

THE PHOTOGRAPHIC AFFECT METER: A NOVEL APPLICATION TO MEASURE  
MOMENTARY EMOTIONAL STATES

A Dissertation

Presented to the Faculty of the Graduate School  
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy

by

John P. Pollak

January 2012

© 2012 John P. Pollak

# THE PHOTOGRAPHIC AFFECT METER: A NOVEL APPLICATION TO MEASURE MOMENTARY EMOTIONAL STATES

John P. Pollak, Ph.D.

Cornell University 2012

Emotion plays an ever-present role in human existence, impacting nearly every behavior and decision in some way. Research in the behavioral sciences is rife with exploration of emotion and the role it plays in everything from business decision making to health-related behavior. However, affect, the feeling or experience of emotion, is complex and presents many challenges to those interested in measuring it, often resulting in a disconnect between the way emotion is experienced and felt and the way that researchers measure it.

This dissertation addresses the challenges of measuring emotion through the presentation of the design, development, and validation of a novel measure of affect. The Photographic Affect Meter, PAM, is a digital measure of affect in which subjects choose from a grid of photos the one that best represents their current emotional state. The objective of PAM is to provide researchers with a means of measuring affect that is brief, reliable, and effective when used in situ. PAM was developed through an extensive iterative design process anchored in Human-Computer Interaction research, drawing inspiration from Affective Computing and Design literature. PAM was then rigorously validated via three separate studies. In the first two studies, subjects were assessed using both PAM and one of three widely accepted measure of affect (PANAS, Russell's Affect Grid, or the Self Assessment Manikin). In the third study, subjects were

induced with negative, neutral, or positive affect and then assessed using PAM. In each of the studies, PAM results were found to be consistent with expectation, establishing the validity and reliability of the measure. While the success of this approach has implications for researchers in Affective Computing, Emotion, Design, and Ecological Momentary Assessment, the primary contribution of this work is the introduction of a novel measure of affect that is ready to be deployed in a wide variety of studies.

## BIOGRAPHICAL SKETCH

John (JP) Pollak, was born in Woodland, California, and moved to Ithaca, New York shortly thereafter, where he grew up and eventually graduated from Ithaca High School as part of the class of 1995. Following high school, JP spent two years at the University of Pennsylvania studying molecular biology and computer science, a year in England playing soccer, and finally two years at Cornell University studying genetics and Human-Computer Interaction. In the years following graduation, JP worked as a software consultant in the life sciences, during which time he developed the software used by nearly every zoo in the world to manage genetic and demographic data, the gold standard software for population viability assessment in conservation, and countless other applications used around the world. This work culminated in a hiring as the Vice President of Product Development at Advanced Warning Systems, Inc., in Carlsbad, California, where JP was responsible for defining and managing the development of the company's entire health care software product line. JP returned to Ithaca and Cornell in 2007 and received a Master of Science degree in Communication shortly thereafter for his work on emotion sharing technologies. He has since continued his work with Dr Geri Gay exploring the use of mobile technologies in health care, with a particular focus on the assessment and sharing of emotion.

## ACKNOWLEDGMENTS

The work presented in this dissertation was only possible because of the support and contribution of faculty, friends, and family members. First and foremost, I am immensely grateful to my advisor, Geri Gay, for providing me the opportunity to conduct this work. She helped to convince me that returning to academia was a good move and, when I returned, she provided me with all of the resources I could possibly need to be successful. I believe she has created a truly unique environment in the Interaction Design Lab, one that fosters creativity and exploration above all else, yet somehow manages to remain focused and productive around strong central themes. Without this environment and Geri's guidance, the work in this dissertation and the fantastic experience I have had here might not have been possible. I am also indebted to my committee members Jeff Hancock and Sahara Byrne for helping guide me through this process. When I returned to graduate studies after years in industry, I may have underestimated the steepness of the learning curve ahead. Jeff and Sahara share much of the responsibility for teaching me what it means to produce scholarly work and ultimately be successful in academia. The work here is a testament to the insight they have provided me on everything from methodology to how to position my work. I am convinced that the members of my committee are three of the best people that I have ever worked with.

I must also thank the wonderful team of researchers that has played no small role in this work over the last few years. Phil Adams—as an undergraduate, Master's student, and ultimately PhD student—has been crucial to every facet of this work since day one and has likewise been a good friend. Andrew Ehrlich, Andy Liang, Vera Khovanskaya, and Zach Porges have each played an important role in data collection, development, and design and are perfect examples of what makes work in the IDL such

a pleasure; they are hard-working and incredibly bright, but also a pure joy to be around. I would also like to thank Amy Gonzales, Eric Baumer, and Dan Cosley whose involvement in the IDL and comments and critiques have certainly helped shape the ideas presented here.

From Weill Cornell Medical College, Doctors Andy Dannenberg, John Leonard, Mary Charlson, and Ellen Ritchie have provided advice, funding, patients, and opportunities that would otherwise be impossible to come by in an Information Science department. I am grateful for the opportunity to work with such talented and caring medical professionals and researchers, and I hope to continue these collaborations into the future.

I would also like to thank those that I consider my cohort in Information Science, Stephen Purpura and Hrönn Brynjarsdóttir. Colleagues and close friends, we've collaborated a little, but mostly, we've shared the ups and downs of graduate school, toasting each other's successes and finding ways to laugh about our failures. Along the way, we did more than our part to keep Bryan's Fine Foods and certain California wineries in business. Patti and Thor Purpura also deserve special thanks, for offering up their home as a testing ground for all manner of bizarre scientific exploration into food and drink that, amazingly, has not (yet) resulted in any significant destruction of property.

I also owe a special thank you to Daniela Retelny and Jamie Guillory—friends and two of the brightest and most driven people I know. They have each contributed to my work, but more importantly, along with Geri and myself, they make up the "Mindless Eating Club." While we have indeed enjoyed many, many fine meals together, I am most appreciative of the support, fun, and balance that we provide one another as we strive to overachieve. I am most hopeful that the club survives its new

geographic challenges and continues to find ways to collaborate and play.

Finally, I offer thanks to my family for supporting me throughout this process. Thank you to my parents for the endless support and to my sister Emily for the endless comic relief and inside jokes. Of course, I thank my wife Caitlin for the love and support and for working even harder than I do. This work truly would not have been possible without her. Lastly I thank my feline companions Shadow and Ellie, who have made the long hours writing at home bearable.

## TABLE OF CONTENTS

Biographical sketch.....	iii
Acknowledgments.....	iv
List of figures.....	viii
List of tables.....	ix
Chapter 1: Introduction.....	1
Chapter 2: Measurement of Affect.....	7
Chapter 3: Design and Development of the Photographic Affect Meter.....	24
Chapter 4: Validation of PAM with Three Widely Used Measures of Affect....	42
Chapter 5: PAM Validation Using Classic Mood Induction.....	65
Chapter 6: General discussion.....	82
Appendices.....	103
References.....	116

## LIST OF FIGURES

Figure 1.1: PAM.....	5
Figure 2.1: Russell’s Circumplex Model of Affect.....	8
Figure 3.1: Russell’s Affect Grid.....	26
Figure 3.2: Bradley and Lang’s Self Assessment Manikin.....	28
Figure 3.3: Images from Lang’s International Affective Picture System.....	29
Figure 3.4: Mapping emotion words into two-dimensional space using Russell’s Circumplex Model of Affect.....	35
Figure 3.5: A sample of PAM images displayed in affective space.....	36
Figure 3.6: PAM scoring.....	37
Figure 3.7: PAM, running on a Google Android-based mobile phone.....	39
Figure 4.1: Two-dimensional affective space as defined by valence and arousal...	47
Figure 4.2: PAM scoring.....	51
Figure 5.1: Plot of Mean Valence by condition.....	74
Figure 5.2: Plot of Mean Arousal by condition.....	75
Figure 5.3: Plot of Mean Positive Affect by condition.....	76
Figure 5.4: Plot of Mean Negative Affect by condition.....	77

