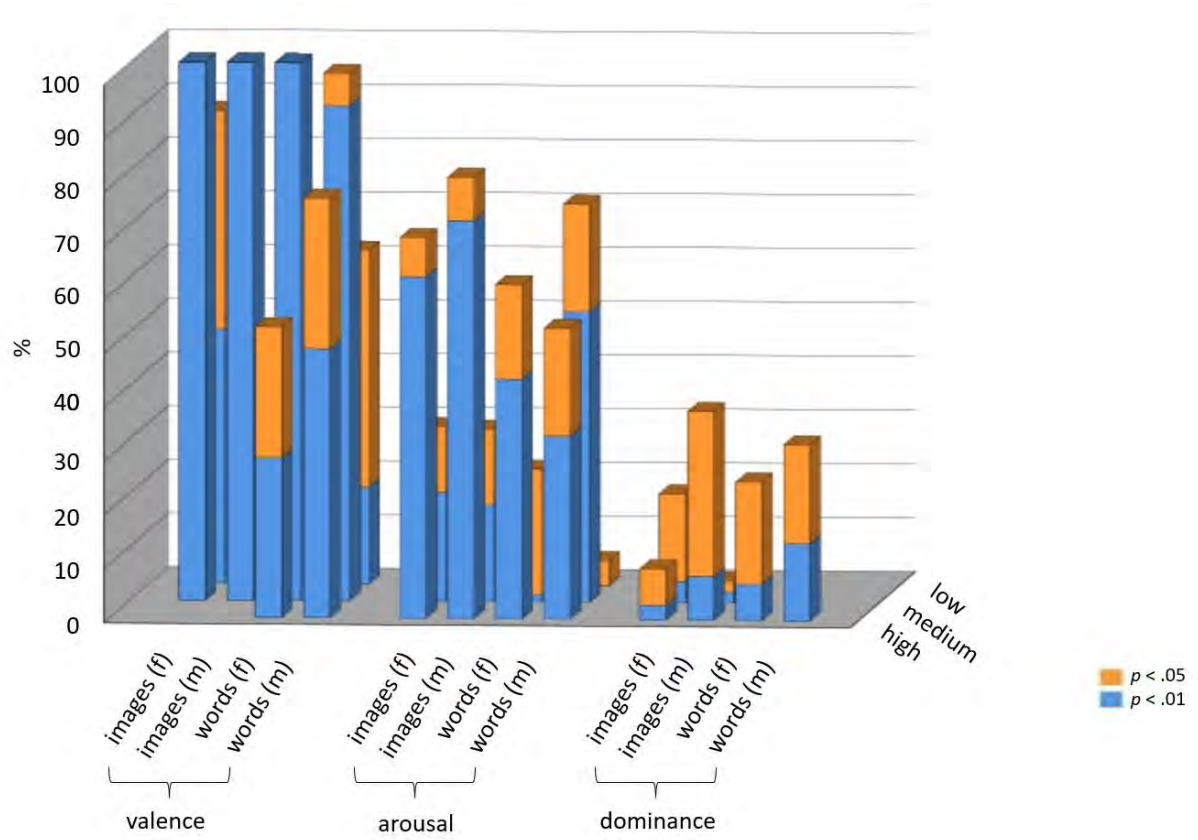


Figure 7. Percentages of participants with significant correlations between idiographic and normative ratings separated by *SD* category (low, medium, high), assessed dimension (valence, arousal, dominance), stimulus type (images, words), and gender (female, male).

Discussion

The present study investigated the prevalence of reliable ES in relation to various factors and in a typical psychology sample of adult participants. An analysis investigating the role of factors such as *dimension* (e.g., valence, arousal, dominance), *dimension category* and *SD category* (e.g., high, medium, low), *stimulus type* (e.g., images and words), as well as *gender* (e.g., female, male) was conducted. Results confirmed that rather than considering normative rating of emotional image and word stimuli generally (un)reliable, the degree of reliability varies with and therefore is dependent on individual factors.

Reliability of Normative Rating Data

Normative rating data from selected sets were originally assessed in the years 2018 (GBPS), 2017 (DIRTI), 2011 (ISAWS), 2008 (Janschewitz), and 1997 to 2008 (IAPS). This means that the normative rating data included in the present study has been assessed at various time points in the past between 2 to and up to 22 years ago. The separation into and comparison of dimensions suggests a greater reliability of valence ratings (e.g., higher correlation between idiographic and normative rating data and greater percentage of participants showing significant

correlations between idiographic and normative rating), compared to arousal and dominance. The latter dimensions correlating with low to medium strength with the normative rating data. Hence, results suggest that the perception of valence remains highly stable within 22 years. This order of dimensions (valence > arousal > dominance) with regard to the strength of correlation between normative and idiographic rating was visible for both types of stimuli, as well as for both genders. Moreover, it remained present when regrouping stimuli based on normative rating *SD* category. In a general tendency, findings were similar when regrouping stimuli based on normative dimension category, however there were exceptions (e.g., low-valanced image stimuli). The low to medium strength of correlations between idiographic and normative rating data based on the factor *dimension category* suggests this factor being the less adequate for stimulus selection among all factors investigated in the present study. In other words, researchers should refrain from selecting stimuli categorized into low/medium/high (valence, arousal, dominance) based on the normative data, as the current findings show that only a few participants show significant correlations between their perception and the normative data. Instead, if they wish to conduct research with ES categorized as low/medium/high on valence, arousal, or dominance, it is highly advised that researchers reassess stimuli for their participant group and categorize stimuli based on the idiographic data.

Separated by type of stimuli, the lower correlation between idiographic and normative valence data for word compared to image stimuli may indicate the tendency that on this dimension word stimuli may be slightly less stable and hence reliable over time, compared to image stimuli. Nevertheless, in the current survey, image stimuli displayed content that was not specifically related to an identifiable point in time, therefore, further research investigating stimuli with iconic content such as the intact Twin Towers or specific fashion is necessary to verify current findings.

Although the strength of correlation between idiographic and normative rating data for arousal was of low to medium strength, stimuli were perceived as *more* arousing assessed today compared to original assessment. This result seems especially surprising considering the ease of access as well as frequency of exposure of content similar to stimuli included in the current survey. That is, as images displaying death, violence, nudity, or diseases are frequently broadcasted on media (e.g., television, newspaper/magazines, billboards) and easily accessible online, one would expect stimuli to be perceived as *less* arousing due to affective habituation (e.g., Dijksterhuis & Smith, 2002; Ferrari et al., 2020). Nevertheless, this may indeed be an explanation valid for the dimension dominance, on which idiographic and normative ratings also correlated with solely medium strength, however, displaying *decreased* perception of

stimulus' dominance in idiographic rating. In other words, stimuli were perceived as less dominant today compared to the original (normative) assessment. Yet, considering the small number of stimuli that were available for the calculations for dominance (n = 25 words, and n = 25 images), further research is necessary to investigate the described effect.

Gender Differences

An overall strong correlation between the ratings of female and male participants collected in the present study (r = .91 to r = .99), support findings of Haberkamp and colleagues (2017) reporting a high correlation between genders for valence and arousal ratings ranging from r = .96 (valence) and r = .91 (arousal) for disgust eliciting pictures, to r = .96 (valence) and r = .68 (arousal) for neutral pictures of the DIRTI set. Yet, numerous included stimuli displayed significant rating differences based on assessed dimension. More detailed analysis revealed that differences between genders were significant for approximately one fourth or less of the stimuli, and more pronounced for image compared to word stimuli. A possible explanation for these findings may be gender differences in brain activity when processing for instance emotional image stimuli (e.g., Kemp et al., 2004; Wrase et al., 2003). Results of a study conducted by Kim and colleagues for instance suggest greater cortical processing of subliminally presented threat-related stimuli compared to male participants (Kim et al., 2013). With regards to previous findings concerning gender-specific use of language displaying females to be using more polite forms, apologies and being more emotional and evaluative compared to males (e.g., Haas, 1979; Mulac et al., 2001) as well as to give more extreme valence ratings than males (Bellezza et al., 1986), the current results seem surprising, especially as a large number of included word stimuli were sexual and taboo words such as insults or abusive language. Nevertheless, these gender differences may vary depending on culture. That is, while in western societies women are believed to be more emotional than men especially regarding emotion expression, differences between the two genders may be more pronounced compared to gender differences in non-Western countries. These gender differences are believed to stem from the culture's sex-specific division of labour as well as associated sex-role ideology (for a review, see Fisher & Manstead, 2000; Fischer et al., 2004).

Taken together, these results suggest that despite overall strong correlations between female and male ratings, emotion perception significantly diverges between genders for specific stimuli. This highlights the need for separate investigations and analyses of female and male participant data. To deepen the understanding in gender differences regarding emotion perception, however, further research is needed to investigate these very specific characteristics of stimuli that are perceived differently by both genders.

Understanding of Dimensions

One aspect that became salient during analyses was the use of scales across dimensions: while valence as well as approach-avoidance ratings had more variability and span across the entire length of the scale, arousal and dominance ratings remained towards the centre of the scale. As this tendency was visible for solely arousal and dominance ratings, rather than indicating the presence of a certain response style among participants (Paulhus, 1991), it may reflect an impeded understanding of these dimensions leading to an indecision due to uncertainty about one's position (see Baumgartner & Steenkamp, 2001). That is, while one is frequently confronted with how positive or negative something is in daily life situations and thus has to decide whether to approach or rather avoid it, the understanding of the concept of valence and approach/avoidance may be good, and ratings in a survey may be provided more easily by using the entire length of a scale. The concepts of arousal and dominance in contrast may be more abstract and less frequently encountered in daily life situations and hence make judgements on these scales more difficult, resulting in a response tendency toward the centre of the scale.

Limitations

Although providing valuable insight into the prevalence of reliable word and image stimuli ratings and highlighting factors that may determine reliability, the current study possessed some limitations: First, participants were recruited online through a crowdsourcing platform (prolific). As participants are financially remunerated after study completion and platform guidelines are comparably strict (e.g., completion of a study that is too fast may indicate the participant's inattention; attention check questions), it is fair to assume that high quality data were collected for the current study. Nevertheless, as data were collected online, there was no option to verify accuracy of demographic data, especially as data was collected during the time of the covid-19 pandemic which may have increasingly driven participants towards this platform. An issue related to this platform is that depending on the participants' profile, available studies are displayed only to the matching participant population. participation through this platform offers financial compensation, participants may be tempted to change their profile information allowing them to receive additional study options. Moreover, with online data collection, participant's attention to the questionnaire without distraction throughout the experimental procedure cannot be assured entirely.

Second, chosen stimuli are solely a small selection from individual sets, which may restrict generalizability of the current findings. As context has been shown to influence interpretation (Cahill, 1975; Prada et al., 2016), individually extracting stimuli from different

sets and reorganizing them, hence creating a new context among stimuli, may have impacted the results. As instructions were kept the same as original instructions, it was impossible to randomize all included stimuli regardless of their set-blocks. If rearrangement and mixing of stimuli from various sets indeed changes the assessment context through changed comparability among stimuli, however, generalizability of normative data may be limited whenever published sets are not used in their entirety and/or displayed among stimuli from other sets. In the present study stimuli sets were selected based on their original assessment scale (all stimuli had been assessed on 9-point scales). This was done to allow comparability of the idiographic to the normative rating data. Additionally, the choice of individual stimuli was based on the provided normative rating data (high/low valence; high/low arousal; high/low *SD*; see *Materials*). As the images from the images from the IAPS had been classified into high arousal/low arousal pleasant photographs and high arousal/ low arousal unpleasant photographs, neutral images were necessary to allow equal distribution across the four blocks of included stimuli (word stimuli assessed on a 9-point Likert scale; word stimuli assessed on a 9-point SAM scale; image stimuli assessed on a 9-point Likert scale; image stimuli assessed on a 9-point SAM scale). Therefore, the remaining 4 images taken from the GBPS were selected as neutral filler images. A replication of study with for instance only high valence or arousal stimuli could help gain additional insight into the question if the selection of stimuli (across the dimension span) may affect the participants' rating in a distinct way.

In a similar vein, despite an overall large sample size the availability of participants was nevertheless restricted by financial resource constrains as well as the accessibility of participants during the covid-19 pandemic. Separating the collected data points for comparison for example across genders automatically reduced the number of data points for individual analyses and hence affected statistical power of the results. In that regards, generalizability of the findings is restricted and replication with a larger participant sample suggested.

Third, previous research has shown that the duration of stimulus presentation influences liking (Marin & Leder, 2016; Reber et al., 1998). In the conducted study, viewing of stimuli was self-paced and hence varied between participants. Therefore, the collected rating data was susceptible to variation caused by differences in duration of stimulus presentation. Additional research is needed to investigate whether stimulus assessment data may differ in dependence of display duration as well as whether differences may vary along with stimulus type beyond the differences between words and images assessed in the current study.

Fourth, inclusion of potentially offensive stimuli remains challenging: while extreme stimuli are of great interest to researchers, ethical guidelines may narrow the range of chosen

stimuli. In this context, tabooing distinct words not only limits generalizability of findings, but also causes a serious threat to the objectivity and exhaustiveness of research and may therefore undermine the original purpose and thus value of research in general.

Final Discussion

The present study sought to answer the two leading questions *Q2.1: Is emotional stimulus reliability determined by factors associated with assessment of the stimuli [such as* dimension *(e.g., valence, arousal, dominance),* dimension category *and SD* category *(e.g., high, medium, low),* stimulus type *(e.g., images and words), or* gender *(e.g., female, male)]?* and *Q2.2: What is the prevalence of reliable emotional stimuli?*. It was the first study to this day aiming to investigate and untangle the influence of various factors (e.g., dimension, dimension/*SD* category, stimulus type) determining stimulus reliability in a typical psychology study of adults, and hence aimed to provide useful data to researchers choosing stimuli from existing sets.

The results suggest several points. Firstly, in relation to the factor *dimension*, valence ratings seem to remain similar to the original published normative ratings throughout a time span of up to 22 years, while arousal and dominance ratings are less reliable. The observed effects were similar for female and male participants. In other words, reliability of arousal and dominance ratings may reduce over time, putting any results of studies relying on these normative ratings at stake of being misinterpreted. Second, a separation into low, medium, and high *SD categories* appears helpful for the selection of stimuli today, however only in relation to valence. That is, for valence, the *SD* of stimuli assessed today correlates with medium to high strength with the normative *SD* and researchers may therefore rely on the normative *SD* category (low, medium, high) when selecting stimuli. The separation of stimuli based on *dimension category*, however, seems redundant. That is, stimuli allocated to the individual categories (e.g., low, medium, high) based on the normative rating display a weak to medium mean correlation between normative and idiographic rating, with only a low percentage of stimuli displaying significant correlations to normative rating when assessed today. This does not mean that researchers should refrain from allocating stimuli into low/medium/high dimension categories, however, that the selection of stimuli from existing sets based on these dimension categories may lead to a selection of unreliable stimuli, as when assessed today stimuli may not fall into the same dimension category. This may be particularly important for researchers who aim to use stimuli of low arousal in order to avoid any physiological effects caused by arousing content (e.g., heart rate, skin temperature, skin conductance amplitude).

Comparison of *stimulus types* revealed a slightly greater reliability of image compared to word stimuli for valence, while correlations of idiographic to normative ratings were slightly lower for image compared to word stimuli on dominance and arousal (except female dominance ratings), indicating, despite a generally low correlation, a (slightly) greater stability of word compared to image stimuli on latter two dimensions. These results could therefore support the line of argument mentioned earlier, that word stimuli are less distinct in their content compared to image stimuli and that a constant "update" of the inner-eye-representation within each individual may hence represent less contrast to the surrounding (e.g., culture, technological development, or societal norms). Finally, the possibility that the recombination of stimuli selected from various individual stimuli sets could have affected stimulus perception, highlights the complexity of the factors that can influence ES perception. Additional research is needed investigating if present findings are also applicable to other stimulus types (e.g., video or audio clips), as well as assessed dimensions (e.g., concreteness, emotion intensity) or specific emotions (e.g., happiness, sadness, anger).

In conclusion, the present results displayed an influence of several factors onto the reliability of stimulus ratings. Hence, the two principal questions can be answered as follows: "*Various factors such as dimension, dimension category, SD category, stimulus type, as well as gender, all influence the reliability of stimuli. The prevalence of reliable stimuli varies along with the sub-category of factors (e.g., low, medium, high for dimension category or SD category; valence, arousal, dominance for dimension).*" Unless researchers are using stimuli in relation to valence ratings as well as stimuli rated for valence with medium *SD*, researchers are highly advised to always reassess stimuli prior to study conduction to ensure reliability of stimuli for their study.

Chapter Four – The Effect of Emotional Stimuli on Recognition Memory in Dependence of Personal Sensitivity

The previous study described in *Chapter Three* has shown that the validity of the rating data provided along with the stimuli may vary in relation to factors such as assessed dimension or dimension category (e.g., *low*/*medium*/*high*). Additionally, previous research suggests that characteristics that are distinct to the assessor (e.g., age, or ethnicity) affect emotion perception. Simultaneously, a large area of research within the field of emotion, regards the relationship between emotion and memory, with numerous study findings suggesting an effect of emotion on memory. If emotion perception is affected by assessor characteristics, while both emotion and memory play an important role in our everyday life, it is important to understand the relation between these three factors (e.g., assessor's characteristic, emotion perception, memory), as additional insight would be able to be directly implemented in learning contexts such schools, or also therapeutic contexts (e.g., trauma therapy). Therefore, the third and final study investigated the influence of an assessor's characteristic (the perceiver's emotional sensitivity) onto the perception of ES as well as recognition memory. An experimental study was conducted, seeking to gain insight into the final research question that was formulated as follows:

> *Q3: What is the relationship between trait sensitivity, emotion perception of stimuli, and recognition memory?*

First, a brief overview will be provided regarding sensory processing sensitivity (SPS), its relation to emotion perception, and recognition memory with involved brain areas, followed by the description and discussion of the conducted study.

Introduction

As described in *Chapter Three*, many factors may affect perception of ES. Among these are perceiver's characteristics such as age (Isaacowitz et al., 2007), gender (Lithari et al., 2010; Nater et al., 2006), and ethnicity (DeBusk & Austin, 2011). Next to these demographic characteristics, however, various researchers have investigated the relationship between emotion perception and personality traits (e.g., Druschel & Sherman, 1999; Vuoskoski & Eerola, 2011; Galea & Lindell, 2016). Rather than being stable, trait personality has been found to change throughout an individual's life (Roberts, 2009) and may be affected by culture (Roberts & Helson, 1997), age (Costa & McCrae, 2006; Roberts et al., 2006), adverse life events such as job loss (Anger et al., 2017), or intentional intervention (Costa & McCrae, 2006).

Personality and Emotion Processing

Usually applied within therapeutic settings but also used by employers (e.g., assessment centers), numerous tests have been created for the assessment of personality traits: examples are the *HEXACO Personality Inventory* (Lee & Ashton, 2004), the Myers-Briggs Type Indicator (Myers, 1962), the *Minnesota Multiphasic Personality Inventory* (Graham, 1987), or the *Revised NEO Personality Inventory* (Costa & McCrae, 1992). The latter – also well-known as the Big Five Inventory (*NEO-FFI* or *NEO-PI-R*) – is a frequently applied test measuring extraversion, agreeableness, conscientiousness, neuroticism, and openness (Costa and McCrae, 1989; 1992). In that regard, research has shown a strong relation between the neuroticism trait and negative affect, as well as extraversion traits and positive affect (Costa & McCrae, 1980; Eaton & Funder, 2001; Hermes et al., 2011; Markon et al., 2005; Tamir & Robinson, 2004). Individuals high in extraversion are characterized by activity, friendliness, warmth, and positive emotions; agreeableness refers to altruism, trustworthiness, and modesty; individuals scoring high on conscientiousness are reliable, hard-working, and deliberate; neuroticism is characterized by vulnerability, anxiousness, and impulsivity; and individuals with high openness have vivid imagination, prefer variety, and love art (McCrae & Costa, 1989).

Individuals who are high in emotional intelligence are skilled at expressing and regulating their emotions (Salovey, 2001), and although the ability to perceive emotions in others can be trained, hence improving emotional intelligence (e.g., Nelis et al., 2009), the effectiveness of training is moderated by personality traits (Herpertz et al., 2016): For example, individuals who are high in agreeableness or conscientiousness benefit more from a training intervention than individuals who are low on these personality traits. Research investigating emotion perception and personality traits has shown a relation between higher accuracy in facial emotion recognition and high scores on extraversion (e.g., Li et al., 2010), conscientiousness, and openness (e.g., Matsumoto et al., 2000). While the score of agreeableness does not seem to have an effect onto accuracy of facial and vocal emotion recognition (e.g., Mill et al., 2009), a high score of neuroticism is related to the difficulty in correctly identifying happy faces (e.g., Andric et al., 2016). Perlman et al., (2009), additionally displayed a significant positive correlation between the level of neuroticism and fixation duration of eyes and mouth of fearful, happy, and sad faces, with moreover, significantly higher correlations for fearful faces compared to happy and sad. Moreover, Vuoskoski and Eerola (2011) displayed correlations between personality traits and distinct emotion perception of musical stimuli: results of their study showed a positive correlation between neuroticism and sadness ratings, as well as a negative correlation between sadness ratings and extraversion. Similarly, Nater et al., (2005)

display a relationship between sensation seeking personality (Zuckerman, 1979) with higher sensation seeking scores being related to a higher state of arousal in response to slow and peaceful music and higher calmness after exposure to fast and arousing (heavy metal) music stimuli. Additionally, a study conducted by Segerstrom (2001), found slower skin conductance latency in highly optimistic people compared to pessimistic participants in response to negative word stimuli. In summary, there is a wealth of evidence that personality characteristics influence how individuals attend to and perceive ES.

Sensory Processing Sensitivity

Pertinent to the research of personality is the construct of *sensory processing sensitivity*, as measured by the 27-item Highly Sensitive Person Scale (HSPS) (Aron, 1996b; Aron & Aron, 1997). Early studies estimated that approximately one fifth of the human population can be characterised as highly sensitive persons (Kagan, 1994). However, distinct from established personality factors such as neuroticism or introversion, it describes the ability to process stimuli and information more strongly and deeply than others (Aron 1996c; Aron & Aron, 1997; Aron et al., 2010; Aron et al., 2012). That is, individuals with high (vs. low) sensitivity for instance adopt a strategy of pausing to analyse before acting, leading to increased responsiveness to subtle, environmental, and social stimuli (e.g., loud noise, changes in temperature) compared to non-HPSs (Aron et al., 2012). In other words, individuals with higher sensory processing sensitivity perceive stimuli of lower intensity more easily than individuals that are non-highly sensitive. Research evidence has displayed a relation between high sensitivity and health, romantic relationships/sexual behaviour, or also parenting: For example, HSPs have been found to reporting more self-perceived stress and more frequently reporting symptoms of ill health (Benham, 2006; Evers et al., 2008); being more prone to experiencing psychological distress (Liss et al., 2006) and stress (Gerstenberg, 2012); being less interested in variety and with fewer bad experiences regarding sexual behaviour (Aron, 2001); and mothers have been found to perceiving home as more chaotic (Wachs, 2013), and parenting as more difficult (Aron et al., 2019).

In line with former research results displaying differences in personality with regards to emotion perception, Jagiellowicz and colleagues (2016) investigated the relation between SPS and emotion perception. In their study, individuals scoring high (vs. low) on SPS perceived positive images as more arousing. Evidence of neural differences associated with high sensory processing sensitivity has shown heightened activation of specific brain areas during change-detection (high-order visual processing) (Jagiellowicz et al., 2011), viewing of positive and negative images (Jagiellowicz et al., 2016), as well as to both, sad and happy emotional states

of others (Acevedo et al., 2014). Surprisingly, however, despite almost 20 % of the population being highly sensitive (Kagan, 1994), the only existing study investigating the relation between assessed emotional stimulus rating and person sensitivity has been conducted by Jagiellowicz et al., (2016) (see above). A recent study conducted by Williams and colleagues (2021) was able to display an association between SPS and the recognition of degraded words (presented in audio), however, used stimuli had not been categorized for emotional value. In conclusion, this means that if indeed emotion perception of stimuli significantly varies in dependence of SPS (e.g., high SPS vs. low SPS), however, researchers do not differentiate between these two groups in their study (e.g., by adjusting the choice of stimuli to the groups, or reassessing stimuli prior to study conduction), researchers risk basing to base their study on (partially) unreliable stimuli and hence creating misleading study results.

The Relation Between Emotion and Memory

Besides an interest in emotion processing in relation to personality characteristics, a large field of interest within the area of emotion research regards the relation between emotion and memory: Research evidence suggests that rather than being a single construct, memory constitutes a number of systems (e.g., Cohen, 1984; Tulving, 1972; Squire, 1992) with different neural substrates (Wood et al., 1980). That is, distinct brain areas are involved for different types of memories: hippocampus, neocortex and amygdala are for instance three areas of the brain involved in declarative memory (Eichenbaum, 2001; Squire & Zola, 1996; Adolphs et al., 2001); the basal ganglia and the cerebellum are involved in procedural memory (Fabbro, 1999); and the prefrontal cortex is involved in complex cognitive functions necessary for the working memory (Curtis & D'Esposito, 2003). Brain areas involved in the process of memory formation through long-term potentiation are among others based on the hippocampal formation (Scoville & Milner, 1957; Penfield & Milner, 1958), but also the amygdala (Maren, 1999). Simultaneously, the amygdala also plays a central role in the processing of emotional memory content, by modulating the influence of stress on memory processes (Roozendaal, 2002, 2003; Ferry & McGaugh, 2000). This may occur in processing of social and emotional stimuli such as pictures and scenes (Hariri et al., 2002; Norris et al., 2004), or the recognition of emotional facial expressions (Adolphs et al., 1998). Similarly, next to its role regarding cognitive functions, parts of the hippocampus have been shown to also relate to stress, emotion, and affect (Fanselow & Dong, 2010).