## LIST OF TABLES

Table 2.1: Selection of scales and systems that measure affect.....	22
Table 4.1: Descriptive statistics for responses to PAM, Affect Grid, and SAM.....	53
Table 4.2: Correlations between affect scores from PAM, Affect Grid, and SAM....	54
Table 4.3: Descriptive statistics for responses to PAM and PANAS.....	55
Table 4.4: Linear regression for PANAS PA on PAM Valence and Arousal.....	56
Table 4.5: Linear regression for PANAS NA on PAM Valence and Arousal.....	57
Table 4.6: Correlations between affective scores from PANAS and PAM.....	58
Table 5.1: Means of each PAM subscale for each video manipulation.....	74

# CHAPTER 1

## INTRODUCTION

Emotion is an omnipresent aspect of human existence that influences nearly every facet of our behavior (R. S. Lazarus & B. N. Lazarus, 1994; Pressman & Cohen, 2005; Slovic, Finucane, & Peters, 2007). As such, scientists often require at least a basic understanding of the emotional state of research subjects. The measurement and study of emotion is commonplace in the behavioral sciences, (Baumeister, Vohs, DeWall, & Zhang, 2007; Bradley & P. J. Lang, 1994; Gross, Richards, & John, 2006; P. J. Lang, 1995; Morris et al., 2010) and is particularly prevalent in health-related research, where it has become clear that numerous relationships exist between emotion and health (Pressman & Cohen, 2005). Given this, there is obviously great need for instruments and methods that reliably assess emotion. While there is no shortage of such instruments, emotion is highly complex and difficult to measure; people rarely understand the whole of their current emotional state and scholars disagree about the very definition of the concept. To this end, rather than assessing the whole of a person's emotional state, researchers typically assess affect, the conscious feeling or experiencing of emotion on the part of an individual.

The measurement of affect is not without challenge. The first challenge is that affect is that it is often considered to exist in two dimensions: valence (unpleasurable to pleasurable or negative to positive) and arousal (low activation to high activation). In this framework, best described by Russell's Circumplex Model of Affect (Russell, 1980), various emotions can be mapped in the two-dimensional affective space defined by valence and arousal. Further, an individual's current affective state can be mapped into these two dimensions, and this is the manner in which many existing measures of affect are reported. In other words, to meaningfully measure and report affect, an instrument

must capture to some extent an individual's state in regards to both valence and arousal dimensions.

A second challenge to the measurement of affect is the dual nature of the emotional experience. On one hand, there is the underlying, generally stable affective condition of the individual. This is also known as trait affect, or dispositional affect, and typically considers how one feels irrespective of day-to-day events, emphasizing longer-term feelings. On the other hand, state affect, or mood, focuses on short-term bout of emotion being experienced in any moment, possibly in response to day-to-day events (George, 1996).

Taking all of this into consideration, most accepted measures of affect are examining emotion in at least one of the two dimensions (valence or arousal) and are either reporting on an individual's state or trait affect. For example, PANAS, the Positive And Negative Affect Schedule, one of the most widely used measures of affect, particularly in the health arena, purposely conflates the valence and arousal dimensions and simply reports scores of how much positive and negative affect an individual is experiencing. By varying the instructions provided to the user (i.e. how you have felt in the last month vs. how you have felt today), PANAS can be used more or less successfully to assess either state or trait affect, although the extent to which one may confound the other is not entirely understood (Watson, Clark, & Tellegen, 1988).

While the assessment of affect (and most other psychological concepts) has traditionally been carried out with pen and paper or more recently computers, the mobile phone presents a new avenue for collecting such data. Ecological Momentary Assessment, or EMA, is a class of data collection methods that leverage the mobile phone's pervasiveness in order to assess subjects in context at opportune moments (L. Barrett & D. Barrett, 2001). Because EMA methods assess subjects in context at the time

of interest, they can possess a significant advantage over standard data collection methods that rely on post-study, recall-based assessment. In particular, the use of EMA permits more sensitive, detailed, and wide-ranging measurements of mood and behavior (Moskowitz & Young, 2006).

Shifting the role of assessment to the mobile phone—which has evolved to be small but powerful computing devices—affords further opportunity beyond those typically discussed around EMA. The field of affective computing provides a framework for examining how computing devices can be used to assess, represent, and communicate affect. While computing devices themselves are inherently unemotional and poor at assessing emotion, we can utilize what we know about how humans and computers interact as well as the functions that computers are particularly good at to perform reliable, contextually-anchored assessments of affect (Picard, 1997).

Despite the substantial amount of work examining the measurement of affect, there remains a disconnect between the way affect is measured and the reality of how we experience it. The assessment of subjects at the beginning and end of a study by self-report (for affect or otherwise) is a standard practice with a long history in experimental and clinical investigations. While this makes sense from a practicality standpoint, there are serious drawbacks to employing this approach when dealing with affect that clearly threaten the validity of findings (Mancuso & Charlson, 1995). First, numerous studies have found reports of specific past events to be distorted (Stone & Shiffman, 2002), generally due to poor memory surrounding a given event or the result of a subject's state at the time of interview (Salovey, Sieber, & Jobe, 1994). Second, the dual nature of emotion as described above makes it extremely difficult to understand what is actually being measured. When affect is measured, what is captured is either the more stable, general affective condition of a subject (trait affect, or dispositional

affect) or a short-term bout of emotion being experienced in that given moment (state affect, or possibly mood) (George, 1996). Hence, when affect is measured at the end of a study, what is generally captured is trait affect, possibly influenced by current or intensely experienced prior state affect, or perhaps just state affect at the time of closeout (Kahneman, 2003). Either way, the momentary emotional experience of the subject during the actual study period is difficult to capture with this single assessment.

This problem does not exist due to a lack of awareness on the part of the research community. Rather, there is a problem with most measures of affect: they are lengthy questionnaire-based methods that are generally only administered at the beginning and end of studies. Based on the literature, what is needed to solve this problem is a tool that (1) reliably measures affect, (2) is unobtrusive and pleasant enough to administer at least daily, and (3) can be administered in situ. This dissertation addresses the need for better measurement of affect by presenting the design, development, validation, and applicability of a computer and mobile-phone based measure of affect called the Photographic Affect Meter, or PAM (See Figure 1.1).

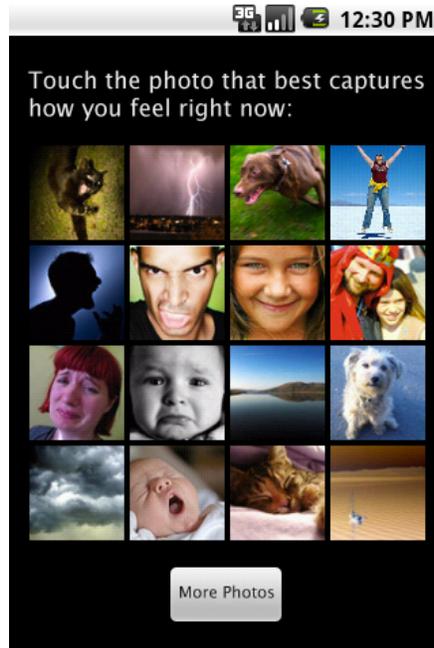


Figure 1.1, PAM, on a Google Android-based mobile phone.

For this work, a decidedly different approach has been taken toward the measurement and assessment of affect. Rather than approaching the problem strictly from the perspective of emotion and measurement, this work uses an iterative, user-centered design process borrowed directly from the Human-Computer Interaction (HCI) literature (Gay & Hembrooke, 2004; Nielsen, 1993). In this approach, appropriate theory (i.e. from work in emotion, affective computing, and EMA) and practical experience are used to frame the design process, but ultimately the most important decisions are made along the way through direct input from potential users.

In PAM, users are asked to select a photo (from a grid of 16 choices) that best represents how they currently feel. The design of PAM takes cues from literature that suggests that photographs and emotion (Chalfen, 1987; J. Mayer, DiPaolo, & Salovey, 1990) are linked, often very explicitly with universally shared meaning (P. J. Lang, 1995). Further research from affective computing suggests that there is an element of

ambiguity, or interpretive flexibility associated with imagery that allows users to feel more engaged with the system and their choices (Mateas, 2001; Sengers, Boehner, Mateas, & Gay, 2008). The selection of photos in PAM was derived through a social process designed to increase the likelihood of finding a range of socially shared meanings in the imagery. The set of images, and ultimately the format of the scale itself was iteratively tested and improved with increasing numbers of subjects. Finally, a rigorous validation process similar to those employed by other accepted measures of affect (e.g. (Bradley & P. J. Lang, 1994; Russell, Anna Weiss, & Mendelsohn, 1989; Watson et al., 1988)) was employed, finding that PAM does indeed satisfy the three objectives laid out above.

This dissertation is structured as follows: Chapter 2 presents an overview of theory and practice regarding the measurement of affect and Ecological Momentary Assessment. Chapter 3 presents the iterative, evidence-based design and development of PAM, anchoring it to classic literature from emotion measurement as well as newer literature on affective computing. Chapter 4 presents validation of PAM in two experiments in which subjects are assessed with both PAM and one of three other widely accepted measures of affect. Chapter 5 presents further validation of PAM through a take on a classic mood induction experiment. Finally, Chapter 6 is a general discussion of these findings, applicability and use of PAM, the implications on theory and practice, and suggestions for future work.

## CHAPTER 2

### MEASUREMENT OF AFFECT

The work presented in this dissertation lies at the intersection of research on measurement of emotion, Ecological Momentary Assessment, and affective computing. Each field independently provides a wealth of knowledge that is relevant to this work; however, combining insights from each does yield an opportunity for advancement. This chapter details the most influential work in each of these areas and identifies an important gap in the literature.

#### *Introduction to the Measurement of Affect*

Inherent to all work examining emotion is the assumption that tools are available to accurately and reliably assess it. Because of the complexity of emotion as a concept, researchers typically measure the extent to which a subject feels or experiences emotion—either underlying or in response to stimuli, a concept known as *affect*. However, there are two aspects to affect that make it particularly challenging to measure and quantify. First, it is generally believed that in order to quantify affect, two to three separate dimensions of measurement are required. Second, there is a duality to the experience of affect over time that does not always clearly delineate.

#### *The Dimensions of Affect*

In measuring or describing affect, researchers commonly use three dimensions first proposed by Wundt over 100 years ago and later empirically supported by Mehrabian and Russell (Mehrabian & Russell, 1974) : *valence* (unpleasurable to pleasurable or negative to positive), *arousal* (low activation to high activation), and *dominance* (submissive to dominant) (Wundt, 1904). The belief, then, is that in order to

adequately describe an individual's affective state, the researcher must in some way capture not one but three different measures. Because of the challenge involved in a three-dimensional model—both in the collection and analysis of data—most current emotion scholars work within a more manageable two-dimensional framework that omits the dominance dimension (Scherer, 2005).

The classic two-dimensional valence-arousal interpretation of affect is perhaps best described by Russell's Circumplex Model of Affect (Russell, 1980). In this model, each of the individual emotions that one might experience has been mapped into two-dimensional, valence-arousal emotional space using repeated testing with semantic differential and various statistical analyses (See Figure 2.1). According to the model, an individual can experience valence and arousal independently, and the specific emotions that are experienced lie at the appropriate intersection. For example, the emotion *excited* is high in arousal and very positive in valence, while the emotion *bored* is low in arousal and negative in valence. Theoretically, an individual's current emotional state can be mapped into this space as well, even though it may encompass a variety of these individual emotions.

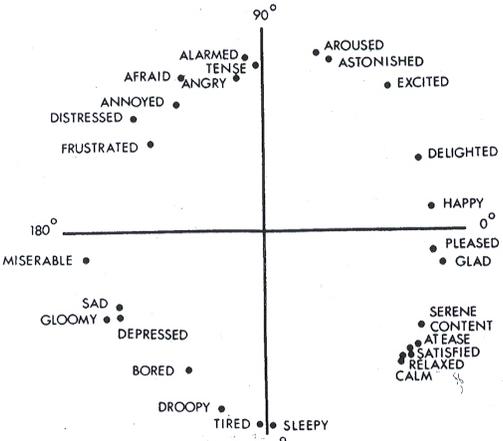


Figure 2.1. Russell's (1980) Circumplex Model of Affect, where the y-axis represents arousal level of the emotion (high arousal at the top, low arousal at the bottom) and the x-axis represents valence (negative valence to the left, positive to the right).

### *State and Trait Affect*

Another widely held notion complicates the measurement of affect: There is a dual nature to the manner in which individuals experience affect over time. On the one hand, there is *trait* affect, the more stable, general emotional disposition of an individual. Trait affect is slow to change and is generally considered to be the result of factors like upbringing, genetics, or exposure to repeated or extreme stimuli. On the other hand, there is *state* affect, the shorter term, often more extreme bouts of emotion experienced by an individual. Unlike trait affect, state affect is generally considered to be the direct response to environmental or social stimuli (George, 1996).

This duality presents two obvious challenges to measurement. First, how does one temporally operationalize state versus trait affect for practical purposes during a research study? A common guideline is that describing affect over the period of a week or longer is a report of trait affect, while anything less is a measure of state affect (Pressman & Cohen, 2005). Second, researchers must be cognizant of the interplay of state and trait emotion and take great care in understanding what it is that they are actually measuring. When emotion is measured at long intervals such as once a week or more, such as at the end of a study period, it is difficult to know for certain what has been captured. Kahneman argues that when subjects recall happiness, a Peak-End model governs their response. In this model, individuals tend to report on their happiness over a period of time based primarily on the most intense moments of happiness in combination with their happiness at the time of reporting (Kahneman, 2003).

### *Measurement Strategies*

Researchers have overcome these difficulties in a number of ways, developing and adapting a variety of strategies—each with their own strengths, weaknesses, and appropriate use cases. These strategies can be broadly classified in a number of ways. First, affect can be measured either by self-report or observation, alternatively labeled as active versus passive involvement. With self-report, or active involvement, the subject is expected to report in some way, generally through a scale administered with pen and paper or computing device, how they are feeling or have felt over some period of time. Self-report is widely used in the social, behavioral, and medical sciences, although it is not without issues, e.g. (Mancuso & Charlson, 1995; Salovey et al., 1994). Observation, on the other hand generally requires little or no involvement from the subject. This class of methods relies on tracking either the behavior or physiology of subject at a given moment and applying algorithms to infer affect. A sampling of these methods will be discussed, however for reasons that will be spelled out later, observational strategies are not the focal point of this dissertation.

Within the self-report classification, it is important to consider the depth, length, or brevity of a method. On one end of the spectrum are deep, introspective methods such as semantic differential (Mehrabian & Russell, 1974) that include many items and require a substantial time commitment to assess the affective experience of the subject. While these strategies require a great deal of time and effort on the part of both the subject and the researcher, they paint a more complete picture of affect. On the other end of the spectrum are brief, abbreviated, and single-item measures of affect. These methods necessarily give up the depth of understanding that more complex measures provide in exchange for greater manageability and the potential for more frequent assessment. Research has weighed in discussing the merits of this spectrum, e.g. (Burisch, 1984; Kahneman, 2003), and like most such debates, individual circumstances

and needs likely outweigh generalizations.