As multiple brain structures such as the hippocampus or the amygdala are implicated in both cognition and emotion, the relation and interdependence of both, has been of great interest to researchers (see Barkus et al., 2010). Prima facie, events associated to stronger emotions

seem to be more easily remembered: For example, Brown and Kulik (1977), demonstrated that emotionally arousing events (such as the assassination of J. F. Kennedy) resulted in lively and detailed memory (flashbulb memories) for one's own circumstances at the time individuals had learned about the event. This type of memory, however, may be confined to negative, rather than highly positive events (Kraha et al., 2014), and, shaped by evolution. The ability to remember negative stimuli may enhance their recognition and hence their avoidance in the future, which in turn promotes survival. An enhanced as well as more detailed memory for negative compared to neutral word stimuli was displayed by a series of experiments conducted by Kensinger and Corkin (2003). The authors additionally found a relative contribution of both valence and arousal increasing memory vividness (with a greater effect for arousal). Moreover, Christianson and Loftus (1991) were able to show that memory for emotional events is much better regarding central compared to peripheral details. This finding is in agreement with the Easterbrook hypothesis (Easterbrook, 1959) proposing a narrowing of attention range with increasing arousal. In turn, this suggests that emotional events automatically attract attention and are hence processed more elaborately which improves memory for central information.

Various types of stimuli such as audio stimuli (see Eerola & Vuoskoski, (2012) for review), images (e.g., Pollatos et al., 2007), and video stimuli (e.g., Schaefer et al., 2010) have been shown to successfully induce emotions. This in turn suggests that emotions triggered through ES may also affect memory. Yet, overall, research findings concerning the relation between memory and emotion are inconclusive. Some research investigating the effect of stress on memory has for instance shown that induced stress impairs declarative memory (Kirschbaum et al., 1996; Schwabe & Wolf, 2010), while other studies have shown that stress immediately after learning improved memory for emotional content (Cahill et al., 2003; Wolf, 2008). Moreover, individual studies have shown a negative effect of emotions on memory for negative emotions such anxiety (Harris, 1999; Harris & Cumming, 2003), or sadness (Chepenik et al., 2007), relatively neutral emotions such as boredom (Goldberg & Todman, 2018), or also positive emotions such as happiness (Storebeck & Clore, 2005). In this regard, the neurobiological perspective suggests that mood valence and arousal, both independently of one another, may modulate memory performance with independent neural areas supporting the influence of arousal and emotional valence on memory. Particularly, memory for valenced information may be supported by a prefrontal-hippocampal network, while memory for arousing items may rely on a neural network involving the amygdala and the hippocampus (Isen et al., 1985; Kensinger & Corkin, 2004). Moreover, Gray (2001) was able to show that spacial and verbal performance are influenced oppositely by emotional states of approach (e.g.,

amusement) and withdrawal (e.g., anxiety): The conducted study displayed an approach state impairing spacial, however, improving verbal performance, and a withdrawal state improving spatial, however, impairing verbal performance. These results suggest a selective modulation of emotion on components of cognitive control. As a fact, the latter findings could indeed provide an explanation for inconclusive or even contradictory study results among existing research.

Aiming to gain a more comprehensive understanding, several meta-analyses investigating the relationship between emotion and memory have been conducted. Nevertheless, these are often very specific within this research area. For example, Murphy and Issacowitz (2008) conducted a meta-analysis regarding memory and attention tasks comparing older and younger adults, concluding that age significantly affected the effects for emotion salience, and that the measurement type appeared to influence the magnitude of effect. Similarly, Mather (2007) acknowledges the presence of contradictory research findings regarding the relation between arousal and memory binding within the field of emotion and memory. Concluding her meta-analysis the author therefore proposes an object-based framework to explain existing contradictory findings. According to this framework, the attention-grabbing nature of an object in visual stimuli is interfering with the working memory and make hence make it more difficult to remember other bound representations.

In conclusion, contradictory research findings concerning the relation between memory and emotion exist and could be highly dependent on the design of the study (e.g., cognitive task, the use of real-life events as material being retrieved, sample size, or participants' age) (Ucros, 1989).

Recognition Memory

One possible way to investigate the relation between emotion and memory is through the assessment of recognition of stimuli previously displayed to participants. *Recognition Memory* hence refers to the process of identifying previously encountered items as studied. This involves achieving a match between information encoded at the time of learning and the information available at the time of retrieval (Tulving, 1983), and may encompass recognition of instances as from a particular class or category (e.g., recognition of a face as a *face*) as well as episodic recognition (e.g., recognition of a face previously encountered in a crowd). In experimental psychology, recognition may be assessed through response accuracy and/or latency in tasks that require subjects to discriminate items encountered in a study phase from new items. Examples are recognition of words from a word-list (e.g., Ratcliff & Murdock, 1976; Kinsbourne & George, 1974), faces (e.g., Harmon, 1973; Fagan, 1972), auditory stimuli (Cohen

et al., 2009) or images (Wichmann et al., 2002). Manipulation of memory load (e.g., number of displayed stimuli), time interval between encoding and repeated exposure, participant concentration (e.g., distraction through noise or additional visual stimuli), or brain physiology (e.g., surgical removal of specific brain areas) have helped researchers gain deeper insight into the episodic recognition memory. Early research conducted by Ratcliff and Murdock (1976) for example has shown that recognition latency increases with increasing stimulus list length and that recognition accuracy increases with decreasing stimulus presentation rate. Harmon (1973) was able to show that visual masking affects stimulus recognition of faces with increased blurring decreasing recognition. However, this was moderated by low-frequency features such as head shape, neck and shoulder geometry and gross hair line (achieved through extreme blurring and removement of facial features) that are sufficient for rather good recognition (e.g., recognition rate of almost 60 %). Cohen et al., (2009) compared recognition memory of audio stimuli and pictures and found auditory recognition memory performance being markedly inferior to visual recognition performance in short-term recognition. In a series of experiments, Wichmann et al., (2002) were able to display an advantage of recognition memory for colored vs. black-and-white image stimuli, as well as an effect of color as a surface property being part of the memory representation. Moreover, Jonesgotman and Zatorre (1993) tested recognition for odor in relation to cerebral excision (unilateral cerebral excision from temporal, frontal, frontotemporal, or centroparietal areas) and showed impaired recognition for individuals with excision from right temporal or right orbitofrontal cortex, hence suggesting the importance of these brain areas for odor memory. All these given examples provide yet a small insight into examples of studies that have investigated recognition memory by using emotional stimuli.

Research Rationale

Emotional stimuli form an important tool within emotion research, and numerous studies investigating the relation between emotion and memory have been conducted in the past. However, as mentioned above, findings are inconclusive, with some study results suggesting memory enhancement and others suggesting memory impediment through emotions (e.g., Cahill et al., 2003; Chepenik et al., 2007; Goldberg & Todman, 2018; Harris, 1999; Harris & Cumming, 2003; Kirschbaum et al., 1996; Schwabe & Wolf, 2010; Storebeck & Clore, 2005; Wolf, 2008). Simultaneously, Aron et al., (2012) suggest a deeper processing of emotional information by HSPs compared to non-HSPs. Deeper processing in turn leads to a more emotional perception of pictures, as well as enhanced concentration and thus promotes long-term retention of stimuli (Soravia et al., 2016). In conclusion, existing evidence suggests that emotion perception, may be affected by trait sensitivity (Jagiellowicz, 2012). With regards to

emotion perception of stimuli between HSPs and non-HSPs, a difference in perception may be expected leading to more extreme valence and arousal ratings by HSPs (vs. non-HSPs). This poses a great risk on the validity of the normative rating data of ES, especially when considering that high sensitivity concerns approximately 20 % of the human population (Kagan, 1994). In fact, not controlling for trait sensitivity when using ES in a study, may be an important reason for existing inconclusive findings within the research field of emotion and memory, as existing normative rating data provided along with ES may not have been valid for studies including HSPs.

Therefore, it is important to investigate the relation between sensory processing sensitivity and perception of ES; the investigation of the relation between sensory processing sensitivity and recognition memory, may help to shed light onto the existing inconclusive research results. To date, no research has investigated the relation between trait sensitivity, emotion perception and recognition memory. Therefore, in the present study, two groups differing in SPS will be compared regarding emotion perception (e.g., valence, arousal) of emotional image stimuli. Moreover, recognition of stimuli will be assessed and compared between the two participant groups, as well as compared across stimuli regarding assessed rating.

Based on previous findings outlined above, the following hypotheses were formulated:

(H1a): Image valence will be perceived as more extreme (more negative for negative stimuli, and more positive for positive stimuli) by participants of higher, compared to lower sensitivity.

(H1b): Image stimuli will be perceived as more arousing by participants of higher, compared to lower sensitivity.

and

(H2a): Recognition will significantly differ between stimuli that were perceived as positive and negative compared neutral stimuli.

(H2b): Recognition will significantly differ between stimuli that were perceived as extremely arousing compared to stimuli perceived as low arousing.

The *Levels of Processing model* proposes that memory is dependent on the depth of processing (Craik & Lockhart, 1972). From this perspective, Aaron and colleagues (2012) suggested that HSPs process emotional information to a deeper level. However, it is unclear whether a deeper processing may indeed lead to an overall better memory of stimuli for HSPs (vs. non-HSPs), or whether an increased perceived emotionality may interact with the effect of emotion onto memory, which in turn could more strongly enhance memory for positive and

impeded memory for negative stimuli in HSPs compared to non-HSPs. Due to contradictory findings suggesting both, enhancement as well as impairment of memory through emotion, no direction of difference will be formulated regarding recognition memory between participants of *higher* compared to *lower* sensitivity. Therefore, the following hypothesis was formulated: *(H3): Recognition memory will significantly differ between participants of higher compared to lower sensitivity.*

Among existing types of stimuli (e.g., images, audio clips, video clips, or words), images were chosen to further investigate this type of stimuli and to be consistent with the previous study (see *Chapter 3*), as well as to minimize previous contact with stimuli (e.g., familiarity of word stimuli). To calculate the number of participants required in these studies, a power analysis was conducted based on comparable previously reported effect sizes taken from Hostler et al., (2018). To detect the smallest observed effect of e.g., positive vs. neutral cues (d = 0.32), a sample size of N = 80 is required.

Method

Participants

A total of 101 participants (47 female, 54 male; 18-69 years old, mean age: 29.13 years, *SD* = 10.73) completed the survey. Of these, 23 participants were excluded: 8 because they did not pass at least 50% of the attention checks, and 15 because they did not complete the second part of the survey. All analyses are based on the remaining 78 participants (34 female, 44 male) who were between 18 and 69 years old (mean age females: 30.94 years, *SD* = 13.44; mean age males: 29.07 years, *SD* = 9.73). Participants were recruited via the webpage Prolific, through advertisements around university as well as word-of-mouth and were paid, or received University Credit Points for their participation. The participation criteria within Prolific were set through filters; these were: a minimum age of 18 years, no medical history, normal/corrected-to-normal vision, as well as fluency of the English language. Participants were from Austria (*n* = 1), Belgium (*n* = 2), Germany (*n* = 1), Greece (*n* = 7), India (*n* = 1), Israel (*n* = 1), Italy (*n* = 8), Turkey (*n* = 1), Peru (*n* = 1), Poland (*n* = 12), Portugal (*n* = 11), Singapore (*n* = 1), South Africa (*n* = 2), Spain (*n* = 4), United Kingdom (*n* = 20), United States of America (*n* = 3), Vietnam (*n* = 1), and Zimbabwe (*n* = 1). All participants were fluent in English, without a history of mental disorders, and had normal or corrected-to normal vision.

Procedure and Materials

Procedure

The current study was set up as a two-part survey taking place online with a one-week interval between Part 1 and Part 2. Prior to the study, participants were informed about the

broad aim of the survey. That is, to avoid effects on memory through deliberate memorization it was not disclosed to participants that the survey would contain a memory task. Confidentiality of personal data was assured, and participants were free to quit participation at any time by closing the web-browser. Names and e-mail contacts of the researchers, as well as contact information in case of distress caused by the content of the survey were provided. After reading the information sheet, participants were asked to tick checkboxes to indicate their agreement to the consent form. Only upon full approval, could participants begin the survey. Demographic data were collected, and participants were asked to carefully read the instructions prior to rating the stimuli. Participants were then introduced to the terms *valence* and *arousal* as well as the rating scales along with an example (see *Appendix A*).

Part 1

Part 1 (*t* 1) consisted of two consecutive blocks: Block 1 containing 100 image stimuli (target stimuli) selected from previously published sets of emotional stimuli (Diconne et al., 2022), and Block 2 containing the *Highly Sensitive Person Scale* (HSPS) (Aron & Aron, 1997). Each image of Block 1 was individually displayed in the top-centre of the computer screen with the rating scales below. Participants were asked to rate each image on the dimensions valence (*"To me this image is… -4 = strongly negative*, to *4 = strongly positive"*) and arousal (*"…and… 0 = not at all arousing*, to *8 = very arousing"*) on 9-point Likert scales (Likert, 1932).[1] Subsequently (Block 2), participants completed the HSPS using 7-point Likert scales (*1 = not at all, 4 = moderately, 7 = extremely*). The rating procedure was self-paced, and no time limit was given. To prevent missing values, scales were set to forced response. Four quality check questions (see similar examples in *Chapter Three*) and an open question *("In your own words, briefly describe the difference between 'valence' and 'arousal'."*) were used in to verify that participants read instructions carefully and paid full attention to the survey.

Following completion of Part 1, participants were informed that an automated message including the weblink to Part 2 of the survey would be sent to them after seven days. Response quality check questions were verified and participants who did not pass at least two (50 %) of the four attention check questions and/or responded incorrectly to the open question were informed so and did not receive an automated message for participation in Part 2 of the survey.

[1] The study was pre-tested with both valence and arousal scales ranging from 0 to 8. However rating responses suggested that participants had difficulties assessing the bidirectional valence (strongly negative – strongly positive) on a unidirectional scale. Therefore, the valence rating scale was changed (from -4 to +4).

Part 2

Part 2 consisted of 200 image stimuli, of which 100 images were the target stimuli from Part 1, and 100 images were distractor stimuli that participants had not seen before. Images were randomized in order and individually displayed at the centre-top of the computer screen. Participants were asked to indicate whether they remember the displayed image from Part 1, completed one week earlier (*"I remember this image from Part 1 of the survey completed one week ago."* - *yes* / *no*) by selecting the button below the image accordingly. Again, this procedure was self-paced, and no time limit was given.

An outline of the study procedure can be found in Figure 8, below.

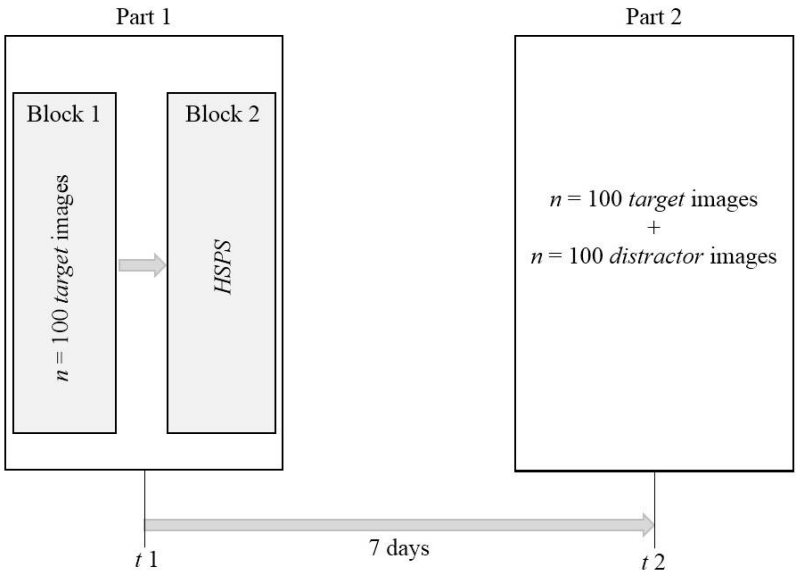


Figure 8. Flow of the study procedure. Order of image stimuli randomized across participants. HSPS = Highly Sensitive Personality Scale (Aron & Aron, 1997).

After completion of the survey, participants were debriefed and given the choice to receive information about study results as well as to enter into a prize draw for a £20 shopping voucher. Students from Manchester Metropolitan University received participation points; participants recruited through Prolific received a participation compensation of £4.30. The study received ethical approval from the Manchester Metropolitan University faculty ethics committee, and data was collected between February and June 2021.

Materials

Similar to the previous study (*Chapter Three*) the included stimuli sets were selected based on the KAPODI database (*Chapter Two*). That is, image stimuli sets were compared regarding their assessment scale, and to ease comparability across sets, only sets originally assessed on 9-point Likert scales were included. Stimuli were chosen from the *Besançon*

Affective Picture Set-Adult (BAPS-Adult; Szymanska et al., 2019), the DIsgust-RelaTed-Images (DIRTI; Haberkamp et al., 2017), the *Military Affective Picture System* (MAPS; Goodman et al., 2016), and the *Nencki Affective Picture Set* (NAPS; Marchewka et al., 2014), which are all freely available emotional image stimuli sets (see *Chapter One*). They will from here on be referred to as *BAPS*, *DIRTI*, *MAPS* and *NAPS*, respectively. N = 25 target stimuli were selected from each set based on normative rating values of valence and arousal provided in the original source: 12 images were selected based on their *valence* rating (*n* = 3: highest rating; *n* = 3: lowest rating; *n* = 3: highest standard deviation (*SD*); *n* = 3: lowest *SD*), and 13 images were selected based on their *arousal* rating (*n* = 3: highest rating; *n* = 3: lowest rating; *n* = 3: highest SD; *n* = 3: lowest SD; *n* = 1: with medium *SD*). For each target image, a distractor image that matched regarding normative rating (e.g., similar or equivalent valence/arousal rating), colour (e.g., grey-scale target image was matched with a grey-scale distractor image) and displayed content (e.g., landscapes were matched with landscapes) was chosen from the same set.

The HSPS shows good psychometric properties (Aron & Aron, 1997) and was found to be reliable in a culturally diverse sample (May et al., 2020). The selection of 23 items of the original 27-item scale was based on Aron (1996a). Items included for example: *"Are you easily overwhelmed by strong sensory input?"* or *"Are you particularly sensitive to the effects of caffeine?"*. The full list of included stimuli (target and distractor stimuli), can be found in the *SM, (Study 3, File L);* the items of the HSPS are included in the *Appendix B*.

Study Design

To answer the above-mentioned research questions, the study was designed as a two-point data collection study. That is, rating data of target stimuli, as well as response of the HSPS was collected first (*t* 1), and recognition of stimuli was assessed seven days later (*t* 1 + 7 days) = *t* 2. Participant's gender (female vs. male), participant's sensitivity (*high* vs. *low*), as well as dimension category (low/medium/high) were the independent variables, and both assessed stimulus rating (*t* 1) as well as assessed recognition (*t* 2) were the dependent variables.

Rating differences regarding the individual stimuli between *low* and *high* SPS as well as between female and male participants were calculated with t-tests; Effects of gender (female/male), SPS level (low/high SPS) and dimension categories (low/medium/high) onto valence and arousal ratings were calculated through 2 × 2 × 3 (gender × SPS level × dimension category) mixed ANOVAs for both dimensions separately. Finally, the effects of gender, SPS level, and dimension category onto stimulus recognition were calculated through 2 × 2 × 3 (gender × SPS level × dimension category) mixed ANOVAs for valence and arousal separately.

A Greenhouse-Geisser correction was implemented where the assumption of sphericity was violated; where effects were significant, post hoc analyses were conducted.

Results

In the current study, 78 participants between the age of 18 and 69 years completed both parts of this two-part survey. Each participant rated each of the 100 target stimuli, completed the HSPS (Part 1), and indicated recognition for 200 image stimuli (Part 2). Completion of both parts of the survey took participants approximately 45 minutes (Part 1 mean: 27.1 minutes, *SD* = 10 minutes; Part 2 mean: 15.9 minutes, *SD* = 5.7 minutes). Raw rating data per participant can be found in the *SM (Study 3, File M)*. Participants scored between 55 and 158 points on the HSPS (mean: 104; *SD* = 19.8) and were separated through median split into two groups (low sensitive person group: *low SPG*, and high sensitive person group: *high SPG*).

Analyses were conducted to investigate the characteristics of the assessed data as well as the effect of the perceived emotion on recognition memory. Due to previous research findings indicating gender differences in perception, data will be reported for female, male, as well as all participants separately. First, the results of the correlations between female and male ratings for each stimulus, as well as between *low* and *high* SPG ratings of each stimulus will be presented for both genders as well as for all participants combined. Second, analyses concerning the effects of the factors gender (female/male), SPS level, and dimension category onto stimuli ratings are presented separately for the dimensions valence and arousal. Finally, recognition (hit rates) for *target* stimuli and analyses concerning the effect of the factors gender, SPS level, and dimension category on hit rates are presented.

Stimulus Rating

Comparison Between Genders

As reported in *Chapter Three*, previous research has displayed differences between genders regarding emotion perception of stimuli (e.g., Kemp et al., 2004; Kuypers, 2017; Memon et al., 2019), therefore, valence and arousal means and *SD* were calculated for all, as well as female and male participants separately. The analyses of female compared to male ratings were conducted for this study again (as in *Chapter Three*) and included to verify if the findings in relation to the current stimulus selection would support or contradict previous study findings reporting gender differences. Moreover, an absolute rating difference score between female and male participants was calculated indicating the direction of rating difference. For each stimulus, a correlation between mean female and male valence and arousal ratings was then calculated indicating stimuli that are perceived significantly differently by both genders (see *Supplementary Material, Study 3, Table N*). Stimulus ratings of female and male

participants strongly correlated on both valence (r = .99; p < .001), and arousal (r = .94; p < .001); rating differed significantly (p < .05) between both genders for n = 9 stimuli for valence, and n = 1 stimulus for arousal.

Comparison Between High and Low Sensitive Person Groups

As mentioned previously, two groups (*low* and *high* SPG) were created through a median split based on participants' HSPS score. Despite the risk of possibly increasing Type II errors, creation of groups through median split is a commonly implemented and acceptable approach (see Iacobucci et al., 2015). Based on gender (female, male, all) × sensitivity (*high* and *low* SPG), six groups were created (see Table 22).

Table 22

Mean HSPS Score per Group

HSPS Group	Mean HSPS Score (*SD*)		
	females	*males*	*all*
low SPG	92.06 (12.43)	85.50 (11.00)	88.15 (11.64)
high SPG	125.88 (13.07)	114.82 (9.05)	119.85 (11.92)

Note. HSPS = Highly Sensitive Person Scale; SPG = sensitive person group. Females: n = 17 per group; males: n = 22 per group; all: n = 39 per group.

Means and standard deviations (*SD*) of stimuli ratings for all six groups were calculated for each stimulus and for both assessed dimensions (valence, arousal) (see *SM, Study 3, File N*).

An independent-samples t-test was conducted to investigate rating differences between the respective *high* and *low* SPGs for each individual stimulus regarding valence and arousal. A Levenes-test for equality of variances was conducted and Welch-correction implemented for all stimuli with variance-heterogeneity between groups. Results (t, p, df) for each stimulus and assessed dimension can be found separately for female, male and all participants in the *SM, Study 3, Table P*. Among the 100 assessed stimuli, valence ratings differed significantly (p < .05), between *high* and *low* SPGs for n = 15 (females), n = 9 stimuli (males), and n = 17 (all). Arousal ratings between these two groups differed significantly for n = 15 (females), n = 1 stimuli (males), and n = 8 (all).

The Effects of the Factors Gender, SPS Level, and Dimension Category on Stimulus Rating

Valence

To allow more in-depth analyses regarding the valence ratings, this dimension was separated into three dimension categories, namely *low*, *medium*, and *high* valence. The cut-off

scores for the categories were: -4 to -1.34 = *low*, -1.33 to 1.33 = *medium*, 1.34 to 4 *high.* There were n = 30 stimuli of *low* valence, n = 45 stimuli of *medium* valence, and n = 25 of *high* valence. An exact stimulus allocation based on the idiographic mean rating from all participants can be found in the *SM*, *Study 3, File Q*. The mean valence ratings for female and male participants can be found in Figure 9.

The 2 × 2 × 3 (gender × SPS level × dimension category) mixed ANOVAs revealed that there was a significant main effect of dimension category ($F(2, 148) = 1490.45$, $p < .001$, partial $\eta^2 = .95$), and SPS level ($F(1, 74) = 7.98$, $p = .01$, partial $\eta^2 = .10$), however, no main effect of gender ($F(1, 74) = 0.07$, $p = .79$, partial $\eta^2 = .001$) in relation to the valence rating. Interactions were significant for dimension category × gender ($F(2, 148) = 6.17$, $p = .003$, partial $\eta^2 = .08$), however not for dimension category × SPS level ($F(2, 148) = 0.94$, $p = .39$, partial $\eta^2 = .01$). The 3-way interaction of dimension category × SPS level × gender also remained non-significant ($F(2, 148) = 0.52$, $p = .60$, partial $\eta^2 = .01$).

Further analyses regarding the interaction between gender and dimension category revealed a significant difference between female and male ratings regarding *high*-valence stimuli ($F(1 , 76) = 0.51$; $p = .05$; mean difference = 0.34). Differences remained non-significant for stimuli of low ($F(1, 76) = 0.07$; $p = .06$; mean difference = -0.24), as well as medium ($F(1, 76) = 0.01$; $p = .88$; mean difference = -0.02) valence.

A pairwise comparison conducted to investigate the main effect of SPS level revealed lower valence ratings from the *high* SPG compared to the *low* SPGs ($p = .005$), (see Figure 10). Both negative and positive stimuli were perceived as more negative by *high* SPGs (vs. *low* SPGs).

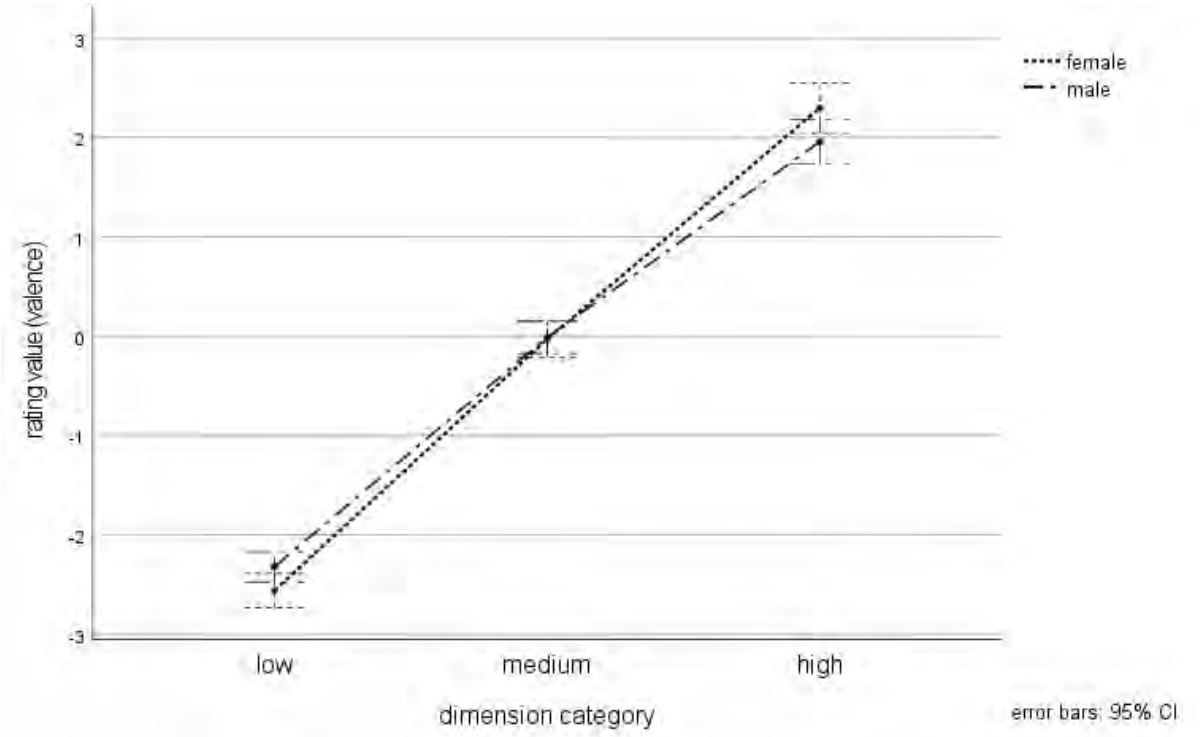


Figure 9. Mean valence ratings of females and males for stimuli of *low*, *medium*, and *high* valence. CI = confidence interval.

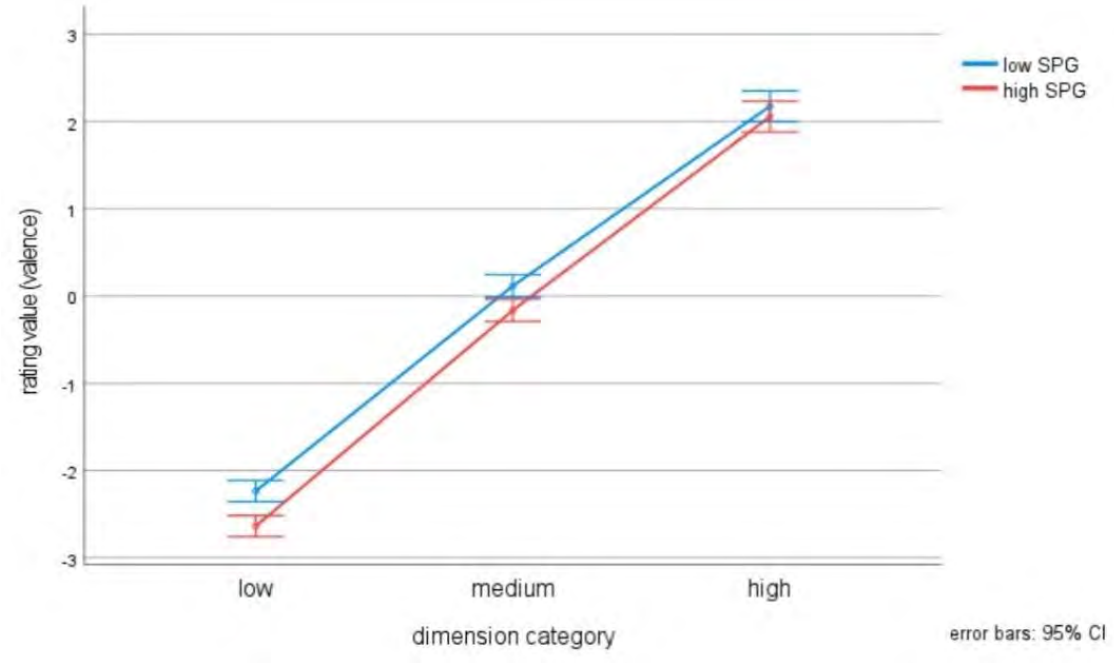


Figure 10. Mean valence ratings of *low* and *high* sensitive person groups (SPG) for stimuli of *low*, *medium*, and *high* valence. CI = confidence interval.

Arousal

Like valence, the dimension arousal was also separated into three dimension categories (*low*, *medium*, and *high* arousal). The cut-off scores for the categories were: 1 to 3.66 = *low*,

3.67 to 6.33 = *medium*, 6,34 to 9 = *high* arousal. There were n = 55 stimuli of *low* arousal, n = 45 stimuli of *medium* arousal, and no stimuli falling into the *high*-arousal stimulus group. An exact stimulus allocation based on the idiographic mean rating from all participants can be found in the *SM*, *Study 3, File Q*. The mean arousal ratings for female and male participants can be found in Figure 11, below.

The 2 × 2 × 3 (gender × SPS level × dimension category) mixed ANOVAs revealed that there was a significant main effect of dimension category ($F(1, 74) = 564.60$, $p < .001$, partial $\eta^2 = .88$), however, no main effect of SPS level ($F(1, 74) = 2.30$, $p = .13$, partial $\eta^2 = .03$) or gender ($F(1, 74) = 0.25$, $p = .62$, partial $\eta^2 = .003$) in relation to the arousal rating. Interactions remained non-significant for dimension category × gender ($F(1, 74) = 3.63$, $p = .06$, partial $\eta^2 = .05$), as well as for dimension category × SPS level ($F(1, 74) = 0.24$, $p = .63$, partial $\eta^2 = .003$). The 3-way interaction of dimension category × SPS level × gender also remained non-significant ($F(1, 74) = 0.21$, $p = .65$, partial $\eta^2 = .003$).

Pairwise comparisons regarding the main effect of dimension category revealed that stimuli of *low* arousal received a significantly lower rating compared to stimuli of *medium* arousal ($p < .001$), (see Figure 12).

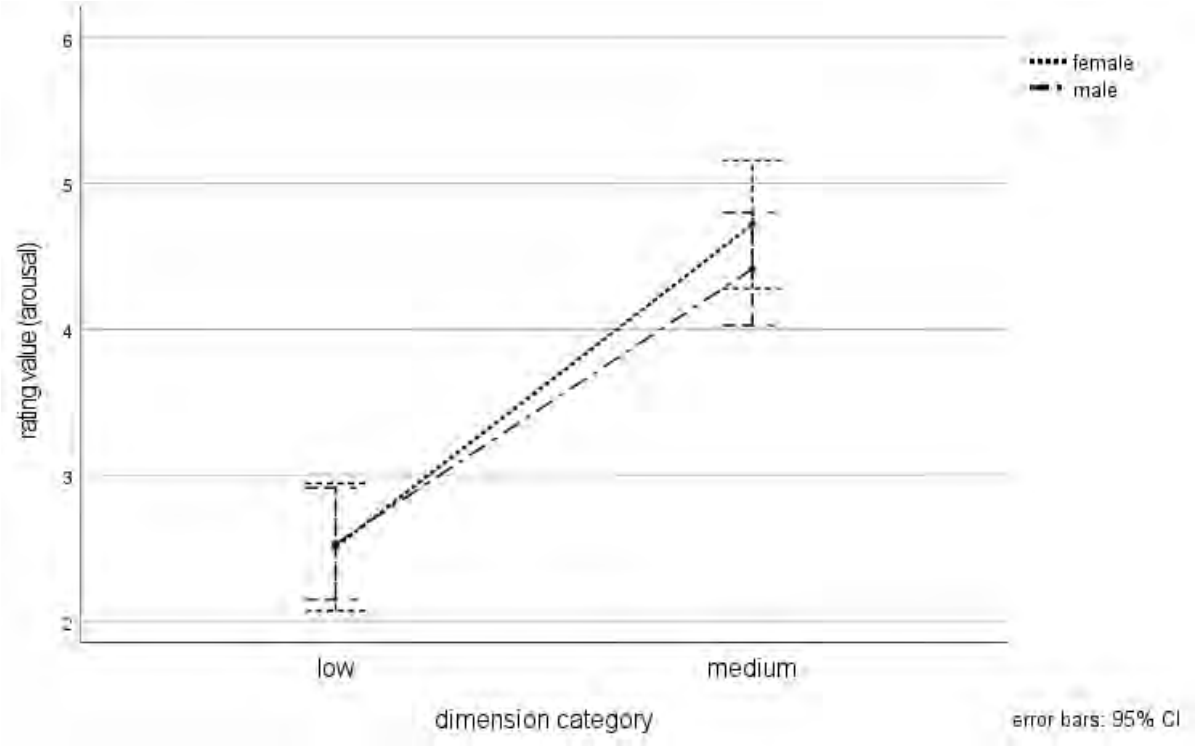


Figure 11. Mean arousal ratings of females and males for stimuli of *low* and *medium* arousal. CI = confidence interval. Note that there were no stimuli falling into the *high* arousal category, as this data is based on the idiographic rating data.

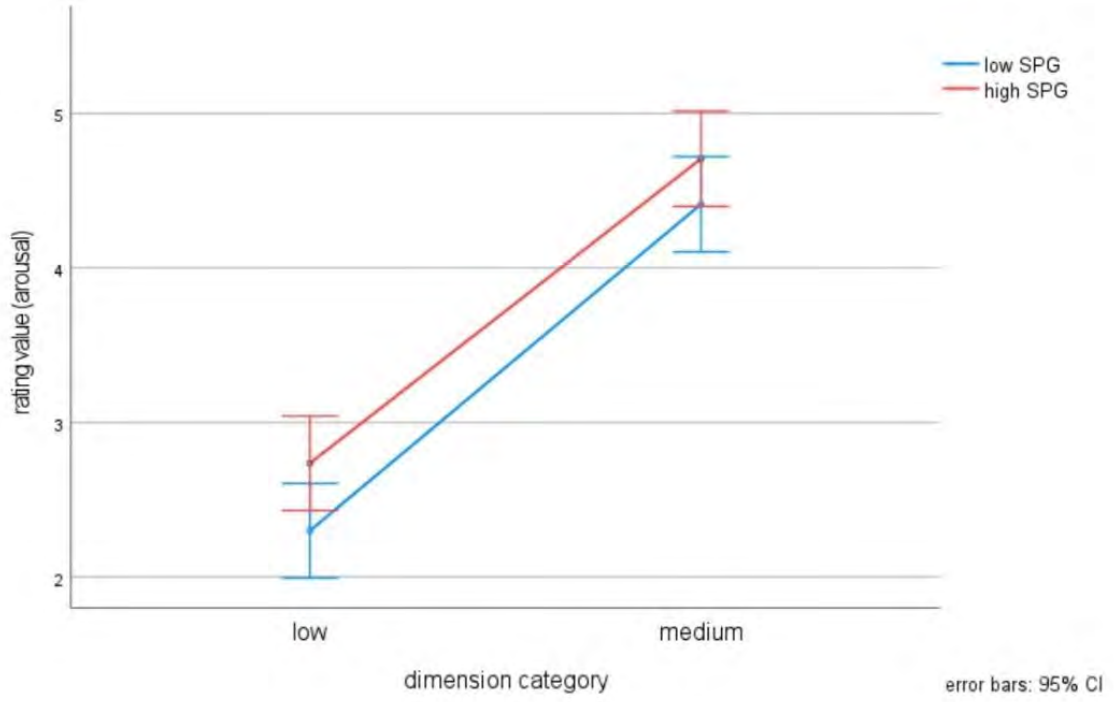


Figure 12. Mean arousal ratings of *low* and *high* sensitive person groups (SPG) for stimuli of *low* and *medium* arousal. CI = confidence interval. Note that there were no stimuli falling into the *high* arousal category as this data is based on the idiographic rating data.

Stimulus Recognition

Prior to analyzing stimulus recognition in relation to stimulus rating, recognition of target stimuli (hits) as well as recognition of distractor stimuli (false positive) were calculated for each of the n = 200 stimuli (see *SM, Study 3, File R*). Target stimuli were recognized by 47.44 % to 97.44 % (mean: 75.19 %) of the participants, while participants recognized between 39-99 % (mean: 75.19 %) of the target stimuli. Separated by *high* and *low* SPGs, hit means were 75.87 % and 74.51 % (all), 73.82 % and 77 % (females), as well as 76.91 % and 73.14 % for males, respectively.

To investigate the relation between perception (e.g., valence and arousal rating) and recognition, calculations had to be made for each individual. That is, mean hit rates with respect to each dimension category (*low*/*medium*/*high* valence/arousal) were calculated for each participant.

The 2 × 2 × 3 (gender × SPS level × dimension category) ANOVA revealed that there was a significant main effect of dimension category ($F(2, 148) = 13.48$, $p < .001$, partial $\eta^2 = .15$), however, no main effect of SPS level ($F(1, 74) = 0.14$, $p = .71$, partial $\eta^2 = .002$) or gender ($F(1, 74) = 0.09$, $p = .76$, partial $\eta^2 = .001$) regarding hit rates in relation to valence. Interactions remained non-significant for dimension category × gender ($F(2, 148) = 0.12$, $p = .88$, partial $\eta^2 = .002$), as well as for dimension category × SPS level ($F(2, 148) = 0.36$, $p = .73$,

partial η^2 = .005) (for mean hit rates in relation to valence and SPS level, see Figure 16, below). There was a significant 3-way interaction of dimension category × SPS level × gender (F(2, 148) = 4.60, p = .01, partial η^2 = .06). Pairwise comparisons regarding this 3-way interaction revealed that SPS level had a significant effect on this interaction for *low*-valence stimuli for females (p = .04) as well as males (p = .03), and regarding stimuli of *medium* valence for females (p = .03).

With regards to arousal, there also was a significant main effect of dimension category (F(2, 144) = 10.31, p < .001, partial η^2 = .13), however, no main effect of gender (F(1, 72) = 0.005, p = .94, partial η^2 < .001) or SPS level (F(1, 74) = 0.002, p = .97, partial η^2 < .001) regarding hit rates. Interactions remained non-significant for dimension category × gender (F(2, 144) = 0.54, p = .59, partial η^2 = .007), as well as for dimension category × SPS level (F(2, 144) = 0.15, p = .86, partial η^2 = .002) (for mean hit rates in relation to arousal and SPS level, see Figure 17, below). There was no significant 3-way interaction of dimension category × SPS level × gender (F(2, 148) = 1.30, p = .28, partial η^2 = .02). Pairwise comparison regarding the main effect of dimension category revealed that stimuli of *low* arousal were recognized significantly less often compared to stimuli of *medium* arousal (p = .003), and compared to stimuli of *high* arousal (p < .001), as well as stimuli of *medium* arousal compared to stimuli of *high* arousal (p = .004), (see Figures 13 and 14, below).

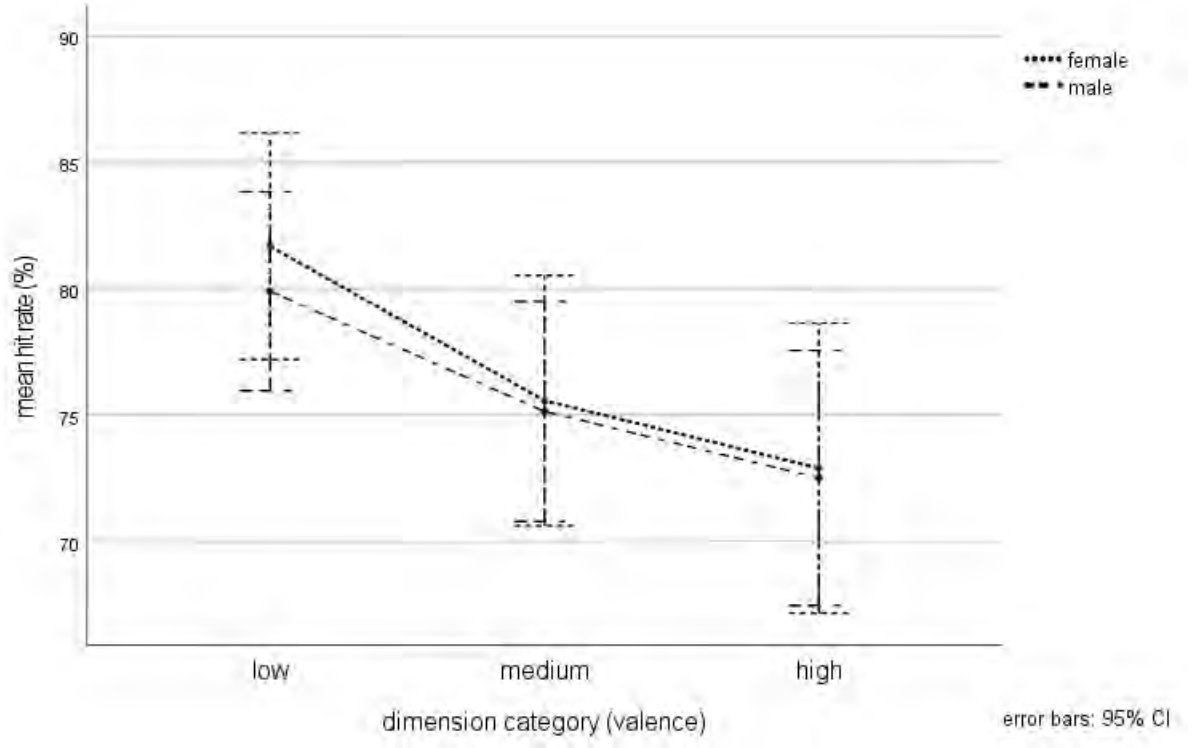


Figure 13. Mean hit rates per valence dimension category (*low*/*medium*/*high*) for females (n = 34) and males (n = 44). CI = confidence interval.

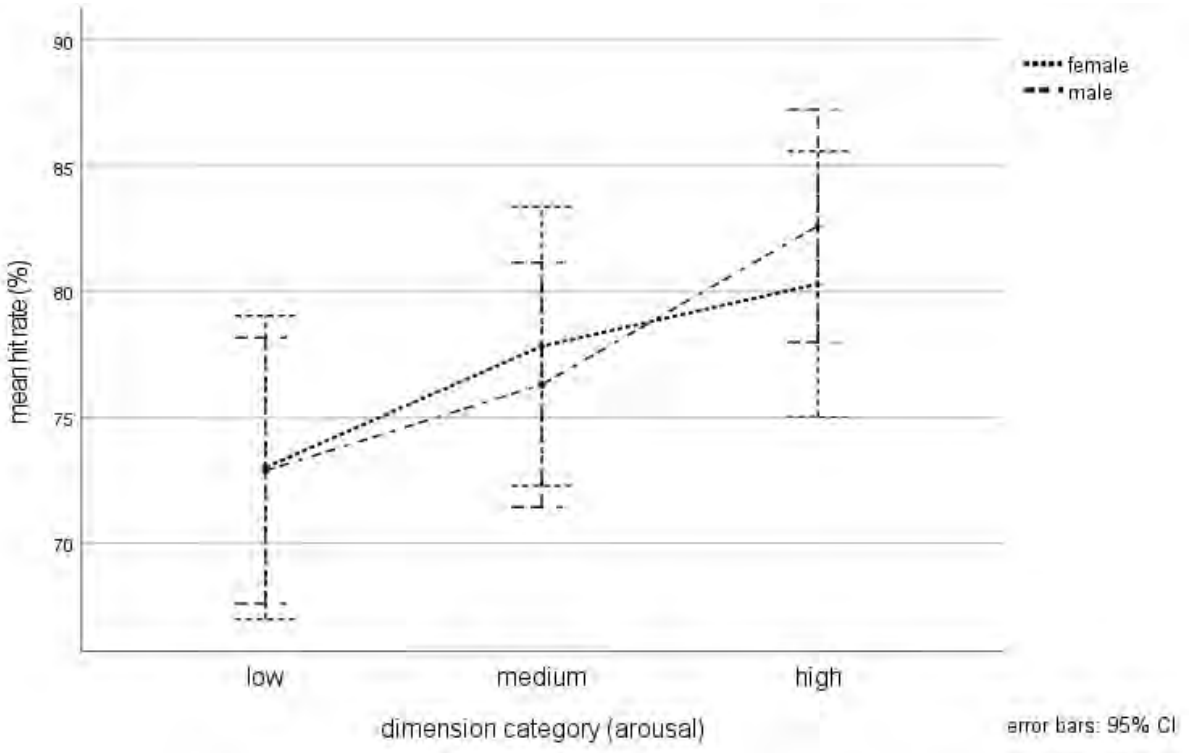


Figure 14. Mean hit rates per arousal dimension category (*low*/*medium*/*high*) for females (n = 34) and males (n = 44). CI = confidence interval.

The hit rate tendency of valence runs contrasting to the hit rate tendency in relation to arousal (see Figure 15). That is, while the hit rate decreases with stimulus valence (e.g., lower hit rates for positive compared to negative stimuli), it increases with stimulus arousal (higher hit rate for stimuli high in arousal compared to stimuli low in arousal). Note that the valence scale (strongly negative – strongly positive) is a bidirectional scale, while arousal (not at all – very arousing) is a one-directional scale. This means that the similar hit rate for the *medium* dimension categories (see Figure 15), refers to medium arousing images, as well as neutrally valent stimuli (the approximate middle between strongly negative and strongly positive).

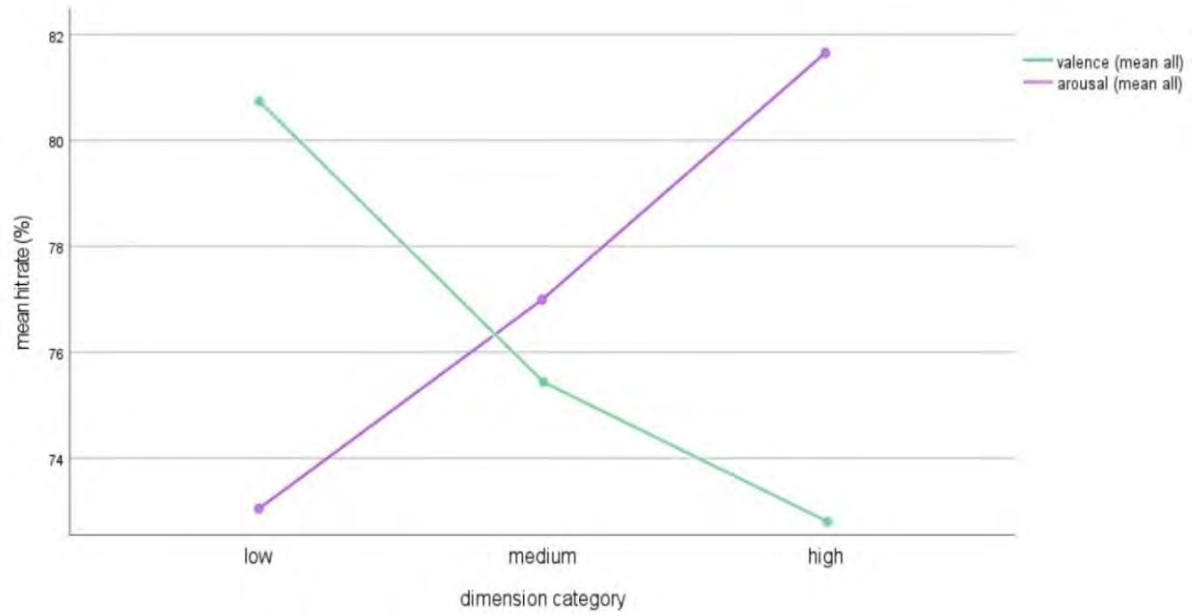


Figure 15. Mean hit rates per dimension category separated by valence and arousal. Note that the valence scale (*strongly negative – strongly positive*) is a bidirectional scale, while arousal (*not at all – very arousing*) is a one-directional scale.