Much diversity exists within these classifications. Strategies differ in terms of method, context, operationalization of affect, number of dimensions considered, timeliness, and even whether or not they produce quantitative output. The next sections of this chapter will detail the most important of these strategies and methods and discuss key strengths and weaknesses. This survey of key strategies will also serve to document much of the progress made toward more effective measurement of affect.

### *Traditional Measures of Affect*

Traditionally, affect has been measured using self-report, primarily with pen and paper and more recently computer adaptations of pen and paper scales. There is no shortage of validated measures in this space, each with their own strengths, weaknesses, and appropriate usage scenarios. The semantic differential work conducted by Mehrabian and Russell (Mehrabian & Russell, 1974) could be considered the among the first forays into what would be recognizable today as tools for measuring affect and effectively solidified the Pleasure-Arousal-Dominance framework for affect. This work, and the resulting scale, involves asking subjects to identify where their current feelings or feelings about a stimulus fall on 18 different nine-point differentials between bipolar adjective pairs (e.g. unhappy-happy for valence, calm-excited for arousal, or influenced-influential for dominance). Measuring affect with Mehrabian and Russell's semantic differential scale yields not only measures of valence, arousal, and dominance, but also point measures for each of the affect dyads that comprise the three primary dimensions. The clear value of this method is the complex and explicit picture it provides.

McNair and Lorr's (McNair & Lorr, 1992) Profile of Moods Scale, or POMS, is a similarly extensive scale that uses semantic differential to measure emotion on six factors (tension-anxiety, depression, anger-hostility, vigor-activity, fatigue, confusion-bewilderment). Unlike nearly every other measure discussed in this review, POMS does not fit neatly within the valence-arousal-dominance view of affect and, in fact, has only one dimension that remotely addresses positively valenced emotions. However, the value in POMS lies in the specificity of each of these factors and the fact that many of them are appropriate for clinical settings; for example, one variant of POMS is frequently used as a gold standard measure for assessing the emotional state of newly diagnosed cancer patients (Cella, 1987).

Watson, et al (Watson et al., 1988) presented a different approach with their Positive And Negative Affect Schedule, or PANAS, essentially combining dimensions of arousal and valence into two measures, one for positive affect (PA) and one for negative affect (NA). PANAS consists of 20 items—single emotion or feeling words that represent positively and negatively valenced feelings as well as arousal/ activation. For the PA scale, higher arousal and more pleasurable selections result in higher scores; low arousal and less pleasure result in a low score. The NA scale functions the same with respects to arousal but features negatively valenced items. More recently, PANAS has been suggested to be more reflective of positive activation than of pleasure, as items such as happiness are not directly assessed (Crawford & Henry, 2004; Watson, Wiese, Vaidya, & Tellegen, 1999), which could lead to misleading interpretation if PA is assumed to be pleasure-driven. Still, presumably because of some combination of the relative simplicity and repeated validation, scholars in numerous fields of study use PANAS extensively.

Semantic differential and POMS might be considered among the most thorough

measures of affect. Both are extremely thorough and provide explicit sub-measures for various specific emotions. However, many subsequent and prominent researchers, including Bradley and Lang (Bradley & P. J. Lang, 1994) and Russell (Russell et al., 1989) have determined that semantic differential is usually too unwieldy both in terms of experimental procedure and eventual analysis of collected data. Likewise PANAS, although it produces simplified measures of positive and negative affect, is still a 20-item scale that requires more than trivial time and effort on the part of the subject. This next group of methods has employed experimentation and factor analyses in attempt to distill these complex measures into more concise and wieldy tools for the researcher.

#### *A Move Toward More Brief Measures of Affect*

While the measures presented thus far provide researchers with in-depth report on emotional state, they are lengthy and time consuming to complete. In most circumstances this is a welcome trade-off, but there are times when methods dictate a more brief measure. Not surprisingly, researchers have sought to streamline the process of measuring affect for some time. In some cases, this has meant adapting or parsing existing scales; for example a heavily parsed version of POMS has been successful in assessing emotional disturbance in newly diagnosed cancer patients (Cella, 1987). A second approach is to develop new measures, which can of course be a more time consuming and arduous task, but may ultimately be more interesting.

One notable brief measure of affect is Russell's Affect Grid (Russell et al, 1989). The Affective Grid is a pen and paper based instrument in which subjects are presented a 9x9 empty grid representing two-dimensional affect space with valence in the x-axis and arousal in the y-axis. Subjects place an X in the location in the grid that represents how they currently feel, and that location can be mapped to a score that correlates

strongly with the Semantic Differential Scale and PANAS Positive. The most significant issue with the Affect Grid is the difficulty that it presents subjects; the instructions are lengthy in order to explain the less-than-clear concepts described above, and even then, subjects must be able to cognitively process their current emotional state and quantify it on a box in a grid.

A second important single-item measure of affect is Bradley and Lang's Self Assessment Manikin, or SAM (Bradley & P. J. Lang, 1994). SAM presents users three sets of five drawings of a simple character, each set representing the range of states in the pleasure, arousal, and dominance dimensions. Subjects identify the character in each set that represents their current state in that particular dimension. SAM scores correlate highly with scores from Semantic Differential, particularly in the Valence and Arousal dimensions but were not reported as validated against PANAS. SAM attempts to simplify the cognitively difficult task of locating ones affective state in two-dimensional space by instead leveraging the human response to imagery.

To this point, this chapter has detailed the most oft used and influential of the traditional measures of affect. From here forward, the discussion will shift toward two relatively new strategies for assessing affect that take advantage of increasing computing power and ubiquity of mobile phones.

### *Ecological Momentary Assessment*

The assessment of subjects at the beginning and end of a study by self-report is a standard practice with a long history in experimental and clinical investigations. This makes sense logistically for a number of reasons including convenience, practicality, managing participant burden, and of course, lack of suitable methods for more frequent sampling. There are, however, two serious drawbacks to employing this approach

when dealing with affect. First, numerous studies have found reports of specific past events to be distorted (Stone & Shiffman, 2002), generally due to poor memory surrounding a given event or the result of a subject's state at the time of interview (Salovey et al., 1994). Such distortion of memory clearly threatens the effectiveness of research (Mancuso & Charlson, 1995).

There are two possible approaches to this problem: take the utmost care when selecting methods and measurement tools to ensure that trait affect is assessed properly (e.g. PANAS has been carefully validated for different time scopes ranging from state to trait (Watson et al., 1988; 1999)) or measure affect more frequently during course of the study. Kahneman (Kahneman, 2003) suggests this second approach—collecting and assembling a large quantity of repeated point measures—can provide a much clearer picture of emotional experience.

There does exist a theoretical framework dealing with the importance of collecting various in-the-moment assessments during research. Ecological Momentary Assessment, or EMA, is a class of data collection methods that attempt to alleviate the problem of recall bias and big picture emphasis by instead assessing subjects in context at opportune moments. Stone and Shiffman (Stone & Shiffman, 2002) provide a roadmap for researchers based on their own research for assessing subjects in this manner. Their work discusses potential tools for the researchers to maximize capture and retention rates, such as wisely selecting sampling methods such as prompt-based reporting or incident-based reporting. In the former, subjects receive prompts at certain times during the day asking them to complete an assessment, and in the latter, subjects are asked to report in when certain events or situations occur in their lives.

Reis and Gable (Reis & Gable, 2000) highlight the importance and benefits of sampling a subject's experience day-to-day using EMA. Salience, recency, and

memorability of events each impact recall and bias and can distort events when reported only at the end of a study. Moskowitz and Young (Moskowitz & Young, 2006) report on the benefits of EMA to their work in clinical pharmacology that is inextricably linked with mood and emotion. They argue that EMA has a significant advantage over standard data collection methods in that it permits more sensitive, detailed, and wide-ranging measurements of mood and behavior. Because of the importance of both time sensitivity and context in their work, they dub EMA the “method of the future” in their field. Along similar lines, Epstein et al (Epstein et al., 2009) have demonstrated success with this approach measuring and modeling affective states at the time of cravings and relapses in substance abusers.

Currently, Ecological Momentary Assessment methods are often conducted on mobile phones. For example, Experience Sampling (L. Barrett & D. Barrett, 2001; Csikszentmihalyi & Hunter, 2003; Larson & Csikszentmihalyi, 1983) is a notable subset of EMA methodology in which users are prompted by messages on their phone to complete some sort of assessment. The popularity of Experience Sampling as a method has even led to the development of an open mobile phone-based platform called MyExperience that researchers can use to administer in situ questioning, collect data off the subject’s phone, and even acquire sensor data (Froehlich, Chen, & Consolvo, 2007). Morris et al (Morris et al., 2010), for example, have used MyExperience with success to track emotion as part of an emotional awareness/ self-regulation intervention.

### *Affective Computing*

In moving to computer- and mobile phone-based assessments of emotion, we can look to the field of affective computing for guidance (Picard, 1997). While a subset of the affective computing literature has focused on the assessment of emotion, the

emphasis has largely been placed on automatic, or passive measurement of emotion. Work along these lines focuses on the role of the computer in deducing a user's affective state and recreating it in some electronic form that can then be decoded by other users. The underlying assumption, then, is that the current emotional state of a user is something that can be inferred from behavior or physiology, or that emotion can be depicted electronically in such a way that users can assess the information presented to them and accurately infer affect.

One avenue of emotional assessment has been to seek out evidence of emotion in an individual's behavior. A particularly successful method along these lines is Pennebaker's (Pennebaker, Zech, & Rime, 2001) Linguistic Inquiry and Word Count (LIWC) software, which analyzes text produced by a subject and reports on the presence, positivity, and negativity of emotion present, among other things. This system has been successful and widely used, but is limited to use where large bodies of text are available for analysis. For example, mining of blog entries on the service [livejournal.com](http://livejournal.com) for the time period surrounding the catastrophe on September 11, 2001, revealed a significant increase in negative emotions present in the writings of over one thousand bloggers (M. A. Cohn, Mehl, & Pennebaker, 2004). We Feel Fine searches new blog posts from a number of services for the phrases that begin with "I feel" and "I am feeling" to take the pulse of the mood of the bloggers of the world (Harris & Kamvar, 2008).

A second avenue of affective measurement is to use changes in an individual's physiology to infer affect. Previous research has linked many aspects of an individual's physiology to various emotional responses. Heart rate, body temperature, galvanic skin response and facial expression have all been examined and found to correspond with specific emotional responses to a variety of stimuli (Buck, Savin, Miller, & Caul, 1972).

In the era of ubiquitous sensing and computing devices, measuring and recording such physiological changes is rapidly becoming reality. For example, researchers working in affective presence have used each of these physiological cues collected from individual sensors to infer and represent affect to others (Nasoz, Alvarez, Lisetti, & Finkelstein, 2004). Facial recognition is clearly a more complex problem, but computer vision has made strides toward making this a reality (Kanade, J. F. Cohn, & Yingli Tian, 2000). There are, unfortunately, several disadvantages to the physiological approach, both technical and contextual. Chief among the technical issues is that generally ideal conditions have to be met for the tools to work at all, such as good lighting, steady cameras, moderate temperatures and humidity, and so on. Contextually speaking, while combining such technologies with mobile phones could provide a great deal of contextual information (Fogg & Eckles, 2008), there is still much ambiguity in what emotion should be represented by a collection of physiological measures alone.

The third avenue of affective measurement is simply use computing devices as a means to collect self-report data. Of course one could simply appropriate existing scales to the computer (and, to be fair, SAM was even originally computer-based (Bradley & P. J. Lang, 1994)). Alternatively, researchers have opted to develop new measures that take greater advantage of the technical capabilities of computers or mobile phones. Morris, et al (Morris et al., 2010), for example, assess affect on mobile phones with a colored two-dimensional Mood Map. The system is essentially an abstract version of the original Affect Grid that has been designed to run on the MyExperience experience sampling application (Froehlich et al., 2007). The simplistic but seemingly effective Mood Map is perfectly suited for such applications, however it has not been validated against reliable measures of affect, and given the level of abstraction of the grid space and arbitrary color use, it may prove difficult to validate.

Other examples of mobile phone-based assessments include Isomursu et al's (Isomursu, Tähti, Väinämö, & Kuutti, 2007) Feedback , although this is more focused at emotional response to a mobile app for use in usability testing and Meschtscherjakov, Weiss, and Scherndl's (Meschtscherjakov, A Weiss, & Scherndl, 2009) emoticon-based work, which is promising, but also not yet validated.

*A counter argument*

Running counter to both of these avenues of research, an emerging position in affective computing challenges the assumption that a computer can easily quantify emotion. In addition to the complexity of emotion discussed to this point, this position further asserts that emotion may rather be an ongoing social interaction between multiple individuals or an individual and the system rather than as a discrete state of an individual that can be somehow decoded by a computer. The traditional approach fails to allow for the fact that first, an individual might only be able to properly formulate their emotions through interaction with another, and second that only over time through ongoing interaction with others can shared meaning for various representations of emotion be constructed (Boehner, DePaula, Dourish, & Sengers, 2005).

In line with this counter-argument, a final avenue of research has focused on capturing individuals' emotional states without necessarily quantifying or analyzing them. For example, MoodJam is a widget that allows users to share emotion on the Web with multicolor representations. Affecter is an experimental webcam and wall-mounted digital display system that streams video with arbitrary and abstract distortion of some kind, leaving the affective interpretation entirely to the user (Sengers et al., 2008). In the mobile space, there are fewer examples yet. One such example of

particular interest is eMoto, a text messaging system in which users shake and squeeze a stylus to generate colors and shapes that make up the background image for the messages they send (Sundström, Ståhl, & Höök, 2007). While these works are not directly comparable to the work in the *measurement* of affect, they may yet provide guidance in designing affective computing systems.

Certainly, the crux of this argument suggests that computers should not be able to infer affect quite so easily as with a simple scale on a mobile phone. A point worth noting is that this argument may appear on the surface to be based on the premise that computers are incapable of quantifying affect, but in reality the core of the argument is that emotion can not and should not be quantified *at all* (Boehner et al., 2005). This is a difficult premise to argue for or against unilaterally, and upon closer examination, even Boehner concedes that in many scenarios and for many research agendas, the quantitative measurement of affect does have a place (Boehner, DePaula, Dourish, & Sengers, 2007):

Although our approach resonates with [the subjective] view, we argue that the way forward for affective computing and affective evaluation is not a debunking of objective approaches in total but a recognition of the limits and liabilities of both objective and subjective accounts of emotion (Boehner et al, 2007, p289).

Where those limitations and liabilities may exist is certainly up for debate, but the fact remains that for a substantial portion of researchers in the behavioral, social, and medical sciences, quantitative measurement of affect is the norm.

### *Summary and Positioning*

To this point, this section has detailed a number of existing measures of affect. Examining this list on the whole (See Table 2.1), particularly in the context of the challenges of measuring affect in experimental settings, one can see a clear gap. On one end of the spectrum, there are a number of rigorously validated, often lengthy, quantitative measures of affect that have been widely accepted and utilized. On the other, novel computer- or mobile phone-based systems measure affect in various ways either eschewing quantitative measurement altogether or not bothering to validate such measurement against known standards. Working towards the center of this spectrum, a gap emerges: there does not appear to be a validated measure of affect that can be used repeatedly, reliably, efficiently, and for methods such as Ecological Momentary Assessment. Observational methods (i.e., not self-report, including physiological measures, LIWC, etc.) show promise, but currently require too much of either the technology, the environment, or the user to be practical and, as such, will be disregarded from the remainder of this discussion.