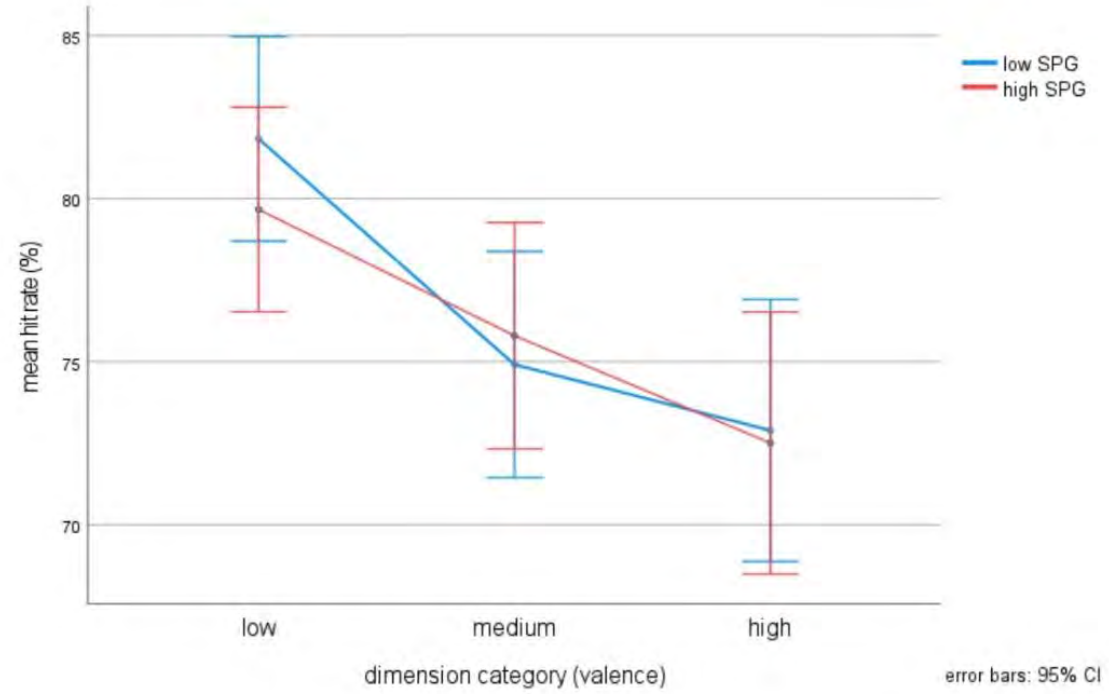


Figure 16. Mean hit rates per valence dimension category (*low*/*medium*/*high*) for *low* and *high* sensitive person groups (SPG). CI = confidence interval.

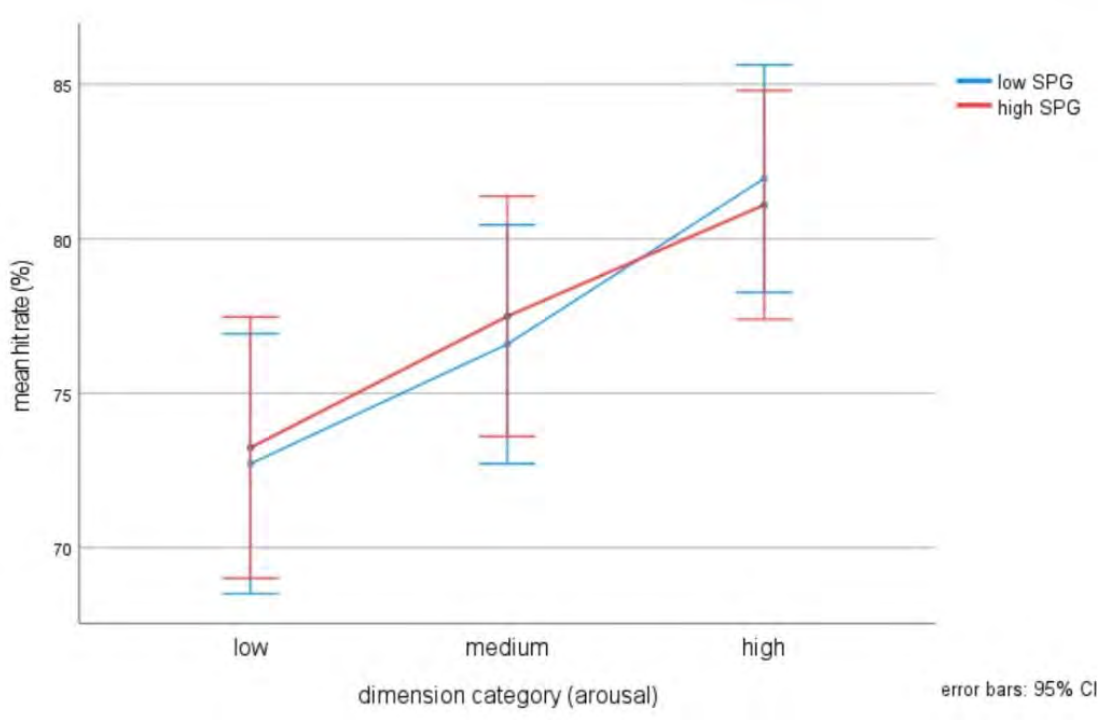


Figure 17. Mean hit rates per arousal dimension category (*low*/*medium*/*high*) for *low* and *high* sensitive person groups (SPG). CI = confidence interval.

Despite no significant effect of interaction between the factors gender and rating dimension, an interesting finding became visible when comparing *low* and *high* SPG separately for female and male participants hit rates: that is, in respect to valence, *low* SPG female participants display higher hit rates for more negative and more positive stimuli compared to neutral stimuli (U-shape tendency), while *high* SPG female participants display a decreasing hit rate with increasing stimulus valence. These tendencies in hit rates for *high* and *low* SPGs are reversed for male participants (U-shape tendency for *high* SPG; decreasing hit rate with increasing stimulus valence for *low* SPG), see Figure 18, below. Similarly, with regards to arousal ratings, *low* SPG females achieved higher hit rates compared to *high* SPG females, however, among male participants, individuals from the *high* SPG achieved higher hit rates compared to individuals from the *low* SPG.

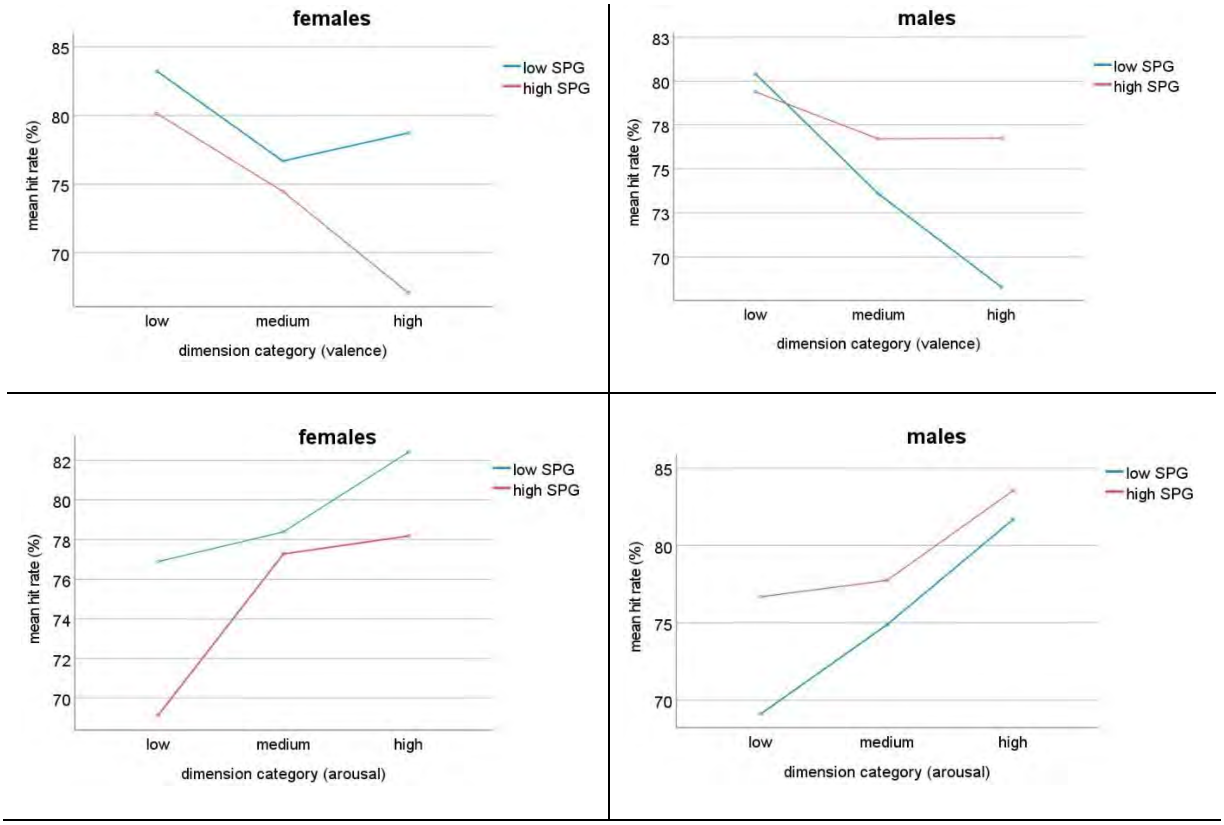


Figure 18. Mean hit rates per rating category for *low* and *high* sensitive person groups (SPG) in relation to valence rating (top) and arousal rating (bottom), (left: females; right: males).

Discussion

Results and Hypotheses

Stimulus Perception

The results of the current study indicate that perception significantly differs between all dimension categories for both valence and arousal (note that based on the idiographic rating data, there were no stimuli of *high*-arousal), and therefore justify the separation into these three dimension categories. Analyses revealed an interaction between the factors gender and dimension category with respect to the valence rating of stimuli, with female participants giving stimuli of *high* valence significantly higher ratings than male participants. In other words, female participants perceived positive stimuli more positive compared to male participants.

There was a main effect of SPS level regarding valence rating, with individuals of the *high* SPG providing lower stimuli ratings. This suggests that this group perceives negative stimuli as more negative, and positive stimuli as less positive, compared to individuals from the *low* SPG. These results, partly support the hypothesis *(H1a): "Image valence is perceived as more extreme, that is more negative for negative stimuli, and more positive for positive stimuli by participants of higher, compared to lower sensitivity."* In other words, the present findings suggest that *low*-valenced stimuli are perceived as more negative by individuals of higher

sensitivity (compared to lower sensitivity), however, stimuli of *high* valence are not perceived as more extreme (more positive) by participants of higher compared to lower sensitivity.

There was no significant effect of SPS level and arousal rating, therefore, the hypothesis *(H1b): "Image stimuli are be perceived as more arousing by participants of higher, compared to lower sensitivity."* Was not supported. Nevertheless, it is important to note that there were no stimuli falling into the *high* arousal category, which is restricting conclusions that may be drawn from the present results.

Furthermore, results did not show and significant effect between SPG and dimension category, nor SPG, dimension category and gender. Therefore, it remains unclear whether male participants from the *high* SPG may indeed not perceive stimulus valence and/or arousal differently compared to males from the *low* SPG, or whether *high* SPG males may be more susceptible to being influenced by social desirability, as suggestible by findings from Kring and Gordon (1998). In their work, authors report males to be less willing than females to report their internal emotional state, either verbally or nonverbally. With regards to the present study, this may be especially true as a large proportion of the presented stimuli included military-related content that may have triggered perception of masculinity (see e.g., Hinojosa, 2010). Nevertheless, present results suggest that in order to rely on homogenous effects of stimuli regarding valence in future studies, particular attention should be granted to participants' sensitivity.

Emotional Stimuli and Recognition Memory

There was a significant main effect of dimension category on the hit rate with significant differences between all dimensions (except *medium* and *high* valence). Again, these results support the advantage of separating stimuli into dimension categories. The recognition tendency was reflected by a hit rate that decreased with increasing stimulus valence and increased with increasing stimulus arousal. Therefore, the hypothesis *(H2a): "Recognition will significantly differ between stimuli that were perceived as positive and negative compared to neutral stimuli."* is only partly supported. Regarding valence, these findings support the assumption that an enhanced recognition for negative stimuli may facilitate avoidance of such stimuli in the future and hence promote survival. Moreover, these results support previous findings reporting attention-grabbing properties of negative information in young adults (e.g., Bebbington et al., 2017) as well as a general negativity bias in humans (Baumeister et al., 2001).

Additionally, results displayed a significant difference between *low* compared to *high* arousal stimuli. Therefore, the hypothesis regarding stimulus recognition in dependence of perceived arousal *(H2b): "Recognition will significantly differ between stimuli that were*

perceived as extremely arousing compared to stimuli perceived as low arousing." can be accepted. These results support previous findings suggesting the positive association between increasing arousal and memory (e.g., Anderson et al., 2006; Bradley et al., 1992; Hamann et al., 1997; Laney et al., 2004; Marx et al., 2008; Mather & Sutherland, 2009). Sharot and Yonelinas (2008; pp. 539) moreover state that "Because consolidation of memory occurs over a period of time, the beneficial effects of arousal on memory should be most apparent following a delay". In fact, their study as well as earlier research (e.g., Kleinsmith & Kaplan, 1963; Kleinsmith et al., 1963) found arousal to improve memory to a greater extent after a delay. A possible explanation for enhanced memory after delay that is specific to arousing material could be the mediating effect of cortisol. In that regard, a study conducted by Kuhlmann and Wolf (2006), displayed a beneficial effect of cortisol for delayed recall (vs. immediate) recall of emotionally arousing stimuli. Moreover, Nielson and Lorber (2009) found that differences between individuals that influence arousal responding (e.g., emotional reappraisal) may interfere with memory modulation. In relation to the present study, additional research is needed investigating the ideal delay to maximise the effect of arousal on recognition memory, as well as investigating whether the ideal delay for memory consolidation differs between individuals of varying trait sensitivity.

Taking into consideration that memory was slightly impaired with increasing stimulus valence, and improved with increasing stimulus arousal, it is surprising that despite significant valence rating differences between *high* and *low* SPG, no significant main effect of SPS level was found in relation to recognition rates. In other words, based on the finding that recognition decreased along with increased stimulus valence, and that individuals from the *high* SPG (vs. *low* SPG) perceived stimuli as significantly less positive, these groups would consequentially be expected to also display significant differences in recognition rates. A possible explanation for the absence of a difference in recognition memory could be that an increased emotional perception of *high* SPG individuals (compared to *low* SPG individuals) during the viewing of the stimuli as well as possibly through stimuli encountered between completion of Part 1 and Part 2 of this study, triggered different brain paths which in turn impeded memory. Further research is therefore necessary to investigate neuropsychological differences between individuals with *high* compared to *low* processing sensitivity.

Finally, the fact that gender had an effect onto dimension categories for rating but not for recognition, raises questions regarding the interaction between emotion perception and gender in relation to recognition memory. In that regard for example Wolf et al., (2001), found

that an increase in stress-induced cortisol negatively affected memory performance among male, but not female participants.

Limitations

Although providing valuable insight into differences in perception of emotional image stimuli between *low* and *high* SPG individuals, as well as the relation between recognition memory and stimulus valence and arousal rating category, the current study is not without limitations: First, participants were recruited online through a crowdsourcing platform (Prolific). As participants are financially remunerated after study completion and platform guidelines are comparably strict (e.g., completion of a study that is too fast may indicate the participant's inattention; participant had to pass at least 50% of the attention check questions), it is fair to assume that high quality data were collected for both parts of the current study. Nevertheless, as data were collected online, there was no option to verify participant's full attention to the questionnaire without distraction throughout the experimental procedure. Inattention during the rating procedure (survey *Part 1*), as well as distraction during the test phase (survey *Part 2*) may have impaired recognition of target stimuli. Additionally, as mentioned for the previous study (*Chapter Three*) data was collected during the time of the covid-19 pandemic which may have increasingly driven participants towards this platform. With studies being displayed to members of the platform in accordance to their profile, participants may be tempted to change their profile information allowing them to receive additional study options. This may in consequence lead to the inclusion of participants that do not fit the required inclusion criteria.

Second, a median split technique was implemented based on the HSPS participants completed in *Part 1* of the survey to create two comparable groups. Due to loss of information about individual variability (Farewell et al., 2004; Humphreys, 1978; MacCallum et al., 2002; Neelamegham, 2001), as participants slightly above or below the median were aggregated to the same group as participants who were farthest above or below the median, the likelihood for a Type II error may have been increased. Results of the current study are therefore to be viewed as preliminary findings regarding this research subject and need validation through experiments comparing participants based on HSPS scores. In a similar vein, and as mentioned in the previous study (*Chapter Three*) the minimum sample size of N = 80 was necessary as calculated through the power analysis. This was not achieved (there were N = 79 participants) as availability of participants was restricted by financial resource constrains as well as the accessibility of participants during the covid-19 pandemic. Therefore, the results are statistically underpowered. Separating the collected data points for comparison across genders,

as well as SPGs further reduced the number of data points for individual analyses and hence additionally affected statistical power of these study results. This highlights the limitation of generalizability of the present research findings, as well as the need for a study replication with a larger sample size.

Third, as mentioned in *Chapter Three*, previous research has shown that the duration of stimulus presentation influences liking (Marin & Leder, 2016; Reber et al., 1998). In the conducted study, viewing of stimuli was self-paced and hence varied between participants. Therefore, the collected rating data was susceptible to variation caused by differences in duration of stimulus presentation. Additionally, the variation in display duration may have affected the recognition of target stimuli. Additional research is needed to investigate whether stimulus assessment data may differ in dependence of display duration as well as the relation between presentation duration and recognition memory. This may for example be achieved by displaying a few stimuli (e.g., images) of a set for a short duration (e.g., 1 second) and a few stimuli for a longer duration (e.g., 5 seconds) and comparing recognition rates across variations of display duration.

Fourth, because of the construction of the study, only target stimuli were assessed prior to the memory test, and no analysis of displayed ethnicities was conducted. As aspects such as own-race bias (for review see Meissner & Brigham, 2001), or stimulus color (Kuhbandner & Pekrun, 2013), have shown to influence memory, and paired distractor stimuli were matched on content (e.g., color or greyscale, landscape, included individuals), no conclusion can be made with regards to the effect of similarity of *distractor* stimuli on recognition of *target* stimuli.

Finally, in the present study only hit rates were analyzed. It cannot be excluded that recognition was influenced by participants' recognition sensitivity or response bias. In future research, consideration of false alarms (e.g., recognition of distractor stimuli) may help gain additional insight.

Final Discussion

The results of the present study found significant main effects of the dimension categories regarding stimulus rating as well as stimulus recognition in relation to both, valence and arousal (except hit rate between *medium* and *high* valence). The tendency of decreasing recognition with increasing stimulus valence and increasing recognition with increasing stimulus arousal, provide further support to previous findings regarding arousal (e.g., Anderson et al., 2006; Bradley et al., 1992; Hamann et al., 1997; Laney et al., 2004; Marx et al., 2008; Mather & Sutherland, 2009) and contradict evidence reporting memory impairment in relation

to arousing stimuli (e.g., Dolcos & McCarthy, 2006). Additionally, they also run contrary to findings regarding stimulus valence (e.g., Kensinger & Corkin, 2003) reporting enhanced as well as more detailed memory for negative compared to neutral word stimuli. In the present study, however, no analyses were made regarding a valence-arousal interaction (e.g., comparisons of low-valence/high-arousal vs. high-valence/high-arousal stimuli).

Nevertheless, the present study sought to answer the question *Q3: "What is the relationship between trait sensitivity, emotion perception of stimuli, and recognition memory?"*. In that regard the results showed, that individuals of *high* sensitivity perceive stimuli as less positive compared to individuals of *low* sensitivity. Nevertheless, no statistically significant difference was found concerning recognition memory (e.g., hit rate) between *low* and *high* SPG. This in turn may suggest different interactions between emotion and memory processes in individuals with *low* compared to *high* SPS in relation to visual emotional content (e.g., images). An additional interesting finding was the reversed effect between female and male participants of *low* and *high* sensitivity regarding recognition rates with respect to valence and arousal. That is, for valence, *high* SPG males showing higher hit rates compared to *low* SPG males, however, *low* SPG females displaying higher hit rates compared to *high* SPG females. This same effect was also visible regarding stimulus arousal, with *high* SPG males showing greater hit rates compared to *low* SPG males, and *low* SPG females displaying higher hit rates compared to *high* SPG females. Although this effect became visible in graphical representations, it remained statistically non-significant. Nevertheless, it may suggest a pattern of responding in relation to trait sensitivity and gender and needs further investigation.

Regarding the investigated factors (e.g., gender, SPS level, dimension category), an effect of interaction was visible only for the factors gender and dimension category in relation to valence ratings of stimuli. In that regard, the analyses showed that female participants gave significantly higher ratings to *high*-valence stimuli compared to male participants. Moreover, there was a significant 3-way interaction reflecting the influence of SPS level onto stimulus valence rating for *low*-valence stimuli for both females and males, as well as for stimuli of *medium* valence among female participants.

Due to the limited number of participants, the present findings may solely be regarded as preliminary findings and are in need of replication with a larger participant sample for each of the groups (e.g., *high*/*low* SPG females; *high*/*low* SPG males). Finally, it would be necessary to expand the present research by investigating the role of stimulus type (e.g., words, video/ audio clips) to investigate if the effects found in the present study are replicable across other types of stimuli.

Chapter Five – General Discussion

With emotions playing in central role in everyday life, and ES being one of the main tool used in emotion-related research, the central aim of the present work was to investigate factors that may affect validity and hence utility of ES within emotion research. In a first step, a comprehensive review of all available stimuli sets and their key characteristics was conducted to gain an overview of existing ES sets. The result of this review is a searchable online database of ES sets, the KAPODI database (*Chapter Two*). In a second step, two studies were conducted to investigate the reliability of ES with regards to different factors that possibly affect stimulus validity (*Chapter Three*), as well as to investigate the relation between trait sensitivity and emotion perception of ES as well as the relation between emotion perception and memory (*Chapter Four*).

With the three conducted studies the present research program sought to answer the following research questions:

Q1: How many emotional stimuli sets are available to the research community? and *Which are the key set characteristics?*
Q2: Do ratings of emotional image and word stimuli remain generally reliable numerous years post publication?
Q3: What is the relationship between trait sensitivity, emotion perception of stimuli, and recognition memory?

Answering the Principal Research Questions

Q1

Numerous sets of emotional stimuli have been published throughout the past few decades, continuously increasing the need for an overview of all existing sets. Although ES sets have been reviewed by researchers previously (e.g., Grühn & Sharifian, 2016; Krumhuber et al., 2017), these attempts focus only on specific sorts of stimuli or did not systematically review existing literature and are hence not comprehensive.

The systematic review conducted within the frame of the present research project is the first comprehensive collection of stimuli sets that are freely accessible or available upon request. The KAPODI database created based on the systematic review contains 364 different publications of ES sets at the time of writing; most of which are presenting new stimuli (compared to new assessment data of an already existing set). To facilitate comparison when

using the database, six subfolders were created based on stimulus type, that is audio, faces, images, video, words, and mixed stimuli.

Despite a careful selection of terms used for the stimuli set search procedure (see *Chapter 2*), an extended-table search revealed another n = 74 publications during the first search, as well as n = 21 publications during the second (updated) search (Figure 1). This is a highly concerning finding, as it displays the difficulty researchers may have to find emotional stimuli sets when searching for suitable stimuli for their own study. Moreover, it highlights the 'iceberg-dilemma' discussed earlier, reflecting an emphasized use of well-known ES sets such as the *ANEW* (Bradley & Lang, 1999a) or the *IAPS* (Lang et al., 1997).

Central set characteristics were extracted and listed for each set. In total, information regarding more than 45 aspects including over 25 key set characteristics is available for each set, based on the information provided in the original publication. Among others, the key set characteristics are type and number of stimuli; number, gender, age and ethnicity of included models; language; rating scale; or number, gender, age and ethnicity of assessors. Coded aspects include information regarding title of publication, authors, assessment approach (e.g., dimensional, categorical), rating scale length, or number of included basic emotions; the entire list can be found in *Chapter Two*, Table 1.

As a result, the KAPODI database listing the numerous existing stimuli sets along with their information regarding stimulus access, helps to significantly shorten the time needed to search existing sets. The extracted key set characteristics hereby provide a clear overview of set content (e.g., number of included stimuli, type of stimuli) as well as set validation procedures (e.g., information regarding rating population) and coded information (e.g., dimensions on which stimuli were assessed). This information bundled in one place, thus saves the researchers' additional time, as they can easily scan the set characteristics without having to read the original publication. Finally, by providing set information separated into over 45 aspects, the database also facilitates comparison across sets and hence selection of suitable sitmuli.