Measure/ System	Output	Affective Framework	Length/ Subject legibility/ Complexity	Validation
Semantic Differential	Scores for Valence, Arousal, Dominance. Scores for each adjective pair.	Valence- Arousal- Dominance model	18 items Intuitive Complex results	Extensively validated using multiple methodologies spanning numerous studies
POMS	Scores for 6 dimensions	Custom	30+ items Intuitive Complex results	Extensively validated using multiple methodologies spanning numerous studies
PANAS	Scores for Positive Affect and Negative Affect Scores for each item are seldom used	Positive – Negative Affect model, intentionally conflates Valence- Arousal model	20 items Intuitive Simple results	Extensively validated using multiple methodologies spanning numerous studies
SAM	Scores for Valence, Arousal, Dominance. Scores for each adjective pair.	Valence- Arousal- Dominance model	3 items Abstract Simple results	Adequately validated
Affect Grid	Scores for Valence and Arousal	Valence- Arousal model	Single item, lengthy instructions Confusing at first Simple results	Adequately validated
LIWC	Scores for presence of emotion, positive or negative emotion, specific emotion words	Numerous dimensions, including Valence	No subject involvement Complex results	Adequately validated
Various Physiological Measures	Widely varied	Often Valence- Arousal, but varied	Minimal subject involvement Complex results	Varied
Mood Map	Scores for Valence and Arousal	Valence- Arousal model	Single item Abstract Simple results	None

Table 2.1, A selection of scales and systems employing the measurement of affect.

What then is needed to address this gap? Based on the above literature, there is a

relatively clear set of criteria. First and foremost, any measure must be proven as a validated and reliable measure of affect. To date rigorous validation of new measures has been commonplace in pen and paper measures as described above, but not so in the affective computing space. While such yet-to-be-validated measures may have other value, they are essentially unusable for researchers looking to capture usable, quantitative affective data. Without such validation and reliability, there is little point in continuing.

The second criterion for a new measure that seeks to fill this gap is that it must be unobtrusive enough that it can be administered at least daily for extended periods of time. As detailed above, many of the existing validated measures are lengthy, unwieldy, or confusing. These are not the attributes of an unobtrusive measure that can be administered with great frequency. Either through good design, brevity, or both, the challenge will be to achieve both validity and this unobtrusiveness that, it would appear, has not yet been realized with previous measures.

Finally, the measure must be able to be administered in situ. In other words, to be truly useful in Ecological Momentary Assessment or other repeated sampling regimes, the new measure must be capable of running on mobile platforms such that it can be administered to subjects at the most opportune moments (Fogg & Eckles, 2008). Ideally, such a measure would be part of a system designed to track, log, transmit, and document responses along with other contextual data, a la (Froehlich et al., 2007) to be of the utmost value to the researcher.

## CHAPTER 3

### DESIGN AND DEVELOPMENT OF THE PHOTOGRAPHIC AFFECT METER

The review of the literature presented in the previous chapter identified three features as crucial for a self-report measure of affect that would be suitable for frequent and contextually anchored use in research studies; the measure must be (1) valid and reliable, (2) brief and unobtrusive, and (3) able to be administered in situ. This chapter details the iterative process through which the Photographic Affect Meter, PAM, was designed and developed in accordance with these three goals. It should become clear that that PAM is as much a computing system or application as much as measure and is treated in that way throughout design and development. Further, a discussion of key design decisions begins to position this work as a novel combination of viewpoints from the quantitative, objective approach to emotion from the classic emotion and psychology literature and the more qualitative, subjective approach from the recent affective computing literature.

#### *Design Inspiration from the Literature*

The previous chapter has reviewed a substantial literature that can be used as a guide for the development of a new measure of affect. Examining this literature through the lens of Ecological Momentary Assessment has dictated the three criteria spelled out above that further refine the task at hand. Pursuant to the task and these three criteria, the literature regarding the measurement of emotion provides methodological fundamentals for measuring affect and previous efforts to create more brief measures of affect serve as a starting point for a new measure. However, the need for brevity as well as the need for the scale to be administered via computing devices demands attention be paid to the Human-Computer Interaction literature, particularly

that pertaining to affective computing. While much of the recent literature in affective computing warns against relying on computers to measure affect, in the end, that same literature can provide significant guidance in the creation of a new measure.

### *Emotion Measurement*

The emotion measurement literature provides background for conceiving, developing, and ultimately evaluating measures of affect. The more widely used measures of affect, including the Positive and Negative Affect Schedule, PANAS (Watson et al., 1988) or the Profile of Moods Scale, POMS (McNair & Lorr, 1992) are built on traditional measurement methods such as semantic differential and Likert-type items. The primary advantage of an approach such as semantic differential is that a large number of polar dimensions can be assessed simultaneously, providing clarity and a wealth of data for each assessment. These data can then be clustered around key dimensions or examined on the whole (McNair & Lorr, 1992; Mehrabian & Russell, 1974). PANAS, for example, reports on two measures—Positive Affect and Negative Affect (Watson et al., 1988). Another advantage of this approach is that through a combination of factor analysis and careful selection of items, sub-scales focusing on a single dimension of affect or more directly pertaining to a certain population can be readily created, e.g. POMS Depression subscales (Malouff & Schutte, 1985) or POMS for distress in cancer patients (Cella, 1987).

The primary drawback to semantic differential and Likert-type scales of affect, particularly in the context of the work presented in this dissertation, is their length. While abbreviated measures do exist, it is intuitively obvious that no single-item Likert scale or semantic dyad can capture the range of human emotion. As discussed in the previous chapter, this challenge has led to alternate approaches that might prove

instructive.

One such approach is that taken by Affect Grid, Russell's attempt at a single-item measure of affect. This approach is markedly different from the previous approaches in that relies on the subject to come to an understanding of the two-dimensional model of affect and be able to assess their own position within those dimensions. First, the subject reads through two pages of instruction plus examples illustrating the concepts of valence and arousal and introducing the concept of a grid representation of two-dimensional affective space. The subject is then responsible for identifying their current affective state within the nine-by-nine grid (figure 3.1). The general premise is that once the subject has undergone this training, this single item can be quickly, repeatedly, and reliably taken. Over time, subjects' assessments of their own state should become more reliable as they grow more familiar with the method. However, Affect Grid is inescapably reliant on the subjects ability to conceptualize affect in terms of valence and arousal and self-rate in these dimensions (Russell et al., 1989).

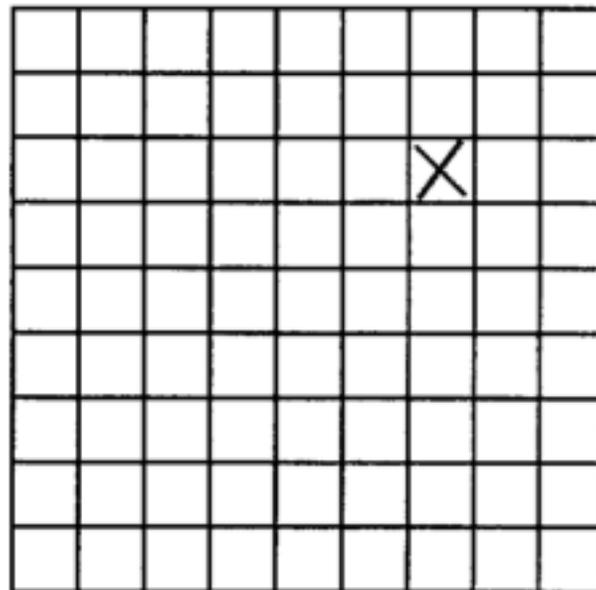


Figure 3.1, Russell's Affect Grid.