Next to an Excel sheet containing all extracted information (available as *Supplementary Material*), an online version of the database additionally allows for an automated search based on search criteria that can be manually modified. Moreover, if researchers wish, they can add their newly created set to the online database. Not only will this allow for the set being more easily found among all other existing sets, but also contribute to the aim of keeping the online KAPODI database continuously updated. This in turn will allow fellow scientists to conduct

their researchers as by state-of-the-art through a consideration of a wide choice of existing stimuli.

As well as presenting a useful tool in itself, the analyses of the existing data within the KAPODI database provide useful information regarding the state, as well as the growth of emotion research as a field. That is, analyses of stimuli sets included in the database revealed a steep increase in set publication throughout the past three decades especially with regards to face stimuli. This development against the backdrop of improvement of technology surely reflects a growing interest in artificial intelligence and human-machine interaction, of which research seems to be predominantly conducted in Asia, followed by Europe. Moreover, the assessment approach used in individual publications (e.g., categorical vs. dimensional; applied rating scale) displayed a favoured approach in relation to type of stimulus. That is, while face stimuli are mostly assessed using a categorical approach as well as using force-choice, word stimuli for instance are mostly assessed through the dimensional approach and by using a Likert-type scale. Although the stimulus quality (e.g., available size for images; audio recording quality) as well as the extensiveness of reported set characteristics (e.g., demographic data of the rating population; demographic data of included models) in publications has increased, meaning that newer publications provide stimuli of higher quality as well as more detailed information, uniformity is still lacking. Therefore, to facilitate the selection process of stimuli for other researchers, scientists presenting new stimuli are advised to include as detailed information as possible. All extracted key set characteristics listed in the KAPODI database could herby serve as a guide.

Q2

With ES playing a central role within emotion research, the validity and thus reliability of ES is indispensable to assure intended effects of ES and hence correct interpretation of study results. Various factors such as gender (e.g., Garrido et al., 2017; Janschewitz, 2008; Warriner et al., 2013; Weierich et al., 2019), ethnicity (e.g., DeBusk & Austin, 2011; Meissner & Brigham, 2001), or age (e.g., Issacowitz et al., 2007) have been shown to influence the perception of emotional stimuli and these aspects of a rating population are hence usually reported along with the assessed data. Nevertheless, the reported information is usually restricted to *assessor*-related factors. However, existing evidence moreover suggests an influence of *stimulus*-related factors such as display size (Codispoti & De Cesarei, 2007) or factors related to study construction such as display duration (e.g., Marin & Leder, 2016), context (e.g., Wogalter, 1998), or assessment scale (e.g., Hasson & Arnetz, 2005) onto stimulus perception. Unfortunately, these aspects are not always reported in detail, and researchers often

use ES without verifying stimulus validity for their own participant sample. This common procedure poses a high risk of misinterpreting study results, if stimuli do not achieve the believed effect in participants.

With stimulus validity being a central necessity when conducting a study using ES, the effect of individual factors on stimulus validity is at great need of investigation. Therefore, the second study within the research project aimed to verify the reliability of stimuli by investigating factors that may be prone to changes in stimulus validity [*dimension* (e.g., valence, arousal, dominance), *dimension category* and *SD category* (e.g., high, medium, low), *stimulus type* (e.g., images and words), and *gender* (e.g., female, male)]. Indeed, results displayed that regarding dimension, solely valence ratings (vs. arousal and dominance) remain relatively stable. In other words, with regards to the other two dimensions namely, arousal and dominance, researchers are highly advised to reassess stimuli prior to study conduction rather than to select ES based on the normative data. Moreover, the present results reinforce the questioning regarding reliability of findings resulting from previously conducted studies in which stimuli had for instance been selected based on or matched for normative data, without reassessment of stimuli. Examples are studies using stimuli matched on arousal (e.g., Gil & Droit-Volet, 2012), valence (e.g., Bonin et al., 2014), or both (e.g., Noulhiane et al., 2007), as well as distinct emotion intensity (e.g., Droit-Volet et al., 2011). The assumption that inconsistent findings within a research area (e.g., memory research in relation to emotion) may be due to lacking validity of stimuli (e.g., through simple use without reassessment for the study in question) is therefore maintained and requires more specific investigation, for instance in the form of meta-analyses. More importantly, however, it highlights the need of careful selection of stimuli through consideration of the validation procedure in the original study.

In-depth analyses investigating the interplay between dimension and stimulus type, as well as dimension and dimension/*SD* category had been conducted to shed light onto the role of these individual factors. The results of the comparison between stimulus type (images vs. words) display a high validity of valence for both stimulus types, however, moreover suggest an overall higher reliability of arousal for word stimuli (vs. images), and a higher reliability of dominance for images (vs. words). Simultaneously, irrespective of stimulus type, the reliability of ratings decreasing from valence to arousal to dominance (in that order) may be an indicator for familiarity with that dimension. That is, as mentioned earlier, participants may be more familiar with the meaning of valence, than with the meaning of arousal; and finally least familiar with the significance of dominance. However, not due to a limited understanding, but rather a less-frequent application of these judgements in everyday life. That is, while individuals

constantly make valence judgements in everyday life situations (How good or bad /positive or negative is this?), conscious arousal and dominance evaluations are made far less frequently and thus be more abstract. In fact, the suggestion made by Stevenson et al., (2011) to further differentiate sexual arousal from general arousal assessments may reflect that participants are not always able to grasp the meaning of arousal without confusing it with sexual arousal, if they are not clearly informed about the differences prior to stimulus assessment. The reliability of stimulus ratings remained low to medium for both stimulus types when controlling for dimension category. In other words, stimuli assessed today are unlikely to fall into the same dimension category as suggested by the normative rating data. Therefore, researchers are advised to refrain from selecting stimuli based on these dimension categories while relying on the normative data without reassessing stimuli for their own study. Although reliability was slightly better in relation to *SD* category, and more acceptable regarding the dimension valence compared to arousal or dominance, researchers should also refrain from relying normative data based on these categories. In conclusion, besides valence ratings, the investigated factors all seem prone to changes in stimulus validity throughout time, and researchers are hence advised to reassess stimuli for their participant sample if they are planning to use them in relation to other dimensions than valence (e.g., arousal).

Finally, a striking finding was revealed when comparing female and male perception of stimuli: Overall perception of both genders highly correlated. Nevertheless, when investigating the individual stimuli up to 25 % of the stimuli were perceived significantly differently between female and male participants. This finding supports previous suggestions to report data separately for both genders, and to moreover select stimuli separately for female and male participants when planning a study. However, more importantly, this direct comparison highlights the misleading character of the assumption that high correlations between rating data of both genders reflect their strong similarity and therefore justifies the use of the same stimuli for both genders in one participant sample without differentiation. That is, especially regarding stimuli sets that contain a large number of stimuli, the correlation between female and male ratings across all stimuli may be high, however, selecting only a few stimuli from this set increases the risk reducing this correlation.

Q3

A large field of research in relations to emotions regards the relation between emotion and memory. Among existing literature concerning this relation, research findings are contradictory with some study results displaying a beneficial effect of certain emotions onto memory (e.g., Brown & Kulik, 1979; Cahill et al., 2003; Chepenik et al., 2007; Goldberg & Todman, 2018;

Harris, 1999; Harris & Cumming, 2003; Kensinger & Corkin, 2003; Storbeck & Clore, 2005; Wolf, 2008), while other studies reveal an impairing effect of emotions onto memory (e.g., Kirschbaum et al., 1996; Schwabe & Wolf, 2010; Strange et al., 2010) as well as a channelling of attentional focus (e.g., Christianson & Loftus, 1991; Easterbrook, 1959; Gray, 2001). Additionally, Aron and colleagues (2012) suggest a deeper processing of emotional information by HSPs compared to non-HSPs. These two broad findings together (contradictory findings on the effect of emotions onto memory, as well as personality influencing the perception of emotion) have opened the question concerning the relation between sensory processing sensitivity, perception of emotional stimuli, and memory, which was finally addressed in the third study of the present research project.

Results displayed an interaction effect of gender and dimension category, showing that female participants perceive stimuli of *high* valence as significantly more positive compared to male participants. Nevertheless, there was no significant difference in recognition rate between both genders concerning stimuli of this dimension category. It is possible, that the covid-19 pandemic with all related restrictions (e.g., social distancing, covering of the face with a mask and hence shielding emotion expression and/or perception) has affected the emotion perception of emotional stimuli. That is, participants may have been inclined to compare the displayed content (e.g., a group of friends laughing together) to the overall global as well as their personal situation and provided stimulus ratings against this background. Especially as many positive image stimuli display social situations (e.g., friends gathering), landscapes, as well as (baby) animals and children, the participants in the present study may have been confronted to the importance of social interaction simultaneously as the fact of not being able to leave the house. This effect may have been more pronounced in women compared to men. Nevertheless, as this remains speculation, a replication study will be needed investigating aspects such as wealth of social contacts or freedom of movement onto stimulus perception. In a similar vein, results showed that the processing sensitivity seems to have an effect onto the valence perception of stimuli that were as a result perceived more negative by individuals with high SPS (vs. low SPS). However, despite differences in perception, the recognition memory did not significantly differ between individuals of differing SPS. This finding could be in line with previous research that has shown that verbalized perception (rating assessment) and physiological responses to stimuli are not always consistent. That is, individuals may perceive a stimulus in a way that is not reflected by their physiological changes in response to the stimulus (Gross & Levenson, 1993). Moreover, voluntary emotion suppression has been shown to influence physiology such as skin conductance and heart rate (e.g., Reynaud et al., 2012). Concerning the present study,

this could indicate that individuals of *high* SPS are more likely to indicate a more extreme perception of negative stimuli (conscious perception), however, these individuals then implement emotion regulation strategies that dampen the effect strong emotions would otherwise have. Due to lacking scientific evidence within this field of research to this day, the above-mentioned effect, as yet, is speculation and requires further investigation.

Nevertheless, analyses of stimulus recognition were also conducted separately for female and male participants. Displayed graphically, this separation revealed an interesting pattern, suggesting – although without statistically significant differences – *inversed* recognition tendencies for female and male participants regarding *low* and *high* SPGs. That is, while among female participants individuals of *low* SPS achieved overall higher hit rates (vs. *high* SPG), *high* SPG (vs. *low* SPG) achieved higher hit rates among male participants. This pattern was visible in relation to valence as well as arousal rating of stimuli. Nevertheless, these findings are based on a relatively small participant sample and therefore in need of replication with a larger sample group. Further investigations could elucidate emotion processing between genders, and – assuming the above-mentioned speculation – help to shed light onto the effect of emotion regulation strategies implemented by females compared to males of *high* SPS.

More broadly, that is independent of SPS level, results showed that recognition (hit rate) increased with increasing arousal and decreased with increasing valence. These results are in line with previous evidence regarding arousal ratings (e.g., Anderson et al., 2006; Bradley et al., 1992; Hamann et al., 1997; Laney et al., 2004; Marx et al., 2008; Mather & Sutherland, 2009), and contradict findings reporting enhanced memory for negative compared to neutrally valenced stimuli (e.g., Kensinger and Corkin, 2003). Additionally, Libkuman and colleagues (2004) found that arousal increased central detail memory for positive and negative stimuli, as well as background detail memory for positive stimuli. Nevertheless, in the present work no analyses were conducted on the stimulus content to investigate central/background information.

Future Research

Based on the findings and limitations regarding the individual studies of the present research, suggestions for future research can be given. These will be expanded upon in the section below.

The KAPODI Database

The creation of a central database of ES certainly facilitates access, comparison and thus selection of suitable stimuli. Nevertheless, sometimes terms or definitions are non-uniformly used across research disciplines which means that existing norms within individual research disciplines can impede collaboration. For example, physiological data (e.g., skin conductance

response; skin temperature) used to train machines for automatic emotion recognition are called *emotional stimuli* within psychology research. Nevertheless, this type of data is more broadly known as *data points* within computer engineering. As ES are used across a wide array of (research) fields (e.g., sociology, anthropology, psychotherapy, computer engineering, medicine), future research should focus on elaborating a way to equate these terms across research disciplines. This could for example be done through an extension of the database in which the certain terms are double-coded and hence displayed in response to more than just one search term.

Moreover, an addition to the database (external or within the database) in which users can give feedback on the use of sets (e.g., applicability of specific stimuli for specific research questions/study constructions), may help fellow researchers ~~to~~ select suitable stimuli even more easily. Additionally, this "feedback-section" could be used as a direct reference to improve or refine the creation of new stimuli according to the needs within research.

Participant Sample

Part of the challenge when conducting a study is the recruitment of a large participant sample allowing generalizations of the findings. The present studies (*Chapter Three*; *Chapter Four*) were conducted with a limited number of participants (n = 125 and n = 78, respectively). A replication of the present studies with a larger participant sample is therefore necessary. This would allow to verify the present results, as well as to improve their representativeness. Moreover, the present studies were conducted within the context of the Covid-19 pandemic. Collecting the data online and through the Prolific platform may have influenced the participant sample. That is, this may on the one hand have helped collecting data more easily due to a great number of available participants, on the other hand however, this may have automatically provided a certain selection of participants (e.g., individuals who due to the economic development throughout the pandemic joined such platforms due to personal financial needs). The pandemic with its restrictions as implemented in certain countries (e.g., social interactions, physical activities) may moreover have influenced the participants' emotional well-being and hence perception of emotional content of stimuli presented to them.

Additionally, the present studies were conducted using image and word stimuli only. The analyses of the KAPODI database content, however, have also shown that video and audio stimuli sets represent 21,48 % of freely existing sets (see *Chapter Two*, Table 2). Therefore, a replication of the present studies including for example these stimuli would allow for comparison of the investigated factors across additional stimuli types and moreover for more specific assertions depending on the stimulus type.

Moreover, the present studies were conducted on an English-speaking adult participant sample. An expansion of research investigating individual factors affecting stimulus reliability in a different culture could help to shed light onto aspects that may influence changes in perception. That is, cultures affected by a rapid development (e.g., technology, education, economy) may be more prone to changes in emotion perception compared to cultures without or with a slow development. In that same vein, in future research it would be interesting to investigate whether the strength of the factors influencing stimulus reliability varies in relation to different age groups (e.g., children, teenagers, young adults, older adults).

Finally, in the second study (see *Chapter Three*), included a participant of non-binary gender, who however, was excluded for analyses comparing females to males. Despite the differences in perception of emotional stimuli based on gender (e.g., Lithari et al., 2010; Nater et al., 2006), the rating of this subject correlated with similar strength to both, female as well as male ratings. In fact, among all sets included in the KAPODI database, not a single publication provides data validated by non-binary participants. Against the background of an increasing visibility, awareness, and acceptance (e.g., non-binary language) of non-binary genders, as well as an estimated prevalence of up to 4.6 % (e.g., Åhs et al., 2018; Kuyper & Wijsen, 2014; Van Caenegem et al., 2015;) with an increasing trend (Twist & de Graaf, 2019), it may become necessary to extend research analyses by this third group in the future when investigating gender differences, as well as when for instance creating and validating new ES sets.

Normative Data

As discussed extensively, various assessor- and study construction -related aspects influence the perception of emotional stimuli and hence the reliability of the normative rating. Nonetheless, even when controlling for these aspects (e.g., gender, age) the normative data is usually reported as a calculated mean across participants (mostly accompanied by data regarding *SD*). Herein lies an issue that is worth raising awareness for and that has also been reported by Montefinese et al., (2014), as well as Schneider et al., (2016): While for instance a mean valence rating of 5 on a 9-pt scale may be considered “neutral” (and hence neither bad/negative nor good/positive) this stimulus treated as a neutral stimulus has not necessarily been perceived as neutral by all participants. This means, while a low *SD* would indicate consensus across raters for one stimulus, another stimulus can receive a mean rating of 5 (hence being “neutral”) however, with high and low rating values from individual participants, which would be reflected in a higher *SD* value. In other words, the *SD* can be seen as an indicator for ambiguity or ambivalence for each stimulus. This is particularly critical, as the experienced level of ambivalence is for instance predictive of the arousal people report (Nordgren et al.,

2006; van Harreveld et al., 2009). Therefore, stimuli with a high standard deviation should be avoided in studies aiming to achieve a standardised effect (e.g., emotion induction) across all participants (note that this does not refer to the fact that stimuli with high *SD* in the normative assessment, may also display high *SD* in idiographic assessment). It is thus indispensable that researchers consider the *SD* beyond a simple source of systematic error when selecting stimuli based on the normative mean rating data in future studies. Moreover, the *SD* should always be reported along with the normative rating data.

Additionally, when normative rating data is assessed, participants are usually asked to indicate their perception of a specific stimulus on a dimensional scale (valence/arousal/dominance) or via indicating a distinct emotion (rating on a scale, or indication of the emotion through forced choice). This procedure assumes that participants can accurately describe their emotions. Nevertheless, words labelling emotions are often language-specific and it is often difficult to translate into one word in another language (Altarriba, 2003). Moreover, the access to one's own emotions that is the ease to control expression and perception of emotions, differs across cultures (Matsumoto, 1989). Due to this influence of culture and/or language, the question whether individuals can indeed accurately describe their emotions remains open for discussion. Nevertheless, the normative data is provided along with the stimuli to give the researcher an idea of the emotional value of each stimulus. To date this is the best available mean indicator and therefore valuable information – at least by the means of providing orientation when selecting emotional stimuli.

Finally, an aspect that is worth mentioning is that some stimuli (e.g., an image of a bungee-jumper) may trigger physiological arousal, while other stimuli may trigger cognitive arousal (e.g., an image of a decomposed animal). Participants may possibly be confused when having to compare their cognitive vs. physiological while being asked to provide arousal ratings. In the present two studies participants were informed about the meaning of the dimension *arousal* and it was highlighted that the referred-to arousal was not sexual arousal. In previous studies instructions may not always have been sufficiently clear to the participants which may be reflected in the normative data. Therefore, a central advice for future research is to provide a clear explanation of the assessed dimension and so ensure that participants are familiar with them.

Conclusion

In conclusion, although ample research has investigated the effect of emotions, and ES have been used within research for more than 60 years, research investigating the use of ES is scarce leaving a lot of unanswered questions. For instance, only little is known about the reasons

for a differing perception of ES across participants, and possibilities for future research seem sheer endless. One main result of the present research project was the creation of the searchable KAPODI database resulting from a systematic review. As the database comprises the largest number of all freely available ES sets to this day, it facilitates access to as well as comparison between individual stimuli sets and may hence serve as a tool for researchers who are searching for suitable ES for their study. In that regard, it clearly contributes to solving the 'iceberg-dilemma' (emphasis of well-known ES sets through amplified use within research causing an overshadowing effect onto smaller, sell-known stimuli sets) and hence allows state-of-the-art research, by giving researchers the opportunity to comprehensively consider the entire array of available stimuli sets. The number of views and downloads of the resulting publication (over 1,500 in the 6 months post-publication, [Diconne et al., 2022]) reflect the interest for such a resource. Moreover, the two conducted studies aimed to identify factors that may affect stimulus reliability – a possible cause for previous inconclusive study results within emotion research. Although the generalizability of the findings from the present research is limited due to a small participant sample, the investigation may make an important contribution to the understanding of emotional stimulus validity as well as factors affecting validity. They can therefore certainly be regarded as a sensible start regarding investigations of ES reliability.

As emotional stimuli are used across a wide array of (research) fields (e.g., sociology, anthropology, psychotherapy, computer engineering, medicine), the strength of the present research project on factors affecting the utility of emotional stimuli in research lays in its implications for science beyond the field of psychology, hence promoting interdisciplinarity.

The hope is that together, the three studies helped to shed light onto open questions of emotion research, and that with previous as well as future research scientific knowledge will grow and serve to the benefit of humanity.

Reference List

Acevedo, B. P., Aron, E. N., Aron, A., Sangster, M. D., Collins, N., & Brown, L. L. (2014). The highly sensitive brain: an fMRI study of sensory processing sensitivity and response to others' emotions. *Brain and Behavior*, *4*(4), 580–594. https://doi.org/10.1002/brb3.242

Adolphs, R., Denburg, N. L., & Tranel, D. (2001). The amygdala's role in long-term declarative memory for gist and detail. *Behavioral Neuroscience*, *115*(5), 983. https://doi.org/10.1037//0735-7044.115.5.983

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature*, *393*(6684), 470–474. https://doi.org/10.1038/30982

Åhs, J. W., Dhejne, C., Magnusson, C., Dal, H., Lundin, A., Arver, S., Dalman, C., & Kosidou, K. (2018). Proportion of adults in the general population of Stockholm County who want gender-affirming medical treatment. *PLoS One*, *13*(10), e0204606. https://doi.org/10.1371/journal.pone.0204606

Altarriba, J. (2003). Does cariño equal 'liking"? A theoretical approach to conceptual nonequivalence between languages. International Journal of Bilingualism, 7, 305-322. https://doi.org/10.1177/13670069030070030501

Anderson, A. K., Wais, P. E., & Gabrieli, J. D. (2006). Emotion enhances remembrance of neutral events past. *Proceedings of the National Academy of Sciences*, *103*(5), 1599–1604. https://doi.org/10.1073/pnas.0506308103

Andric, S., Maric, N. P., Knezevic, G., Mihaljevic, M., Mirjanic, T., Velthorst, E., & van Os, J. (2016). Neuroticism and facial emotion recognition in healthy adults. *Early Intervention in Psychiatry*, *10*(2), 160–164. https://doi.org/10.1111/eip.12212

Anger, S., Camehl, G., & Peter, F. (2017). Involuntary job loss and changes in personality traits. *Journal of Economic Psychology*, *60*, 71–91. https://doi.org/10.1016/j.joep.2017.01.007

Aron, A., Ketay, S., Hedden, T., Aron, E. N., Rose Markus, H., & Gabrieli, J. D. (2010). Temperament trait of sensory processing sensitivity moderates cultural differences in neural response. *Social Cognitive and Affective Neuroscience*, *5*(2-3), 219–226. https://doi.org/10.1093/scan/nsq028

Aron, E. N. (1996a). *Hoog sensitieve personen: Hoe blijf je overeind als de wereld je overweldigt* [The highly sensitive person: How to thrive when the world overwhelms you?]. Utrecht: A. W. Bruna Uitgevers B.V..

Aron, E. N. (1996b). Counseling the highly sensitive person. *Counseling and Human Development*, *28*(9), 1–7.

Aron, E. N. (1996c). *The highly sensitive person*. New York, NY: Broadway.

Aron, E. N. (2001). *The highly sensitive person in love: Understanding and managing relationships when the world overwhelms you*. Harmony.

Aron, E. N., & Aron, A. (1997). Sensory-processing sensitivity and its relation to introversion and emotionality. *Journal of Personality and Social Psychology*, *73*(2), 345. https://doi.org/10.1037//0022-3514.73.2.345

Aron, E. N., Aron, A., & Jagiellowicz, J. (2012). Sensory processing sensitivity: A review in the light of the evolution of biological responsivity. *Personality and Social Psychology Review*, *16*(3), 262–282. https://doi.org/10.1177/1088868311434213

Aron, E. N., Aron, A., Nardone, N., & Zhou, S. (2019). Sensory processing sensitivity and the subjective experience of parenting: An exploratory study. *Family Relations*, *68*(4), 420–435. https://doi.org/10.1111/fare.12370

Balzus, L., Klawohn, J., & Kathmann, N. (2021). Feeling bad about being wrong: Affective evaluation of performed actions and its trial-by-trial relation to autonomic arousal. *Emotion*, *21*(7), 1402–1416. https://doi.org/10.1037/emo0000995

Barkus, C., McHugh, S. B., Sprengel, R., Seeburg, P. H., Rawlins, J. N. P., & Bannerman, D. M. (2010). Hippocampal NMDA receptors and anxiety: at the interface between cognition and emotion. *European Journal of Pharmacology*, *626*(1), 49–56. https://doi.org/10.1016/j.ejphar.2009.10.014

Barrington, B. L. (1963). A list of words descriptive of affective reactions. *Journal of Clinical Psychology*, *19*(2), 259–262. https://doi.org/10.1002/1097-4679(196304)19:2<259::AID-JCLP2270190238>3.0.CO;2-3

Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, *36*(2), 253–263. https://doi.org/10.1017/S0048577299971664

Battocchi, A., Pianesi, F., & Goren-Bar, D. (2005). The properties of DaFEx, a database of kinetic facial expressions. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3784 LNCS, 558–565. https://doi.org/10.1007/11573548_72

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323.