Rather than relying on subjects to understand affect, Bradley and Lang take the opposite approach with the graphic-based Self Assessment Manikin, SAM. SAM is presented essentially without instruction, asking subjects to choose from a series of cartoon characters the one that best represents how they feel (Figure 3.2). This step is repeated three times, first for arousal, then valence, then dominance. The premise is that the cartoons convey the key feelings associated with each dimension of affect and that they will resonate with subjects to the point that selecting from the choices is intuitive and reliable. This has advantages over semantic differential or Likert-type responses in that literacy or “putting emotions into words” are not necessary, but has obvious issues if subjects find the cartoons confusing (Bradley & P. J. Lang, 1994).

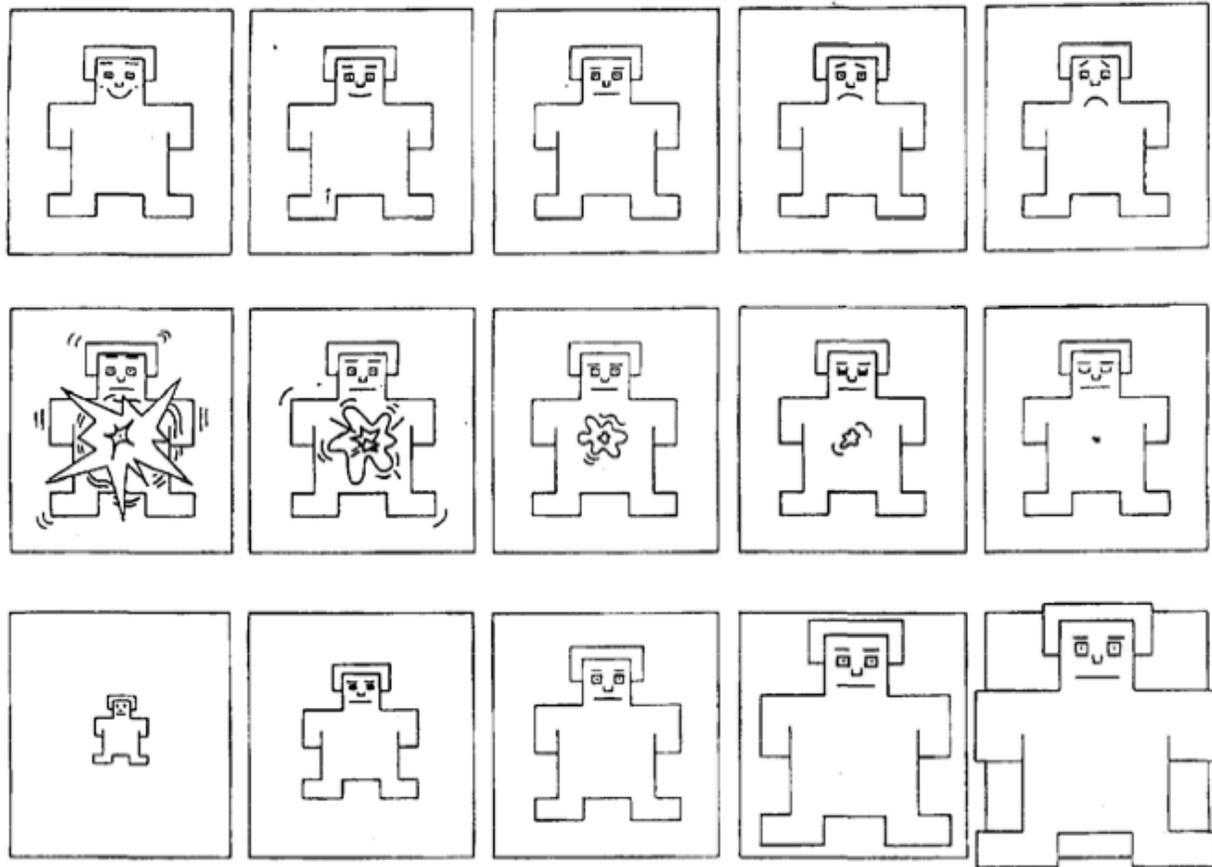


Figure 3.2. Bradley and Lang's Self Assessment Manikin. From top to bottom, Valence, Arousal, and Dominance dimensions.

Lang takes the concept of graphically representing affect one step further with his International Affective Picture System, IAPS. While IAPS is actually intended to induce affect rather than measure it, the work is still relevant for a number of reasons. Essentially, IAPS is an archive of photos that represent a validated instrument for eliciting a variety of emotional response across a range of cultures. Lang has assembled collection of images that is comprised of highly suggestive imagery including a range of subject matter from extreme violence to vicious animals to sleeping babies (Figure 3.3). In various experiments, subjects were exposed to the imagery and physiological

measures that are often mapped to emotional response, such as Galvanic Skin Response and EKG response, were collected. The final assortment of images in IAPS was comprised of those that elicited a consistent response across the study population (P. J. Lang, 1995).



Figure 3.3, Images from Lang's International Affective Picture System, IAPS.

The findings of Lang's work are relevant in two important ways. First, it is important to note that photographs can have a shared emotional meaning across large groups of individuals. Second, it is important that these photographs actually elicit a predictable emotional response from those viewing them. In other words, not only do certain images have the same affective meaning, but also that meaning can be predicted. Of course, IAPS is not intended as a measure of affect. It has not been validated as such and, further, the fact that the images in the library are known to *illicit* an affective response—obviously not a desirable trait when attempting to measure affect. However, this work does suggest that it may be possible to capitalize on the emotional legibility of imagery for the purpose of measurement. It is actually around this point—the emotional legibility of images, that a link to the affective computing literature will shortly become clear.

## *Affective Computing*

As discussed in the previous chapter, much of the recent literature in affective computing (notably (Boehner et al., 2005; 2007; Sengers et al., 2008)) has argued against the use of computers to measure affect. But the arguments presented in this literature, along with some of the critically designed systems, offer insight into the design of systems aiming to do just that. Sengers argues for the design of systems that allow users to create their own representations of emotion and meaning without computer intervention. A means of doing so is introducing ambiguity into a system's representation of emotion. Giving the user more control of representation and interpretation of emotion can pave the way for more meaningful interactions, as users are apt construct meaning where there might otherwise have been none (Leahu, Schwenk, & Sengers, 2008; Sengers et al., 2008). This suggests that there may be value in giving potential subjects a role in the co-creation of the representations of emotion used in the scale and ultimately the scale itself.

Gaver, et al examines the nature of ambiguity and its potential for generating new experiences or reflections. In examples provided from computational systems and famous works of art, ambiguity is defined as the interpretive relationship between the user and the artifact or system (Gaver, Beaver, & Benford, 2003) . Ambiguity signals and invites open interpretation, creating a system that is readily appropriable and encourages new reflection and new experiences (Sengers et al., 2004). Mateas uses the term interpretive flexibility for systems open to interpretation or appropriation (Mateas, 2001). In interpretively flexible systems, meaning is negotiated between the user, designer, and the computational intelligence of the system itself (Boehner et al., 2005). The importance of this point is that if the designers of the system, or the scale, simply choose and test the representations of emotion, something will be lost. Rather, the

system should be co-created with a variety of users in some way accounting for differences in interpretation.

Finally, research in user experience and experiential design suggests that designers need to build with the expectation that the complexity of human life and experience cannot be fully understood (Wright & McCarthy, 2008). This work begs the question of what roles do users, designers, and systems play in creating the meaning and experience of a system? Such equivocality appears to be at odds with the development of a tool for reliable measure of affect, yet, this is the task at hand. How then does one incorporate elements of co-creation and co-design, interpretive flexibility, and ambiguity—all ultimately highly subjective constructs—into the design of an objective, reliable measure of affect?