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Baveye, Y., Bettinelli, J. Dellandréa, E., Chen L., & Chamaret, C. (2013). A Large Video Database for Computational Models of Induced Emotion. Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, pp. 13–18, https://doi.org/10.1109/ACII.2013.9

Bebbington, K., MacLeod, C., Ellison, T. M., & Fay, N. (2017). The sky is falling: evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior*, *38*(1), 92–101. https://doi.org/10.1016/j.evolhumbehav.2016.07.004

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720. https://doi.org/10.1109/34.598228

Bellezza, F. S., Greenwald, A. G., & Banaji, M. R. (1986). Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*, *18*(3), 299–303. https://doi.org/10.3758/BF03204403

Benham, G. (2006). The highly sensitive person: Stress and physical symptom reports. *Personality and Individual Differences*, *40*(7), 1433–1440. https://doi.org/10.1016/j.paid.2005.11.021

Bertels, J., Deliens, G., Peigneux, P., & Destrebecqz, A. (2014). The Brussels Mood Inductive Audio Stories (MIAS) database. *Behavior Research Methods*, *46*(4), 1098–1107. https://doi.org/10.3758/s13428-014-0445-3

Bertels, J., Kolinsky, R., & Morais, J. (2009). Norms of emotional valence, arousal, threat value and shock value for 80 spoken French words: Comparison between neutral and emotional tones of voice. *Psychologica Belgica*, *49*(1), 19. https://doi.org/10.5334/pb-49-1-19

Bireta, T. J., Guitard, D., Neath, I., & Surprenant, A. M. (2021). Valence does not affect serial recall. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *75*(1), 35. https://doi.org/10.1037/cep0000239

Boiger, M., Ceulemans, E., De Leersnyder, J., Uchida, Y., Norasakkunkit, V., & Mesquita, B. (2018). Beyond essentialism: Cultural differences in emotions revisited. *Emotion*, *18*(8), 1142–1162. https://doi.org/10.1037/emo0000390

Bolton, J. E., & Wilkinson, R. C. (1998). Responsiveness of pain scales: a comparison of three pain intensity measures in chiropractic patients. *Journal of Manipulative and Physiological Therapeutics*, *21*(1), 1–7.

Bonin, P., Gelin, M., & Bugaiska, A. (2014). Animates are better remembered than inanimates: Further evidence from word and picture stimuli. *Memory & Cognition*, *42*(3), 370–382. https://doi.org/10.3758/s13421-013-0368-8

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Bradley, M. M., & Lang, P. J. (1999a). Affective norms for English words (ANEW): Instruction manual and affective ratings (Vol. 30, No. 1, pp. 25–36). Technical report C-1, the center for research in psychophysiology, University of Florida.

Bradley, M. M., & Lang, P. J. (1999b). International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2). Gainesville, FL: University of Florida, Center for Re-

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition*, *18*(2), 379. https://doi.org/10.1037//0278-7393.18.2.379

Bradley, M. M., Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1): 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition*, *5*(1), 73–99. https://doi.org/10.1016/0010-0277(77)90018-X

Brunier, G., & Graydon, J. (1996). A comparison of two methods of measuring fatigue in patients on chronic haemodialysis: visual analogue vs Likert scale. *International Journal of Nursing Studies*, *33*(3), 338–348. https://doi.org/10.1016/0020-7489(95)00065-8

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. Proc. Interspeech 2005, 1517–1520, https://doi.org/10.21437/Interspeech.2005-446

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neuroscience*, *21*(7), 294–299. https://doi.org/10.1016/s0166-2236(97)01214-9

Cahill, L., Gorski, L., & Le, K. (2003). Enhanced human memory consolidation with post-learning stress: Interaction with the degree of arousal at encoding. *Learning & Memory*, *10*(4), 270–274. https://doi.org/10.1101/lm.62403

Cahill, M. C. (1975). Interpretability of graphic symbols as a function of context and experience factors. *Journal of Applied Psychology*, *60*(3), 376–380. https://doi.org/10.1037/h0076624

Calvo, M. G., Fernández-Martín, A., Recio, G., & Lundqvist, D. (2018). Human observers and automated assessment of dynamic emotional facial expressions: KDEF-dyn database validation. *Frontiers in Psychology*, *9*(OCT), 1–12. https://doi.org/10.3389/fpsyg.2018.02052

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology, 39* (1/4), 106-124. https://doi.org/10.2307/1415404

Cardos, R., Predatu, R., & David, O. (2016). Development and validation of a cartoon based set of children's facial emotion stimuli for the REThink online therapeutic game. *Cognitive Behavioral Coaching* (Iccbc 2016), 39–43.

Carvalho, S., Leite, J., Galdo-Álvarez, S., & Gonçalves, Ó. F. (2012). The emotional movie database (EMDB): A self-report and psychophysiological study. *Applied Psychophysiology and Biofeedback*, *37*(4), 279–294. https://doi.org/10.1007/s10484-012-9201-6

Chen, S. H., Kennedy, M., & Zhou, Q. (2012). Parents' expression and discussion of emotion in the multilingual family does language matter? *Perspectives on Psychological Science*, *7*(4), 365–383. https://doi.org/10.1177/1745691612447307

Chepenik, L. G., Cornew, L. A., & Farah, M. J. (2007). The influence of sad mood on cognition. *Emotion*, *7*(4), 802. https://doi.org/10.1037/1528-3542.7.4.802

Christensen, J. F., Lambrechts, A., & Tsakiris, M. (2019). The Warburg Dance Movement Library—The WADAMO Library: A Validation Study. *Perception*, *48*(1), 26–57. https://doi.org/10.1177/0301006618816631

Christianson, S. Å., & Loftus, E. F. (1991). Remembering emotional events: The fate of detailed information. *Cognition & Emotion*, *5*(2), 81–108. https://doi.org/10.1080/02699939108411027

Codispoti, M., & De Cesarei, A. (2007). Arousal and attention: Picture size and emotional reactions. *Psychophysiology*, *44*(5), 680–686. https://doi.org/10.1111/j.1469-8986.2007.00545.x

Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences*, *106*(14), 6008–6010. https://doi.org/10.1073/pnas.0811884106

Cohen, R. L., Netley, C., & Clarke, M. A. (1984). On the Generality of the short-term memory-reading ability relationship. *Journal of Learning Disabilities*, *17*(4), 218–221. https://doi.org/10.1177/002221948401700406

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. *The Handbook of Emotion Elicitation and Assessment*, *1*(3), 203–221.

Cosker, D., Krumhuber, E., & Hilton, A. (2011). A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. Proceedings of the IEEE International Conference on Computer Vision, 2296–2303. https://doi.org/10.1109/ICCV.2011.6126510

Cosme, G., Tavares, V., Nobre, G., Lima, C., Sá, R., Rosa, P., & Prata, D. (2021). Cultural differences in vocal emotion recognition: a behavioural and skin conductance study in Portugal and Guinea-Bissau. *Psychological Research*, *86*(2), 597–616. https://doi.org/10.1007/s00426-021-01498-2

Costa Jr, P. T., & McCrae, R. R. (2006). Age changes in personality and their origins: comment on Roberts, Walton, and Viechtbauer (2006). *Psychological Bulletin*, *132*(1), 26–28. https://doi.org/10.1037/0033-2909.132.1.26

Costa, P. T. J., & McCrae, R. R. (1992). *The NEO-PI-R: Revised Personality Inventory (NEO-PI-R)*. Odessa, FL: Assessment Resources.

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *Journal of Personality and Social Psychology*, *38*(4), 668. https://doi.org/10.1037//0022-3514.38.4.668

Costa, P. T., & McCrae, R. R. (1989). *NEO five-factor inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

Costantini, G., Iadarola, I., Paoloni, A., & Todisco, M. (2014). EMOVO corpus: An Italian emotional speech database. Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, 3501–3504.

Cowie, R., Douglas-Cowie, E., Savvidou, S., & McMahon, E. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. 19-24. Paper presented at Speech and Emotion: Proceedings of the ISCA workshop, Newcastle, United Kingdom.

Cowie, R., Sawey, M., Doherty, C., Jaimovich, J., Fyans, C., & Stapleton, P. (2013). Gtrace: General trace program compatible with emotionML. In Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013 (pp. 709–710). https://doi.org/10.1109/ACII.2013.126

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Crane, E., & Gross, M. (2007). Motion capture and emotion: Affect detection in whole body movement. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4738 LNCS, 95–101. https://doi.org/10.1007/978-3-540-74889-2_9

Cullen, A., & Harte, N. (2018). A longitudinal database of Irish political speech with annotations of speaker ability. *Language Resources and Evaluation*, *52*(2), 401–432. https://doi.org/10.1007/s10579-017-9401-z

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/S1364-6613(03)00197-9

Dalrymple, K. A., Gomez, J., & Duchaine, B. (2013). The dartmouth database of children's faces: Acquisition and validation of a new face stimulus set. *PLoS ONE*, *8*(11), 1–7. https://doi.org/10.1371/journal.pone.0079131

Davis, E., Greenberger, E., Charles, S., Chen, C., Zhao, L., & Dong, Q. (2012). Emotion experience and regulation in China and the United States: how do culture and gender shape emotion responding? *International Journal of Psychology*, *47*(3), 230–239. https://doi.org/10.1080/00207594.2011.626043

Deacon, R. M., Bannerman, D. M., & Rawlins, J. N. P. (2002). Anxiolytic effects of cytotoxic hippocampal lesions in rats. *Behavioral Neuroscience*, *116*(3), 494. https://doi.org/10.1037//0735-7044.116.3.494

DeBusk, K. P. A., & Austin, E. J. (2011). Emotional intelligence and social perception. *Personality and Individual Differences*, *51*(6), 764–768. https://doi.org/10.1016/j.paid.2011.06.026

Deng, Y., Yang, M., & Zhou, R. (2017). A new standardized emotional film database for Asian culture. *Frontiers in Psychology*, *8*, 1941. https://doi.org/10.3389/fpsyg.2017.01941

Diconne, K., Kountouriotis, G. K., Paltoglou, A. E., Parker, A., & Hostler, T. J. (2022). Presenting KAPODI – The Searchable Database of Emotional Stimuli Sets. *Emotion Review*, *14*(1), 84–95. https://doi.org/10.1177/17540739211072803

Dijksterhuis, A. P., & Smith, P. K. (2002). Affective habituation: subliminal exposure to extreme stimuli decreases their extremity. *Emotion*, *2*(3), 203. https://doi.org/10.1037/1528-3542.2.3.203

Dissanayake, C., Sigman, M., & Kasari, C. (1996). Long-term stability of individual differences in the emotional responsiveness of children with autism. *Journal of Child Psychology and Psychiatry*, *37*, 461–468. https://doi.org/10.1111/j.1469-7610.1996.tb01427.x

Dolcos, F., & McCarthy, G. (2006). Brain systems mediating cognitive interference by emotional distraction. *Journal of Neuroscience*, *26*(7), 2072–2079. https://doi.org/10.1523/JNEUROSCI.5042-05.2006

Droit-Volet, S., Fayolle, S. L., & Gil, S. (2011). Emotion and time perception: effects of film-induced mood. *Frontiers in Integrative Neuroscience*, *5*, 33. https://doi.org/10.3389/fnint.2011.00033

Druschel, B. A., & Sherman, M. F. (1999). Disgust sensitivity as a function of the Big Five and gender. *Personality and Individual Differences*, *26*(4), 739–748. https://doi.org/10.1016/S0191-8869(98)00196-2

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review*, *66*(3), 183. https://doi.org/10.1037/h0047707

Eaton, L. G., & Funder, D. C. (2001). Emotional experience in daily life: valence, variability, and rate of change. *Emotion*, *1*(4), 413. https://doi.org/10.1037/1528-3542.1.4.413

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES-a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, *42*(1), 351–362. https://doi.org/10.3758/BRM.42.1.351

Eddie, D., & Bates, M. E. (2017). Toward validation of a borderline personality disorder–relevant picture set. *Personality Disorders: Theory, Research, and Treatment*, *8*(3), 255. https://doi.org/10.1037/per0000173

Edelman, M., Bourdieu, P., Thompson, J. B., Raymond, G., & Adamson, M. (1992). Language and Symbolic Power. *Contemporary Sociology*, *21*(5), 717. https://doi.org/10.2307/2075589

Eerola, T., & Vuoskoski, J. K. (2012). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, *30*(3), 307–340. https://doi.org/10.1525/mp.2012.30.3.307

Eichenbaum, H. (2001). The hippocampus and declarative memory: cognitive mechanisms and neural codes. *Behavioural Brain Research*, *127*(1-2), 199–207. https://doi.org/10.1016/S0166-4328(01)00365-5

Ekkekakis, P. (2012). Affect, Mood, and Emotion. *Measurement in Sport and Exercise Psychology*, 321-332.

Ekman, P. (1976). *Pictures of facial affect*. Consulting Psychologists Press.

Ekman, P. (1982). *Methods for measuring facial action*. In K. R. Scherer & P Ekman (Eds.), Handbook of Methods in Nonverbal Behaviour Research (pp. 45–135). New York: Cambridge University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion.* New York: Wiley.

Ekman, P., & Friesen, W. V. (1978). *The Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, CA.

Ekman, P., Friesen, W., & Hager, J. (2002). *Facial Action Coding System*. Manual and Investigator's Guide, Salt Lake City, UT: Research Nexus

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86–88. https://doi.org/10.1126/science.164.3875.86

Esposito, A., Riviello, M. T., & Di Maio, G. (2009). The COST 2102 Italian Audio and Video Emotional Database. *Frontiers in Artificial Intelligence and Applications*, *204*, 51–61. https://doi.org/10.3233/978-1-60750-072-8-51

Evers, A., Rasche, J., & Schabracq, M. J. (2008). High sensory-processing sensitivity at work. *International Journal of Stress Management*, *15*(2), 189. https://doi.org/10.1037/1072-5245.15.2.189

Fabbro, F., Naatanen, R., & Kujala, T. (1999). The neurolinguistics of bilingualism. *Nature*, *398*(6728), 577–577.

Fagan III, J. F. (1972). Infants' recognition memory for faces. *Journal of Experimental Child Psychology*, *14*(3), 453–476. https://doi.org/10.1016/0022-0965(72)90065-3

Fairfield, B., Ambrosini, E., Mammarella, N., & Montefinese, M. (2017). Affective norms for italian words in older adults: Age differences in ratings of valence, arousal and dominance. *PLoS ONE*, *12*(1), 1–22. https://doi.org/10.1371/journal.pone.0169472

Fanselow, M. S., & Dong, H. W. (2010). Are the dorsal and ventral hippocampus functionally distinct structures? *Neuron*, *65*(1), 7–19. https://doi.org/10.1016/j.neuron.2009.11.031

Farewell, V. T., Tom, B. D. M., & Royston, P. (2004). The impact of dichotomization on the efficiency of testing for an interaction effect in exponential family models. *Journal of the American Statistical Association*, *99*(467), 822–831. https://doi.org/10.1198/016214504000001169

Ferrari, V., Mastria, S., & Codispoti, M. (2020). The interplay between attention and long-term memory in affective habituation. *Psychophysiology*, *57*(6), e13572. https://doi.org/10.1111/psyp.13572

Ferré, P., Guasch, M., Moldovan, C., & Sánchez-Casas, R. (2012). Affective norms for 380 Spanish words belonging to three different semantic categories. *Behavior Research Methods*, *44*(2), 395–403. https://doi.org/10.3758/s13428-011-0165-x

Ferry, B., & McGaugh, J. L. (2000). Role of amygdala norepinephrine in mediating stress hormone regulation of memory storage. *Acta Pharmacologica Sinica*, *21*(6), 481-493.

Fischer, A. H., & Manstead, A. S. (2000). The relation between gender and emotions in different cultures. *Gender and Emotion: Social Psychological Perspectives*, *1*, 71-94. https://doi.org/10.1017/CBO9780511628191.005

Fischer, A. H., Rodriguez Mosquera, P. M., Van Vianen, A. E., & Manstead, A. S. (2004). Gender and culture differences in emotion. *Emotion*, *4*(1), 87. https://doi.org/10.1037/1528-3542.4.1.87

Frowd, C. D., Matuszewski, B. J., Shark, L. K., & Quan, W. (2009). Towards a comprehensive 3D dynamic facial expression database. Proceedings of the 9th WSEAS International Conference on Signal, Speech and Image Processing, SSIP ’09, Proc. 9th WSEAS Int. Conf. Multimedia, Internet and Video Technologies, MIV ’09, 113–119.

Galea, S., & Lindell, A. K. (2016). Do the Big Five personality traits predict individual differences in the left cheek bias for emotion perception? *Laterality: Asymmetries of Body, Brain and Cognition*, *21*(3), 200–214. https://doi.org/10.1080/1357650X.2016.1146738

Garrido, M. V., Lopes, D., Prada, M., Rodrigues, D., Jerónimo, R., & Mourão, R. P. (2017). The many faces of a face: Comparing stills and videos of facial expressions in eight dimensions (SAVE database). *Behavior Research Methods*, *49*(4), 1343–1360. https://doi.org/10.3758/s13428-016-0790-5

Ge, Y., Zhao, G., Zhang, Y., Houston, R. J., & Song, J. (2019). A standardised database of Chinese emotional film clips. *Cognition and Emotion*, *33*(5), 976–990. https://doi.org/10.1080/02699931.2018.1530197

Gerstenberg, F. X. (2012). Sensory-processing sensitivity predicts performance on a visual search task followed by an increase in perceived stress. *Personality and Individual Differences*, *53*(4), 496–500. https://doi.org/10.1016/j.paid.2012.04.019

Gil, S., & Droit-Volet, S. (2012). Emotional time distortions: the fundamental role of arousal. *Cognition & Emotion*, *26*(5), 847–862. https://doi.org/10.1080/02699931.2011.625401

Gilman, T. L., Shaheen, R., Nylocks, K. M., Halachoff, D., Chapman, J., Flynn, J. J., Matt, L. M., & Coifman, K. G. (2017). A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation. *Behavior Research Methods*, *49*(6), 2061–2082. https://doi.org/10.3758/s13428-016-0842-x

Glascher, J., Adolphs, R., 2003. Processing of the arousal of subliminal and supraliminal emotional stimuli by the human amygdala. *Journal of Neuroscience*, *23*(32), 10274–10282. https://doi.org/10.1523/JNEUROSCI.23-32-10274.2003

Golan, O., Baron-Cohen, S., & Hill, J. (2006a). The Cambridge mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *Journal of Autism and Developmental Disorders*, *36*(2), 169–183. https://doi.org/10.1007/s10803-005-0057-y

Goldberg, R., & Todman, M. (2018). Induced boredom suppresses the recall of positively valenced information: A preliminary study. *Psychological Thought*, *11*(1). https://doi.org/10.5964/psyct.v11i1.249

Goodman, A. M., Katz, J. S., & Dretsch, M. N. (2016). Military Affective Picture System (MAPS): A new emotion-based stimuli set for assessing emotional processing in military populations. *Journal of Behavior Therapy and Experimental Psychiatry*, *50*, 152–161. https://doi.org/10.1016/j.jbtep.2015.07.006

Graham, J. R. (1987). *The MMPI: A practical guide*. Oxford University Press.

Gray, J. A. (1982). Précis of The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system. *Behavioral and Brain Sciences*, *5*(3), 469–484. https://doi.org/10.1017/S0140525X00013066

Gray, J. R. (2001). Emotional modulation of cognitive control: Approach–withdrawal states double-dissociate spatial from verbal two-back task performance. *Journal of Experimental Psychology: General*, *130*(3), 436. https://doi.org/10.1037//0096-3445.130.3.436

Greenberg, L.S. (2002). *Emotion-focused therapy: Coaching clients to work through their feelings*. Washington, DC: American Psychological Association.

Gross, J. J. (2015). Emotion Regulation: Current Status and Future Prospects. *Psychological Inquiry*, *26*(1), 1–26. https://doi.org/10.1080/1047840X.2014.940781

Gross, J. J., & Levenson, R. W. (1993). Emotional suppression: physiology, self-report, and expressive behavior. *Journal of Personality and Social Psychology*, *64*(6), 970. https://doi.org/10.1037//0022-3514.64.6.970

Gross, J. J., & Levenson, R. W. (1995). Emotion Elicitation using Films. *Cognition and Emotion*, *9*(1), 87–108. https://doi.org/10.1080/02699939508408966

Grühn, D., & Sharifian, N. (2016). *Lists of emotional stimuli*. In Emotion measurement (pp. 145–164). Woodhead Publishing. https://doi.org/10.1016/b978-0-08-100508-8.00007-2

Gunes, H., & Piccardi, M. (2006, August). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In 18th International Conference on Pattern Recognition (ICPR'06), 1, 1148–1153. https://doi.org/10.1109/icpr.2006.39

Gur, R. C., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., Macy, L., ... & Gur, R. E. (2002). A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. *Journal of Neuroscience Methods*, *115*(2), 137–143. https://doi.org/10.1016/s0165-0270(02)00006-7

Gutiérrez-Maldonado, J., Rus-Calafell, M., & González-Conde, J. (2014). Creation of a new set of dynamic virtual reality faces for the assessment and training of facial emotion recognition ability. *Virtual Reality*, *18*(1), 61–71. https://doi.org/10.1007/s10055-013-0236-7

Haas, A. (1979). Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, *86*(3), 616. https://psycnet.apa.org/doi/10.1037/0033-2909.86.3.616

Haberkamp, A., Glombiewski, J. A., Schmidt, F., & Barke, A. (2017). The DIsgust-RelaTed-Images (DIRTI) database: Validation of a novel standardized set of disgust pictures. *Behaviour Research and Therapy*, *89*, 86–94. https://doi.org/10.1016/j.brat.2016.11.010

Hamann, S. B., Cahill, L., McGaugh, J. L., & Squire, L. R. (1997). Intact enhancement of declarative memory for emotional material in amnesia. *Learning & Memory*, *4*(3), 301–309. https://doi.org/10.1101/lm.4.3.301

Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, *17*(1), 317–323. https://doi.org/10.1006/nimg.2002.1179

Harmon, L. D. (1973). The recognition of faces. *Scientific American*, *229*(5), 70–83.

Haro, J., Ferré, P., Boada, R., & Demestre, J. (2017). Semantic ambiguity norms for 530 Spanish words. *Applied Psycholinguistics*, *38*(2), 457–475. https://doi.org/10.1017/S0142716416000266

Harris, L. M. (1999). Mood and prospective memory. *Memory*, *7*(1), 117–127. https://doi.org/10.1080/741943717

Harris, L. M., & Cumming, S. R. (2003). An examination of the relationship between anxiety and performance on prospective and retrospective memory tasks. *Australian Journal of Psychology*, *55*(1), 51–55. https://doi.org/10.1080/00049530412331312874

Hasson, D., & Arnetz, B. B. (2005). Validation and findings comparing VAS vs. Likert scales for psychosocial measurements. *International Electronic Journal of Health Education*, *8*, 178–192.

Hayes, M. H., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*(1), 98–99.

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, *19*(2), 49–71. https://doi.org/10.1080/10478400802049936

Hermes, M., Hagemann, D., Naumann, E., & Walter, C. (2011). Extraversion and its positive emotional core—Further evidence from neuroscience. *Emotion*, *11*(2), 367. https://doi.org/10.1037/a0021550

Herpertz, S., Schütz, A., & Nezlek, J. (2016). Enhancing emotion perception, a fundamental component of emotional intelligence: Using multiple-group SEM to evaluate a training program. *Personality and Individual Differences*, *95*, 11–19. https://doi.org/10.1016/j.paid.2016.02.015

Hewig, J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., & Bartussek, D. (2005). A revised film set for the induction of basic emotions. *Cognition and Emotion*, *19*(7), 1095–1109. https://doi.org/10.1080/02699930541000084

Hinojosa, R. (2010). Doing hegemony: Military, men, and constructing a hegemonic masculinity. *Journal of Men's Studies*, *18*, 179 –194. http://dx.doi.org/10.3149/jms.1802.179

Holland, C. A. C., Ebner, N. C., Lin, T., & Samanez-Larkin, G. R. (2019). Emotion identification across adulthood using the Dynamic FACES database of emotional expressions in younger, middle aged, and older adults. *Cognition and Emotion*, *33*(2), 245–257. https://doi.org/10.1080/02699931.2018.1445981

Hostler, T. J., Wood, C., & Armitage, C. J. (2018). The influence of emotional cues on prospective memory: a systematic review with meta-analyses. *Cognition and Emotion*, *32*(8), 1578–1596. https://doi.org/10.1080/02699931.2017.1423280

Humphreys, L. G. (1978). Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. *Journal of Educational Psychology*, *70*(6), 873–876. https://psycnet.apa.org/doi/10.1037/0022-0663.70.6.873

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., & Popovich, D. L. (2015). The median split: Robust, refined, and revived. *Journal of Consumer Psychology*, *25*(4), 690–704. https://doi.org/10.1016/j.jcps.2015.06.014

Imbir, K. K. (2015). Affective norms for 1,586 polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods*, *47*(3), 860–870. https://doi.org/10.3758/s13428-014-0509-4

Imbir, K. K. (2016). Affective norms for 718 polish short texts (ANPST): Dataset with affective ratings for valence, arousal, dominance, origin, subjective significance and source dimensions. *Frontiers in Psychology*, *7*(JUL), 1–5. https://doi.org/10.3389/fpsyg.2016.01030

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

Isen, A. M. (1985). Asymmetry of happiness and sadness in effects on memory in normal college students: Comment on Hasher, Rose, Zacks, Sanft, and Doren. *Journal of*

Experimental Psychology: General, *114*(3), 388. https://psycnet.apa.org/doi/10.1037/0096-3445.114.3.388

Izard, C. E. (2007). Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280. https://doi.org/10.1111/j.1745-6916.2007.00044.x

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, *2*(3), 260-280. https://doi.org/10.1111/j.1745-6916.2007.0004

Jagiellowicz J. (2012). The relationship between the temperament trait of sensory processing sensitivity and emotional reactivity. New York: Doctoral Dissertation at Stony Brook University. Available at http://dspace.sunyconnect.suny.edu/bitstream/handle/1951/59701/Jagiellowicz_grad.sunysb_0771E_10998.pdf?sequence=1.

Jagiellowicz, J., Aron, A., & Aron, E. N. (2016). Relationship between the temperament trait of sensory processing sensitivity and emotional reactivity. *Social Behavior and Personality: an International Journal*, *44*(2), 185–199. https://doi.org/10.2224/sbp.2016.44.2.185

Jagiellowicz, J., Xu, X., Aron, A., Aron, E., Cao, G., Feng, T., & Weng, X. (2011). The trait of sensory processing sensitivity and neural responses to changes in visual scenes. *Social Cognitive and Affective Neuroscience*, *6*(1), 38–47. https://doi.org/10.1093/scan/nsq001

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior Research Methods*, *40*(4), 1065–1074. https://doi.org/10.3758/BRM.40.4.1065

Jonesgotman, M., & Zatorre, R. J. (1993). Odor Recognition Memory in Humans: Role of Right Temporal and Orbitofrontal Regions. *Brain and Cognition, 22*(2), 182–198. https://doi.org/10.1006/brcg.1993.1033

Kagan, J. (1994). Galen's prophecy: Temperament in human nature. New York: Basic Books. https://doi.org/10.4324/9780429500282

Kamp, S. M., Potts, G. F., & Donchin, E. (2015). On the roles of distinctiveness and semantic expectancies in episodic encoding of emotional words. *Psychophysiology*, *52*(12), 1599–1609. https://doi.org/10.1111/psyp.12537

Kanske, P., & Kotz, S. A. (2011). Cross-modal validation of the Leipzig Affective Norms for German (LANG). *Behavior Research Methods*, *43*(2), 409–413. https://doi.org/10.3758/s13428-010-0048-6

Kanske, P., & Kotz, S. A. (2012). Auditory affective norms for German: Testing the influence of depression and anxiety on valence and arousal ratings. *PLoS ONE*, *7*(1), 1–6. https://doi.org/10.1371/journal.pone.0030086

Katsimerou, C., Albeda, J., Huldtgren, A., Heynderickx, I., & Redi, J. A. (2016). Crowdsourcing empathetic intelligence: The case of the annotation of EMMA database for emotion and mood recognition. *ACM Transactions on Intelligent Systems and Technology*, *7*(4). https://doi.org/10.1145/2897369

Keefe, B. D., Villing, M., Racey, C., Strong, S. L., Wincenciak, J., & Barraclough, N. E. (2014). A database of whole-body action videos for the study of action, emotion, and untrustworthiness. *Behavior Research Methods*, *46*(4), 1042–1051. https://doi.org/10.3758/s13428-013-0439-6

Kemp, A. H., Silberstein, R. B., Armstrong, S. M., & Nathan, P. J. (2004). Gender differences in the cortical electrophysiological processing of visual emotional stimuli. *NeuroImage*, *21*(2), 632–646. https://doi.org/10.1016/j.neuroimage.2003.09.055

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, *31*(8), 1169–1180. https://doi.org/10.3758/bf03195800

Kensinger, E. A., & Corkin, S. (2004). Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proceedings of the National Academy of Sciences*, *101*(9), 3310–3315. https://doi.org/10.1073/pnas.0306408101

Kensinger, E. A., & Schacter, D. L. (2006). Processing emotional pictures and words: Effects of valence and arousal. *Cognitive, Affective, & Behavioral Neuroscience*, *6*(2), 110–126. https://doi.org/10.3758/CABN.6.2.110

Kensinger, E. A., Garoff-Eaton, R. J., & Schacter, D. L. (2007). How negative emotion enhances the visual specificity of a memory. *Journal of Cognitive Neuroscience*, *19*(11), 1872–1887. https://doi.org/10.1162/jocn.2007.19.11.1872

Keshtiari, N., Kuhlmann, M., Eslami, M., & Klann-Delius, G. (2015). Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD). *Behavior Research Methods*, *47*(1), 275–294. https://doi.org/10.3758/s13428-014-0467-x

Kim, E. Y., Lee, S. H., Park, G., Kim, S., Kim, I., Chae, J. H., & Kim, H. T. (2013). Gender difference in event related potentials to masked emotional stimuli in the oddball task. *Psychiatry investigation*, *10*(2), 164. https://doi.org/10.4306/pi.2013.10.2.164

Kinsbourne, M., & George, J. (1974). The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(1), 63–69. https://doi.org/10.1016/S0022-5371(74)80031-9

Kirschbaum, C., Wolf, O. T., May, M., Wippich, W., & Hellhammer, D. H. (1996). Stress-and treatment-induced elevations of cortisol levels associated with impaired declarative memory in healthy adults. *Life Sciences*, *58*(17), 1475–1483. https://doi.org/10.1016/0024-3205(96)00118-x

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5*, 345-379. https://doi.org/10.1007/BF00992553

Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *41*(4), 1027–1038. https://doi.org/10.1109/tsmcb.2010.2103557

Kleinsmith, L. J., & Kaplan, S. (1963). Paired-associate learning as a function of arousal and interpolated interval. *Journal of Experimental Psychology, 65*(2), 190–193. https://doi.org/10.1037/h0040288

Kleinsmith, L. J., Kaplan, S., & Trate, R. D. (1963). The relationship of arousal to short- and long-term verbal recall. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 17*(4), 393–397. https://doi.org/10.1037/h0083278

Klüver, H., & Bucy, P. C. (1937). "Psychic blindness" and other symptoms following bilateral temporal lobectomy in rhesus monkeys. *American Journal of Physiology*, *119*, 352–353.

Kraha, A., Talarico, J. M., & Boals, A. (2014). Unexpected positive events do not result in flashbulb memories. *Applied Cognitive Psychology*, *28*(4), 579–589. https://doi.org/10.1002/acp.3039

Kring, A. M., & Gordon, A. H. (1998). Sex Differences in Emotion: Expression, Experience, and Physiology. *Journal of Personality and Social Psychology*, *74*(3), 686–703. https://doi.org/10.1037/0022-3514.74.3.686

Krumhuber, E. G., Skora, L., Küster, D., & Fou, L. (2017). A review of dynamic datasets for facial expression research. *Emotion Review*, *9*(3), 280–292. https://doi.org/10.1177/1754073916670022

Kuhbandner, C., & Pekrun, R. (2013). Joint effects of emotion and color on memory. *Emotion, 13*(3), 375–379. https://doi.org/10.1037/a0031821

Kuhlmann, S., & Wolf, O. T. (2006). Arousal and cortisol interact in modulating memory consolidation in healthy young men. *Behavioral Neuroscience*, *120*(1), 217. https://doi.org/10.1037/0735-7044.120.1.217

Kuyper, L., & Wijsen, C. (2014). Gender identities and gender dysphoria in the Netherlands. *Archives of Sexual Behavior*, *43*(2), 377–385. https://doi.org/10.1007/s10508-013-0140-y

Kuypers, K. P. (2017). Emotional empathic responses to dynamic negative affective stimuli is gender-dependent. *Frontiers in Psychology*, *8*, 1491. https://doi.org/10.3389/fpsyg.2017.01491

Laney, C., Campbell, H. V., Heuer, F., & Reisberg, D. (2004). Memory for thematically arousing events. *Memory & Cognition*, *32*(7), 1149–1159. https://doi.org/10.3758/BF03196888

Lang, P. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications. *Technology in Mental Health Care Delivery Systems*, 119–137.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, *1*, 39–58.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual (Report No. A-8). Gainesville, FL: University of Florida, NIMH Center for the Study of Emotion and Attention.

Lang, P. J., Bradley, M. M., Cuthbert, B. N. (1999). International affective picture system (IAPS): Instruction manual and affective ratings. The center for research in psychophysiology, University of Florida.

Lange, C. G., & James, W. (1922). The Emotions (Volume I). *Retrieved from: archive. org/details/emotionsvolumei007644mbp*. https://doi.org/10.1037/10735-000

Lassalle, A., Pigat, D., O'Reilly, H., Berggen, S., Fridenson-Hayo, S., Tal, S., Elfström, S., Råde, A., Golan, O., Bölte, S., Baron-Cohen, S., & Lundqvist, D. (2019). The EU-Emotion Voice Database. *Behavior Research Methods*, *51*(2), 493–506. https://doi.org/10.3758/s13428-018-1048-1

Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., & Althoff, J. (2010). Presenting the VENEC Corpus: Development of a Cross-Cultural Corpus of Vocal Emotion Expressions and a Novel Method of Annotating Emotion Appraisals. Lrec 2010 - Seventh International Conference on Language Resources and Evaluation, 7, 53–57.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lepping, R. J., Atchley, R. A., & Savage, C. R. (2016). Development of a validated emotionally provocative musical stimulus set for research. *Psychology of Music*, *44*(5), 1012–1028. https://doi.org/10.1177/0305735615604509

Li, J., Tian, M., Fang, H., Xu, M., Li, H., & Liu, J. (2010). Extraversion predicts individual differences in face recognition. *Communicative & Integrative Biology*, *3*(4), 295–298. https://doi.org/10.4161/cib.3.4.12093

Li, Y., Tao, J., Chao, L., Bao, W., & Liu, Y. (2017). CHEAVD: a Chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, *8*(6), 913–924. https://doi.org/10.1007/s12652-016-0406-z

Libkuman, T., Stabler, C., & Otani, H. (2004). Arousal, valence, and memory for detail. *Memory*, *12*(2), 237–247. https://doi.org/10.1080/09658210244000630

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 44–53.

Likforman-Sulem, L., Esposito, A., Faundez-Zanuy, M., Clemencon, S., & Cordasco, G. (2017). EMOTHAW: A Novel Database for Emotional State Recognition from Handwriting and Drawing. *IEEE Transactions on Human-Machine Systems*, *47*(2), 273–284. https://doi.org/10.1109/THMS.2016.2635441

Liss, M., Saulnier, C., Fein, D., & Kinsbourne, M. (2006). Sensory and attention abnormalities in autistic spectrum disorders. *Autism*, *10*(2), 155–172. https://doi.org/10.1177/1362361306062021

Lithari, C., Frantzidis, C. A., Papadelis, C., Vivas, A. B., Klados, M. A., Kourtidou-Papadeli, C., Pappas, C., Ioannides, A. A., & Bamidis, P. D. (2010). Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. *Brain Topography*, *23*(1), 27–40. https://doi.org/10.1007/s10548-009-0130-5

Litten, V., Roberts, L. D., Ladyshewsky, R. K., Castell, E., & Kane, R. (2020). Empathy and psychopathic traits as predictors of selection into business or psychology disciplines. *Australian Journal of Psychology*, *72*(1), 93–105. https://doi.org/10.1111/ajpy.12263

Little, A. C. (2013). The influence of steroid sex hormones on the cognitive and emotional processing of visual stimuli in humans. *Frontiers in Neuroendocrinology*, *34*(4), 315–328. https://doi.org/10.1016/j.yfrne.2013.07.009

LoBue, V., & Thrasher, C. (2014). The Child Affective Facial Expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, *5*(OCT), 1–8. https://doi.org/10.3389/fpsyg.2014.01532

López-Caneda, E., & Carbia, C. (2018). The Galician Beverage Picture Set (GBPS): A standardized database of alcohol and non-alcohol images. *Drug and Alcohol Dependence*, *184*, 42–47. https://doi.org/10.1016/j.drugalcdep.2017.11.022

Lubis, N., Heck, M., Sakti, S., Yoshino, K., & Nakamura, S. (2018). Processing negative emotions through social communication: Multimodal database construction and analysis. 2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017, 2018-Janua, 79–85. https://doi.org/10.1109/ACII.2017.8273582

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989x.7.1.19

Maffei, A., & Angrilli, A. (2019). E-MOVIE - Experimental MOVies for Induction of Emotions in neuroscience: An innovative film database with normative data and sex differences. *PLoS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0223124

Magalhães, S. D. S., Miranda, D. K., Miranda, D. M. D., Malloy-Diniz, L. F., & Romano-Silva, M. A. (2018). The Extreme Climate Event Database (EXCEED): Development of a picture database composed of drought and flood stimuli. *PloS ONE*, *13*(9), e0204093. https://doi.org/10.1371/journal.pone.0204093

Marcell, M. M., Borella, D., Greene, M., Kerr, E., & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology*, *22*(6), 830–864. https://doi.org/10.1076/jcen.22.6.830.949

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods*, *46*(2), 596-610. https://doi.org/10.3758%2Fs13428-013-0379-1

Maren, S. (1999). Long-term potentiation in the amygdala: a mechanism for emotional learning and memory. *Trends in Neurosciences*, *22*(12), 561–567. https://doi.org/10.1016/s0166-2236(99)01465-4

Marin, M. M., & Leder, H. (2016). Effects of presentation duration on measures of complexity in affective environmental scenes and representational paintings. *Acta Psychologica*, *163*, 38–58. https://doi.org/10.1016/j.actpsy.2015.10.002

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. *Journal of Personality and Social Psychology*, *88*(1), 139. https://doi.org/10.1037/0022-3514.88.1.139

Marx, B. P., Marshall, P. J., & Castro, F. (2008). The moderating effects of stimulus valence and arousal on memory suppression. *Emotion*, *8*(2), 199. https://doi.org/10.1037/1528-3542.8.2.199

Mather, M. (2007). Emotional Arousal and Memory Binding: An Object-Based Framework. *Perspectives on Psychological Science*, *2*(1), 33–52. https://doi.org/10.1111/j.1745-6916.2007.00028.x

Mather, M., & Sutherland, M. (2009). Disentangling the effects of arousal and valence on memory for intrinsic details. *Emotion Review*, *1*(2), 118-119. http://dx.doi.org/10.1177/1754073908100435

Matsumoto, D. (1989). Cultural influences on the perception of emotion. *Journal of Cross-Cultural Psychology*, *20*(1), 92-105. https://doi.org/10.1177/0022022189201006

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., Yrizarry, N., Loewinger, S., Uchida, H., Yee, A., Amo, L., & Goh, A. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian brief affect recognition test (JACBART). *Journal of Nonverbal Behavior*, *24*(3), 179–209. https://doi.org/10.1023/A:1006668120583

Matsumoto, D., Yoo, S. H., Nakagawa, S., Alexandre, J., Altarriba, J., Anguas-Wong, A. M., Arriola, M., Bauer, L. M., Bond, M. H., Cabecinhas, R., Chae, J., Comunian, A. L., DeGere, D. N., de Melo Garcia Bley, L., Fok, H. K., Friedlmeier, W., Garcia, F. M., Ghosh, A., Granskaya, J. V., … Yoo, S. H. (2008). Culture, Emotion Regulation, and Adjustment. *Journal of Personality and Social Psychology*, *94*(6), 925–937. https://doi.org/10.1037/0022-3514.94.6.925

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., & Cohn, J. F. (2013). DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, *4*(2), 151–160. https://doi.org/10.1109/T-AFFC.2013.4

May, A. K., Norris, S. A., Richter, L. M., & Pitman, M. M. (2020). A psychometric evaluation of the Highly Sensitive Person Scale in ethnically and culturally heterogeneous South African samples. *Current Psychology*, 1–15. https://doi.org/10.1007/s12144-020-00988-7

McCrae, R. R., & Costa Jr, P. T. (1989). Rotation to maximize the construct validity of factors in the NEO Personality Inventory. *Multivariate Behavioral Research*, *24*(1), 107–124. https://doi.org/10.1207/s15327906mbr2401_7

McCurrie, C. H., Crone, D. L., Bigelow, F., & Laham, S. M. (2018). Moral and Affective Film Set (MAAFS): A normed moral video database. *PloS ONE*, *13*(11), e0206604. https://doi.org/10.1371/journal.pone.0206604

McDuff, D., Amr, M., & Kaliouby, R. El. (2019). AM-FED+: An Extended Dataset of Naturalistic Facial Expressions Collected in Everyday Settings. *IEEE Transactions on Affective Computing*, *10*(1), 7–17. https://doi.org/10.1109/TAFFC.2018.2801311

McGaugh, J. L. (2000) Memory: A century of consolidation. *Science*, *287*(5451), 248–251. https://doi.org/10.1126/science.287.5451.248

McGaugh, J. L., McIntyre, C. K., & Power, A. E. (2002). Amygdala modulation of memory consolidation: interaction with other brain systems. *Neurobiology of Learning and Memory*, *78*(3), 539–552. https://doi.org/10.1006/nlme.2002.4082

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Memon, S. A., Dhamyal, H., Wright, O., Justice, D., Palat, V., Boler, W., ... & Singh, R. (2019). Detecting gender differences in perception of emotion in crowdsourced data. arXiv preprint arXiv:1910.11386.

Miccoli, L., Delgado, R., Guerra, P., Versace, F., Rodríguez-Ruiz, S., & Fernández-Santaella, M. C. (2016). Affective pictures and the open library of affective foods (OLAF): Tools to investigate emotions toward food in adults. *PLoS ONE*, *11*(8), 1–13. https://doi.org/10.1371/journal.pone.0158991

Miccoli, L., Delgado, R., Rodriguez-Ruiz, S., Guerra, P., García-Mármol, E., & Fernández-Santaella, F. (2014). Meet OLAF, a good friend of the IAPS! the Open Library of Affective Foods: A tool to investigate the emotional impact of food in adolescents. *PLoS ONE*, *9*(12), 1–22. https://doi.org/10.1371/journal.pone.0114515

Mill, A., Allik, J., Realo, A., & Valk, R. (2009). Age-related differences in emotion recognition ability: a cross-sectional study. *Emotion*, *9*(5), 619. https://doi.org/10.1037/a0016562

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, and Computers*, *36*(4), 630–633. https://doi.org/10.3758/BF03206543

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, *46*(3), 887–903. https://doi.org/10.3758/s13428-013-0405-3

Moors, A. (2009). Theories of emotion causation: A review. *Cognition and Emotion*, *23*(4), 625-662. https://doi.org/10.1080/02699930802645739

Morris, W. M. (1992). A functional analysis of the role of mood in affective systems. In M.s. Clark (Ed.), *Review of Personality and Social Psychology* (Vol. 13, pp. 256-293). Newbury Park, CA: Sage.

Mulac, A., Bradac, J. J., & Gibbons, P. (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research*, *27*(1), 121–152. https://doi.org/10.1111/j.1468-2958.2001.tb00778.x

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1-9.

Murphy, N. A., & Isaacowitz, D. M. (2008). Preferences for emotional information in older and younger adults: a meta-analysis of memory and attention tasks. *Psychology and Aging*, *23*(2), 263. https://doi.org/10.1037/0882-7974.23.2.263

Myers, I. B. (1962). *The Myers-Briggs Type Indicator: Manual (1962)*. Consulting Psychologists Press. https://doi.org/10.1037/14404-000

Nalloor, R., Bunting, K. M., & Vazdarjanova, A. (2012). Encoding of emotion-paired spatial stimuli in the rodent hippocampus. *Frontiers in Behavioral Neuroscience*, *6*, 27. https://doi.org/10.3389/fnbeh.2012.00027

Nater, U. M., Abbruzzese, E., Krebs, M., & Ehlert, U. (2006). Sex differences in emotional and psychophysiological responses to musical stimuli. *International Journal of Psychophysiology*, *62*(2), 300–308. https://doi.org/10.1016/j.ijpsycho.2006.05.011

Nater, U. M., Krebs, M., & Ehlert, U. (2005). Sensation seeking, music preference, and psychophysiological reactivity to music. *Musicae Scientiae*, *9*(2), 239–254. https://doi.org/10.1177%2F102986490500900205

Nazareth, D. S., Jansen, M. P., Truong, K. P., Westerhof, G. J., & Heylen, D. (2019). MEMOA: Introducing the Multi-Modal Emotional Memories of Older Adults Database. 2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, 697–703. https://doi.org/10.1109/ACII.2019.8925462

Neelamegham, R. (2001). Treating an individual difference predictor as continuous or categorical. *Journal of Consumer Psychology*, *10*(1-2), 49–51.

Neiberg, D., & Elenius, K. (2008). Automatic recognition of anger in spontaneous speech. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2755–2758.

Nelis, D., Quoidbach, J., Mikolajczak, M., & Hansenne, M. (2009). Increasing emotional intelligence:(How) is it possible? *Personality and Individual Differences*, *47*(1), 36–41. https://doi.org/10.1016/j.paid.2009.01.046

Nielson, K. A., & Lorber, W. (2009). Enhanced post-learning memory consolidation is influenced by arousal predisposition and emotion regulation but not by stimulus valence or arousal. *Neurobiology of Learning and Memory*, *92*(1), 70–79. https://doi.org/10.1016/j.nlm.2009.03.002

Nieuwenhuis-Mark, R. E., Schalk, K., & De Graaf, N. (2009). Free recall and learning of emotional word lists in very elderly people with and without dementia. *American Journal of Alzheimer's Disease and Other Dementias*, *24*(2), 155–162. https://doi.org/10.1177/1533317508330561

Nordgren, L. F., Van Harreveld, F., & Van Der Pligt, J. (2006). Ambivalence, discomfort, and motivated information processing. *Journal of Experimental Social Psychology*, *42*(2), 252–258. https://doi.org/10.1016/j.jesp.2005.04.004

Norris, C. J., Chen, E. E., Zhu, D. C., Small, S. L., & Cacioppo, J. T. (2004). The interaction of social and emotional processes in the brain. *Journal of Cognitive Neuroscience*, *16*(10), 1818–1829. https://doi.org/10.1162/0898929042947847

Noulhiane, M., Mella, N., Samson, S., Ragot, R., & Pouthas, V. (2007). How emotional auditory stimuli modulate time perception. *Emotion*, *7*(4), 697. https://doi.org/10.1037/1528-3542.7.4.697

O'Toole, A. J., Harms, J., Snow, S. L., Hurst, D. R., Pappas, M. R., Ayyad, J. H., & Abdi, H. (2005). A video database of moving faces and people. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, *27*(5), 812–816. https://doi.org/10.1109/TPAMI.2005.90

Okon-Singer, H., Kofman, O., Tzelgov, J., & Henik, A. (2011). Using international emotional picture sets in countries suffering from violence. *Journal of Traumatic Stress*, *24*(2), 239–242. https://doi.org/10.1002/jts.20600

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3), 197. https://doi.org/10.1037/h0055737

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (No. 47). University of Illinois press. https://doi.org/10.1017/s0008413100018740

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, *5*, 210. https://doi.org/10.1186/s13643-016-0384-4

Palumbo, R., Di Domenico, A., Fairfield, B., & Mammarella, N. (2021). When twice is better than once: increased liking of repeated items influences memory in younger and older adults. *BMC psychology*, *9*(1), 1-10. https://doi.org/10.1186/s40359-021-00531-8

Parsons, C. E., Young, K. S., Craske, M. G., Stein, A. L., & Kringelbach, M. L. (2014). Introducing the Oxford Vocal (OxVoc) Sounds database: a validated set of non-acted affective sounds from human infants, adults, and domestic animals. *Frontiers in Psychology*, *5*, 562. https://doi.org/10.3389/fpsyg.2014.00562

Paulhus, D. L. (1991). Measurement and control of response bias. *Measurement of Personality and Social Psychological Attitudes*, *1.* http://dx.doi.org/10.1016/B978-0-12-590241-0.50006-X

Penfield, W., & Milner, B. (1958). Memory deficit produced by bilateral lesions in the hippocampal zone. *AMA Archives of Neurology & Psychiatry*, *79*(5), 475–497. https://doi.org/10.1001/archneurpsyc.1958.02340050003001

Perlman, S. B., Morris, J. P., Vander Wyk, B. C., Green, S. R., Doyle, J. L., & Pelphrey, K. A. (2009). Individual differences in personality predict how people look at faces. *PloS One*, *4*(6), e5952. https://doi.org/10.1371/journal.pone.0005952

Perunovic, W. Q. E., Heller, D., & Rafaeli, E. (2007). Within-person changes in the structure of emotion: The role of cultural identification and language. *Psychological Science*, *18*(7), 607–613. http://dx.doi.org/10.1111/j.1467-9280.2007.01947.x

Petridis, S., Martinez, B., & Pantic, M. (2013). The MAHNOB laughter database. *Image and Vision Computing*, *31*(2), 186–202. https://doi.org/10.1016/j.imavis.2012.08.014

Pollak, S. D., Camras, L. A., & Cole, P. M. (2019). Progress in understanding the emergence of human emotion. *Developmental Psychology*, *55*(9), 1801–1811. https://doi.org/10.1037/dev0000789

Pollatos, O., Kopietz, R., Linn, J., Albrecht, J., Sakar, V., Anzinger, A., ... & Wiesmann, M. (2007). Emotional stimulation alters olfactory sensitivity and odor judgment. *Chemical Senses*, *32*(6), 583-589. https://doi.org/10.1093/chemse/bjm027

Prada, M., Garrido, M. V., Camilo, C., & Rodrigues, D. L. (2018). Subjective ratings and emotional recognition of children's facial expressions from the CAFE set. *PLoS ONE*, *13*(12). https://doi.org/10.1371/journal.pone.0209644

Prada, M., Rodrigues, D., Silva, R. R., & Garrido, M. V. (2016). Lisbon Symbol Database (LSD): Subjective norms for 600 symbols. *Behavior Research Methods*, *48*(4), 1370–1382. https://doi.org/10.3758/s13428-015-0643-7

Pronk, T., van Deursen, D. S., Beraha, E. M., Larsen, H., & Wiers, R. W. (2015). Validation of the Amsterdam Beverage Picture Set: A Controlled Picture Set for Cognitive Bias Measurement and Modification Paradigms. *Alcoholism: Clinical and Experimental Research*, *39*(10), 2047–2055. https://doi.org/10.1111/acer.12853

Provost, E. M., Shangguan, Y., & Busso, C. (2015). UMEME: University of Michigan emotional mcgurk effect data set. *IEEE Transactions on Affective Computing*, *6*(4), 395–409. https://doi.org/10.1109/TAFFC.2015.2407898

Quas, J. A., & Lench, H. C. (2007). Arousal at encoding, arousal at retrieval, interviewer support, and children's memory for a mild stressor. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *21*(3), 289–305. https://doi.org/10.1002/acp.1279

Ramalingam, V. V., Pandian, A., Jaiswal, A., & Bhatia, N. (2018). Emotion detection from text. *Journal of Physics: Conference Series*, *1000*(1). https://doi.org/10.1088/1742-6596/1000/1/012027

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190. https://doi.org/10.1037//0033-295x.83.3.190

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*(1), 45–48. https://doi.org/10.1111%2F1467-9280.00008

Reeves, B., Lang, A., Kim, E. Y., & Tatar, D. (1999). The effects of screen size and message content on attention and arousal. *Media Psychology*, *1*(1), 49–67. http://dx.doi.org/10.1207/s1532785xmep0101_4

Reynaud, E., El-Khoury-Malhame, M., Blin, O., & Khalfa, S. (2012). Voluntary emotion suppression modifies psychophysiological responses to films. *Journal of Psychophysiology*, *26*(3), 116. https://doi.org/10.1027/0269-8803/a000074

Ribeiro, R. L., Pompéia, S., & Bueno, O. F. A. (2005). Comparison of Brazilian and American norms for the international affective picture system (IAPS). *Brazilian Journal of Psychiatry*, *27*(3), 208–215. https://doi.org/10.1590/s1516-44462005000300009

Riegel, M., Wierzba, M., Wypych, M., Żurawski, Ł., Jednoróg, K., Grabowska, A., & Marchewka, A. (2015). Nencki Affective Word List (NAWL): the cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behavior Research Methods*, *47*(4), 1222–1236. https://doi.org/10.3758/s13428-014-0552-1

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality*, *43*(2), 137–145. https://doi.org/10.1016/j.jrp.2008.12.015

Roberts, B. W., & Helson, R. (1997). Changes in culture, changes in personality: The influence of individualism in a longitudinal study of women. *Journal of Personality and Social Psychology*, *72*(3), 641. http://dx.doi.org/10.1037//0022-3514.72.3.641

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1. https://doi.org/10.1037/0033-2909.132.1.1

Rolls, E. T. (1990). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion*, *4*(3), 161-190. https://doi.org/10.1080/02699939008410795

Roozendaal, B. (2002). Stress and memory: opposing effects of glucocorticoids on memory consolidation and memory retrieval. *Neurobiology of Learning and Memory*, *78*(3), 578–595. https://doi.org/10.1006/nlme.2002.4080

Roozendaal, B. (2003). Systems mediating acute glucocorticoid effects on memory consolidation and retrieval. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *27*(8), 1213–1223. https://doi.org/10.1016/j.pnpbp.2003.09.015

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

Russell, J. A., & Feldman Barrett, L. (2009). Core affect. In D. Sander & KR Scherer (Eds.), *The Oxford Companion to Emotion and the Affective Sciences* (pp. 104). New York: Oxford University Press.

Sabini, J., & Silver, M. (2005). Ekman's basic emotions: Why not love and jealousy?. *Cognition and Emotion*, *19*(5), 693-712. https://doi.org/10.1080/02699930441000481

Sacco, A. M., de Paula Couto, M. C. P., & Koller, S. H. (2016). Construction and validation of the White, Pardo, and Black children picture set (BIC-multicolor). *Psychology and Neuroscience*, *9*(1), 68–78. https://doi.org/10.1037/pne0000040

Salovey, P. (2001). Applied emotional intelligence: Regulating emotions to become healthy, wealthy, and wise. In J. Ciarrochi, J. P. Forgas, & J. D. Mayer (Eds.), Emotional intelligence in everyday life: A scientific inquiry (pp. 168–184). Psychology Press.

Samaria, F. S., & Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. IEEE Workshop on Applications of Computer Vision - Proceedings, 138–142. https://doi.org/10.1109/acv.1994.341300

Samson, A. C., Kreibig, S. D., Soderstrom, B., Wade, A. A., & Gross, J. J. (2016). Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and Emotion*, *30*(5), 827–856. https://doi.org/10.1080/02699931.2015.1031089

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., & Akarun, L. (2008). Bosphorus database for 3D face analysis. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5372 LNCS, 47–56. https://doi.org/10.1007/978-3-540-89991-4_6

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379-399. https://doi.org/10.1037/h0046234

Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, *24*(7), 1153–1172. http://dx.doi.org/10.1080/02699930903274322

Schandry, R. (2011). Emotionen. In *Biologische Psychologie: Ein Lehrbuch* (3rd ed.). Weinheim, Deutschland: Beltz PVU.

Scherer, K. R., & Brosch, T. (2009). Culture-specific appraisal biases contribute to emotion dispositions. *European Journal of Personality: Published for the European Association of Personality Psychology*, *23*(3), 265–288. https://doi.org/10.1002%2Fper.714

Schmidtke, D. S., Schröder, T., Jacobs, A. M., & Conrad, M. (2014). ANGST: affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods*, *46*(4), 1108–1118. https://doi.org/10.3758/s13428-013-0426-y

Schmidtmann, G., Jennings, B. J., Sandra, D. A., Pollock, J., & Gold, I. (2020). The McGill Face Database: Validation and Insights Into the Recognition of Facial Expressions of Complex Mental States. *Perception*, *49*(3), 310–329. https://doi.org/10.1177/0301006620901671

Schneider, I. K., Veenstra, L., van Harreveld, F., Schwarz, N., & Koole, S. L. (2016). Let's not be indifferent about neutrality: Neutral ratings in the International Affective Picture System (IAPS) mask mixed affective responses. *Emotion*, *16*(4), 426. https://doi.org/10.1037/emo0000164

Schön, D., Boyer, M., Moreno, S., Besson, M., Peretz, I., & Kolinsky, R. (2008). Songs as an aid for language acquisition. *Cognition*, *106*(2), 975–983. https://doi.org/10.1016/j.cognition.2007.03.005

Schwabe, L., & Wolf, O. T. (2010). Learning under stress impairs memory formation. *Neurobiology of Learning and Memory*, *93*(2), 183–188. https://doi.org/10.1016/j.nlm.2009.09.009

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21. https://doi.org/10.1136/jnnp.20.1.11

Segerstrom, S. C. (2001). Optimism and attentional bias for negative and positive stimuli. *Personality and Social Psychology Bulletin*, *27*(10), 1334–1343. https://doi.org/10.1177%2F01461672012710009

Shankland, R., Favre, P., Corubolo, D., Méary, D., Flaudias, V., & Mermillod, M. (2019). Food-Cal: development of a controlled database of high and low calorie food matched with non-food pictures. *Eating and Weight Disorders*, *24*(6), 1041–1050. https://doi.org/10.1007/s40519-019-00687-8

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, *100*(2), 139–156. https://psycnet.apa.org/doi/10.1037/0033-2909.100.2.139

Sharot, T., & Yonelinas, A. P. (2008). Differential time-dependent effects of emotion on recollective experience and memory for contextual information. *Cognition*, *106*(1), 538–547. https://doi.org/10.1016/j.cognition.2007.03.002

Sianipar, A., van Groenestijn, P., & Dijkstra, T. (2016). Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology*, *7*, 1907. https://doi.org/10.3389/fpsyg.2016.01907

Sloan, D. M., Sege, C. T., McSweeney, L. B., Suvak, M. K., Shea, M. T., & Litz, B. T. (2010). Development of a borderline personality disorder-relevant picture stimulus set. *Journal of Personality Disorders*, *24*(5), 664–675. https://doi.org/10.1521/pedi.2010.24.5.664

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, *44*(1), 256–269. https://doi.org/10.3758/s13428-011-0131-7

Soares, A. P., Pinheiro, A. P., Costa, A., Frade, C. S., Comesaña, M., & Pureza, R. (2013). Affective auditory stimuli: Adaptation of the international affective digitized sounds (IADS-2) for European Portuguese. *Behavior Research Methods*, *45*(4), 1168–1181. https://doi.org/10.3758/s13428-012-0310-1

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, *3*(1), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

Song, T., Zheng, W., Lu, C., Zong, Y., Zhang, X., & Cui, Z. (2019). MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, *7*, 12177–12191. https://doi.org/10.1109/access.2019.2891579

Soravia, L. M., Witmer, J. S., Schwab, S., Nakataki, M., Dierks, T., Wiest, R., ... & Jann, K. (2016). Prestimulus default mode activity influences depth of processing and recognition in an emotional memory task. *Human Brain Mapping*, *37*(3), 924–932. http://dx.doi.org/10.1002/hbm.23076

Spinhoven, P., Elzinga, B. M., Hovens, J. G. F. M., Roelofs, K., van Oppen, P., Zitman, F. G., & Penninx, B. W. J. H. (2011). Positive and negative life events and personality traits in predicting course of depression and anxiety. *Acta Psychiatrica Scandinavica*, *124*(6), 462–473. https://doi.org/10.1111/j.1600-0447.2011.01753.x

Squire, L. R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195. https://doi.org/10.1037/0033-295x.99.2.195

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences*, *93*(24), 13515–13522. https://doi.org/10.1073%2Fpnas.93.24.13515

Stadthagen-González, H., Ferré, P., Pérez-Sánchez, M. A., Imbault, C., & Hinojosa, J. A. (2018). Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust,

anger, fear, and sadness. *Behavior Research Methods*, *50*(5), 1943–1952. https://doi.org/10.3758/s13428-017-0962-y

Stevenson, R. A., Stevenson, L. D., Rupp, H. A., Kim, S., Janssen, E., & James, T. W. (2011). Incorporating emotions specific to the sexual response into theories of emotion using the Indiana sexual and affective word set. *Archives of Sexual Behavior*, *40*(1), 59–78. https://doi.org/10.1007/s10508-010-9669-1

Storbeck, J., & Clore, G. L. (2005). With sadness comes accuracy; with happiness, false memory: Mood and the false memory effect. *Psychological Science*, *16*(10), 785–791. https://doi.org/10.1111/j.1467-9280.2005.01615.x

Strange, B. A., Kroes, M. C., Fan, J. E., & Dolan, R. J. (2010). Emotion causes targeted forgetting of established memories. *Frontier in Behavioral Neuroscience*, *4*, 175. https://doi.org/10.3389/fnbeh.2010.00175

Sutton, T. M., & Altarriba, J. (2016). Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behavior Research Methods*, *48*(2), 686–728. https://doi.org/10.3758/s13428-015-0598-8

Sylvester, T., Braun, M., Schmidtke, D., & Jacobs, A. M. (2016). The Berlin affective word list for children (kidBAWL): exploring processing of affective lexical semantics in the visual and auditory modalities. *Frontiers in Psychology*, *7*, 969. https://doi.org/10.3389%2Ffpsyg.2016.00969

Szwoch, M. (2014). On Facial Expressions and Emotions RGB-D Database. *Communications in Computer and Information Science*, *424*, 384–394. https://doi.org/10.1007/978-3-319-06932-6_37

Szymanska, M., Comte, A., Tio, G., Vidal, C., Monnin, J., Smith, C. C., Nezelof, S., & Vulliez-Coady, L. (2019). The Besançon affective picture set-adult (BAPS-Adult): Development and validation. *Psychiatry Research*, *271*(March 2018), 31–38. https://doi.org/10.1016/j.psychres.2018.11.005

Szymanska, M., Monnin, J., Noiret, N., Tio, G., Galdon, L., Laurent, E., Nezelof, S., & Vulliez-Coady, L. (2015). The Besançon Affective Picture Set-Adolescents (the BAPS-Ado): Development and validation. *Psychiatry Research*, *228*(3), 576–584. https://doi.org/10.1016/j.psychres.2015.04.055

Tamir, M., & Robinson, M. D. (2004). Knowing good from bad: the paradox of neuroticism, negative affect, and evaluative processing. *Journal of Personality and Social Psychology*, *87*(6), 913. https://doi.org/10.1037/0022-3514.87.6.913

Torkamani-Azar, M., Kanik, S. D., Vardan, A. T., Aydin, C., & Cetin, M. (2019). Emotionality of Turkish language and primary adaptation of affective English norms for Turkish. *Current Psychology,* 38, 273–294. https://doi.org/10.1007/s12144-018-0119-x

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

Tse, C. S., & Altarriba, J. (2009). The word concreteness effect occurs for positive, but not negative, emotion words in immediate serial recall. *British Journal of Psychology*, *100*(1), 91. https://doi.org/10.1348/000712608x318617

Tu, Y. Z., Lin, D. W., Suzuki, A., & Goh, J. O. S. (2018). East Asian young and older adult perceptions of emotional faces from an age- and sex-fair East Asian facial expression database. *Frontiers in Psychology*, *9*(NOV), 2358. https://doi.org/10.3389/fpsyg.2018.02358

Tulving, E. (1972). Episodic and Semantic Memory. Episodic and semantic memory. In E. Tulving & W. Donaldson, Organization of memory (pp. 381–403). Academic Press.

Tulving, E. (1983). *Elements of episodic memory*. Oxford Psychology Series.

Twist, J., & de Graaf, N. M. (2019). Gender diversity and non-binary presentations in young people attending the United Kingdom's National Gender Identity Development Service. *Clinical Child Psychology and Psychiatry*, *24*(2), 277–290. https://doi.org/10.1177/1359104518804311

Ucros, C. G. (1989). Mood state-dependent memory: A meta-analysis. *Cognition and Emotion*, *3*(2), 139–169. https://doi.org/10.1080/02699938908408077

Vaiman, M., Wagner, M. A., Caicedo, E., & Pereno, G. L. (2017). Development and validation of an Argentine set of facial expressions of emotion. *Cognition and Emotion*, *31*(2), 249–260. https://doi.org/10.1080/02699931.2015.1098590

Valstar, M., & Pantic, M. (2010, May). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect (p. 65).

Van Caenegem, E., Wierckx, K., Elaut, E., Buysse, A., Dewaele, A., Van Nieuwerburgh, F., De Cuypere, G., & T'Sjoen, G. (2015). Prevalence of gender nonconformity in Flanders, Belgium. *Archives of Sexual Behavior*, *44*(5), 1281–1287. https://doi.org/10.1007/s10508-014-0452-6

Van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving Faces, Looking Places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion*, *11*(4), 907–920. https://doi.org/10.1037/a0023853

Van Dyck, E., Vansteenkiste, P., Lenoir, M., Lesaffre, M., & Leman, M. (2014). Recognizing induced emotions of happiness and sadness from dance movement. *PloS ONE*, *9*(2), e89773. https://doi.org/10.1371/journal.pone.0089773

Van Harreveld, F., van der Pligt, J., & de Liver, Y. N. (2009). The agony of ambivalence and ways to resolve it: introducing the MAID model. *Personality and Social Psychology Review*, *13*, 45– 61. http://dx.doi.org/10.1177/1088868308324518

Vermeulen, N., Bayot, M., Mermillod, M., & Grynberg, D. (2019). Alexithymia disrupts the beneficial influence of arousal on attention: Evidence from the attentional blink. *Personality Disorders: Theory, Research, and Treatment*, *10*(6), 545.

Võ, M. L., Jacobs, A. M., & Conrad, M. (2006). Cross-validating the Berlin affective word list. *Behavior Research Methods*, *38*(4), 606–609. https://doi.org/10.3758/BF03193892

Volkova, E., De La Rosa, S., Bülthoff, H. H., & Mohler, B. (2014). The MPI emotional body expressions database for narrative scenarios. *PLoS ONE*, *9*(12), 1–28. https://doi.org/10.1371/journal.pone.0113647

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron*, *30*(3), 829–841. https://doi.org/10.1016/s0896-6273(01)00328-2

Vuoskoski, J. K., & Eerola, T. (2011). The role of mood and personality in the perception of emotions represented by music. *Cortex*, *47*(9), 1099–1106. https://doi.org/10.1016/j.cortex.2011.04.011

Wachs, T. D. (2013). Relation of maternal personality to perceptions of environmental chaos in the home. *Journal of Environmental Psychology*, *34*, 1–9. https://doi.org/10.1016/j.jenvp.2012.11.003

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Weierich, M. R., Kleshchova, O., Rieder, J. K., Reilly, D. M., & Vazire, S. (2019). The complex affective scene set (COMPASS): Solving the social content problem in affective visual stimulus sets. *Collabra: Psychology*, *5*(1), 53. https://doi.org/10.1525/collabra.256

Westermann, R., Spies, K., Stahl, G., & Hesse, F.W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, *26*(4), 557–580. https://doi.org/10.1002/(sici)1099-0992(199607)26:4<557::aid-ejsp769>3.0.co;2-4

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 509–520. http://dx.doi.org/10.1037//0278-7393.28.3.509

Wierzba, M., Riegel, M., Wypych, M., Jednorwóg, K., Turnau, P., Grabowska, A., & Marchewka, A. (2015). Basic emotions in the nencki affective word list (NAWL be): New method of classifying emotional stimuli. *PLoS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0132305

Williams, J. M., Carr, M., & Blagrove, M. (2021). Sensory processing sensitivity: Associations with the detection of real degraded stimuli, and reporting of illusory stimuli and paranormal experiences. *Personality and Individual Differences*, *177*, 110807. https://doi.org/10.1016/j.paid.2021.110807

Williams, M. A., McGlone, F., Abbott, D. F., & Mattingley, J. B. (2005). Differential amygdala responses to happy and fearful facial expressions depend on selective attention. *Neuroimage*, *24*(2), 417–425. https://doi.org/10.1016/j.neuroimage.2004.08.017

Wingenbach, T. S. H., Ashwin, C., & Brosnan, M. (2016). Validation of the Amsterdam Dynamic Facial Expression Set ' Bath Intensity Variations (ADFES-BIV): A Set of Videos Expressing Low, Intermediate, and High Intensity Emotions. *PLoS ONE*, *11*(1), 1–28. https://doi.org/10.1371/journal.pone.0147112

Wolf, O. T. (2008). The influence of stress hormones on emotional memory: relevance for psychopathology. *Acta Psychologica*, *127*(3), 513–531. https://doi.org/10.1016/j.actpsy.2007.08.002

Wolf, O. T., Schommer, N. C., Hellhammer, D. H., McEwen, B. S., & Kirschbaum, C. (2001). The relationship between stress induced cortisol levels and memory differs between men and women. *Psychoneuroendocrinology*, *26*(7), 711–720. https://doi.org/10.1016/s0306-4530(01)00025-7

Wolff, J. S., & Wogalter, M. S. (1998). Comprehension of Pictorial Symbols: Effects of Context and Test Method. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(2), 173–186. https://doi.org/10.1518%2F001872098779480433

Wood, F., Taylor, B., Penny, R., & Stump, D. (1980). Regional cerebral blood flow response to recognition memory versus semantic classification tasks. *Brain and Language*, *9*(1), 113–122. https://doi.org/10.1016/0093-934X(80)90075-9

Wrase, J., Klein, S., Gruesser, S. M., Hermann, D., Flor, H., Mann, K., ... & Heinz, A. (2003). Gender differences in the processing of standardized emotional visual stimuli in humans: a functional magnetic resonance imaging study. *Neuroscience Letters*, *348*(1), 41–45. https://doi.org/10.1016/s0304-3940(03)00565-2

Wyllie, J., Carlson, J., & Rosenberger III, P. J. (2014). Examining the influence of different levels of sexual-stimuli intensity by gender on advertising effectiveness. *Journal of Marketing Management*, *30*(7-8), 697–718. https://doi.org/10.1080/0267257X.2013.838988

Xue, Y. L., Mao, X., & Zhang, F. (2006). Beihang university facial expression database and multiple facial expression recognition. *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, *2006*(August), 3282–3287. https://doi.org/10.1109/ICMLC.2006.258460

Yan, W. J., Wu, Q., Liu, Y. J., Wang, S. J., & Fu, X. (2013). CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013. https://doi.org/10.1109/FG.2013.6553799

Yingliang, M. A., Paterson, H. M., & Pollick, F. E. (2006). A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior Research Methods*, *38*(1), 134–141. https://doi.org/10.3758/bf03192758

Yoshie, M., & Sauter, D. A. (2019). Cultural Norms Influence Nonverbal Emotion Communication: Japanese Vocalizations of Socially Disengaging Emotions. *Emotion (Washington, D.C.)*, *20*(3), 513–517. https://doi.org/10.1037/emo0000580

Zafeiriou, S., Papaioannou, A., Kotsia, I., Nicolaou, M., & Zhao, G. (2016). Facial Affect "In-the-Wild": A Survey and a New Database. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1487–1498. https://doi.org/10.1109/CVPRW.2016.186

Zajonc, R. B. (1968). Attitudinal Effects of Mere Exposure. *Journal of Personality and Social Psychology*, *9*(2 PART 2), 1–27. https://doi.org/10.1037/h0025848

Zammuner, V. L. (2011). People's active emotion vocabulary: Free listing of emotion labels and their association to salient psychological variables. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6800 LNCS(December), 449–460. https://doi.org/10.1007/978-3-642-25775-9_40

Appendix

A. Introduction to the terms “valence” and “arousal” as well as encoding instructions

The part of the survey you complete today consist of two tasks: a rating of images and a completion of a short questionnaire about yourself.

In this first task you will be presented with different images displayed individually on your device screen.

Please rate each image using the scales below it.

There will be a short description of the scales and instructions on how to use them.

You will find that you get into a rhythm when rating the images, and we find that most people take just a few seconds to look at and rate each image.

There is no right or wrong answer. However, it is important that you select the answers according to **your very personal** perception of each image.

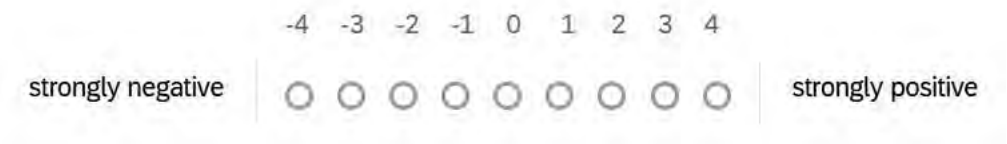
For each image displayed to you, please indicate which option best represents how you perceive the image by using the scales below it.

One rating will be for the valence (from -4 to +4), this means how **negative/unpleasant** or **positive/pleasant** you found an image to be.
The second rating will be for how **activating** or **arousing** you found the image to be (from 0 to 8). This refers to how you feel in your body, when viewing the image. If you feel increased physiological activation, this is what we mean. However, note, that this does not refer to sexual arousal.

Please pay attention:
For an image that you find **neither positive nor negative**, you would select **0** for **valence** (which is in the middle of the scale),
and
if you find the content to be **neither disturbing nor exciting** in any way, you would select **0** for **arousal** (which is on the far left of the scale).

Here you can see what the scales will look like:

How **negative/unpleasant or positive/pleasant** is the image to you?
Give a -4 to +4 rating:
(0 represents "not negative or positive")



How **exciting** is the image to you? Consider how much the image **grabs your attention** and **increases your physiological activation**.
Give a 0-8 rating:

(4 represents "medium arousing")

B. HSPS Questionnaire Items

Item 1	Are you easily overwhelmed by strong sensory input?
Item 2	Do you seem to be aware of subtleties in your environment?
Item 3	Do other people's moods affect you?
Item 4	Do you tend to be more sensitive to pain?
Item 5	Do you find yourself needing to withdraw during busy days, into bed or into a darkened room or any place where you can have some privacy and relief from stimulation?
Item 6	Are you particularly sensitive to the effects of caffeine?
Item 11	Does your nervous system sometimes feel so frazzled that you just have to go off by yourself?
Item 12	Are you conscientious?
Item 13	Do you startle easily?
Item 14	Do you get rattled when you have a lot to do in a short amount of time?
Item 15	When people are uncomfortable in a physical environment do you tend to know what needs to be done to make it more comfortable (like changing the lighting or the seating)?
Item 16	Are you annoyed when people try to get you to do too many things at once?
Item 17	Do you try hard to avoid making mistakes or forgetting things?
Item 18	Do you make a point to avoid violent movies and TV shows?
Item 19	Do you become unpleasantly aroused when a lot is going on around you?
Item 20	Does being very hungry create a strong reaction in you, disrupting your concentration or mood?
Item 21	Do changes in your life shake you up?
Item 22	Do you notice and enjoy delicate or fine scents, tastes, sounds, works of art?
Item 23	Do you find it unpleasant to have a lot going on at once?
Item 24	Do you make it a high priority to arrange your life to avoid upsetting or overwhelming situations?
Item 25	Are you bothered by intense stimuli, like loud noises or chaotic scenes?
Item 26	When you must compete or be observed while performing a task, do you become so nervous or shaky that you do much worse than you would otherwise?
Item 27	When you were a child, did parents or teachers seem to see you as sensitive or shy?