### *Key Design Decisions*

From the beginning, an objective of this work has been to leverage lessons on the subjectivity of emotion learned from affective computing in the *design* of the new measure all the while relying on the classic emotion literature to objectively specify then validate the measure. In this context, that means an approach that would (1) use a representation of emotion that offered opportunity for interpretation and personal meaning and (2) allow users (rather than the designers or the system) to determine the meaning of each representation of emotion. Following this, rigorous quantitative validation similar to that used in the validation of previous scales such as PANAS (Watson et al., 1988), Affect Grid (Russell et al., 1989), or SAM (Bradley & P. J. Lang, 1994).

Along these lines, the first important consideration was choosing a medium for representing emotion. Text-based instruments are obtrusive in the amount of time they

take to complete, and even simplified measures like Russell's Affect Grid (Russell et al., 1989) can be unwieldy. Others, such as SAM (Bradley & P. J. Lang, 1994) and Mood Map (Morris et al., 2010) instead make use of graphical representations. Along these lines, color has been a popular medium for representing emotion in affective computing work, e.g. (Sundström et al., 2007). A body of research does connect color with emotion (E. A. Mayer et al., 2008; Naz & Helen, 2004) but suggests that interpretation of color would prove to be too equivocal to be useful for assessment purposes (D'Andrade & Egan, 1974).

Photographs might offer a richer and more engaging representation of affect. The link between photographs and emotion is well researched, with evidence suggesting that photos themselves can be emotionally charged and can have universal emotional legibility as described above (P. J. Lang, 1995). Among family and friends, photos are often shared as a means of conveying and recalling the emotions. Photos documenting shared experiences, special moments, or events others might have missed carry mutually understood emotional meaning (Sondhi & Sloane, 2007). This has even been found to be the case when sharing photos over mobile phones (Kun & Marsden, 2007).

On the other hand, research has found that photos can also have very personal meaning based on prior or shared experiences (Chalfen, 1987). Further, photos themselves can represent a wide range of interpretive flexibility from images with more specific emotional content (such as an expressive human face) to something more ambiguous (a drop of water rippling in a glass). This tension between reproducible, shared meaning on the one hand and personal, interpretive flexibility on the other hand make photos a clear choice for the representation of emotion for this work.

As for the presentation of the photos to the subject, there were two

considerations. First, the number of photos presented to the subject should be maximized to increase choice. Of course, the constraints of the mobile platform limit this significantly. Testing on a variety of smart phones including Android-based devices, iPhone, and Blackberry indicated that only 16-24 images could be displayed at a time—no more than four across if the images were roughly square—and still be clearly legible. If a goal of the system was to account for individual differences and personal meaning found in photos, more photos would need to be made available. As such, the decision was made that a “More Photos” or refresh button would be added, allowing subjects to select from a new batch of photos if none could be found in the original batch. The need for this button—along a submit button—necessitated additional screen space, leaving room for a four-by-four grid, 16 photos, to be displayed at a given time. Such a layout does present another advantage, namely that it might allow for simple mapping of the photos into two-dimensional affective space a la Russell’s Affect Grid (Russell et al., 1989) or Mood Map (Morris et al., 2010).

### *Developing a Corpus of Affective Photographs*

With the selection of photos as the medium for representing affect, the question remained of what photos would be appropriate for such a task. IAPS was an obvious choice given the extensive validation and work done around it (P. J. Lang, 1995). However, as described above, in order for these photos to consistently evoke such responses, the subject matter they depict falls at extreme ends of the spectrum (including violence, sexuality, etc.). Further, the IAPS photos have been chosen to provoke emotional response, whereas the goal of PAM is assessment. Pilot testing found both of these issues to be problematic; subjects felt that most of the extreme images would not represent their day-to-day emotional state and further felt that the

photos influenced their state.

The online photo-sharing service Flickr was the next obvious choice given the vast number of images available and its developer-friendly API. Further, because of the social nature of Flickr, many of the photos in the service have associated keywords and tags that can be used in explicit searches. Using the Flickr API, roughly 9,000 Creative Commons-licensed images that had been tagged by the community of Flickr users with one of Russell's 28 words of affect (Russell, 1980) were downloaded. These tagged images would be likely to contain a wide variety of emotional content. Further, because a massive community of users actually determined which photos were considered emotional, it could be argued that the representation of emotion in this system was arrived at by the users themselves, keeping to the recommendations of Sengers (Sengers et al., 2008).

After removing images with inappropriate content, 7,714 photos remained. To begin to evaluate these images, they were then inserted into Aurora, a mobile phone-based emotion sharing system in which users select images that represent their current emotional state many times each day and share them with peers. Aurora had been previously tested with color as a medium for sharing emotion, and the majority of users found the photographs to be a richer, more expressive medium for sharing emotion (Gay, Pollak, Adams, & Leonard, 2011). Perhaps more importantly, after 70 individuals ages 18-55 used Aurora for a period of two weeks, it became clear that while users frequently selected certain images to represent their emotional states, others were completely ignored. Continuing with the co-creation theme, the most frequently selected photos were identified with the hope that these photos were most likely to contain imagery that subjects most readily associated with emotional states.

### *Development of the Photographic Affect Meter*

With a corpus of communally derived, emotionally charged photos in hand and a rough design for the system, it was time to begin testing and initial development of PAM. Approximately the top 100 most frequently selected photos from the Aurora pilot studies were identified and categorized based on Russell emotion word from which they were originally identified on Flickr. Using the grid metaphor and two-dimensional affect space presented by Affect Grid (Russell et al., 1989) and the mapping of emotion words in Russell's Circumplex Model of Affect, images were assigned to specific grid cells based on their emotion-word tag (Figure 3.4).

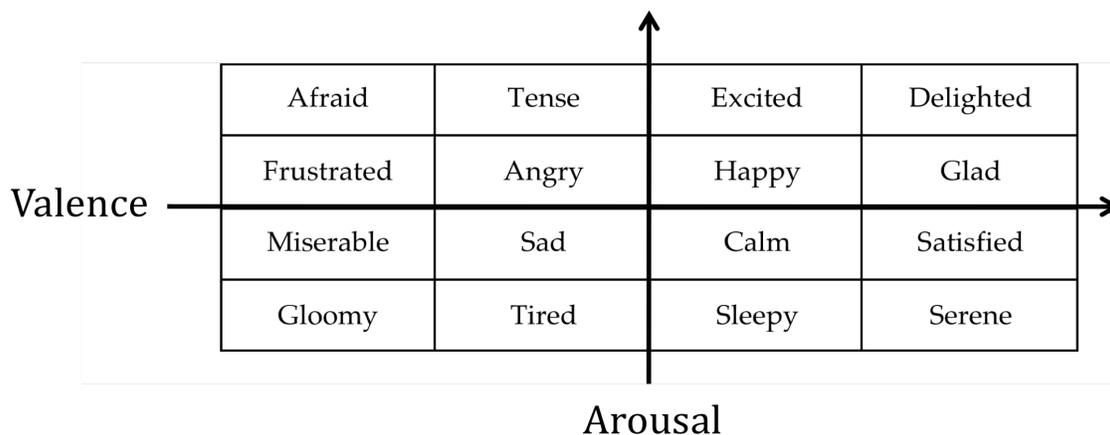


Figure 3.4, Mapping emotion words into two-dimensional space using Russell's Circumplex Model of Affect and Affect Grid yields starting positions for the most frequently used emotion photos.

With this prototype of PAM, 70 more testers completed 200 tests in which they first took PAM then subsequently were assessed with PANAS. Following this pilot, photos that were never selected or those that represented extreme outliers (e.g. a photo

intended to represent low arousal and negative affect was selected and followed by a high PANAS Positive Affect score) were discarded. The remaining 48 images—three per grid cell—would make up the PAM image set. Again, it should be noted that the methods employed to arrive at this point are quite consistent with the recommendations spelled out in the affective computing literature. Aside from the choice of using photographs as a medium for representing emotion and the general premise of the system, the core of the scale itself is entirely user-derived.

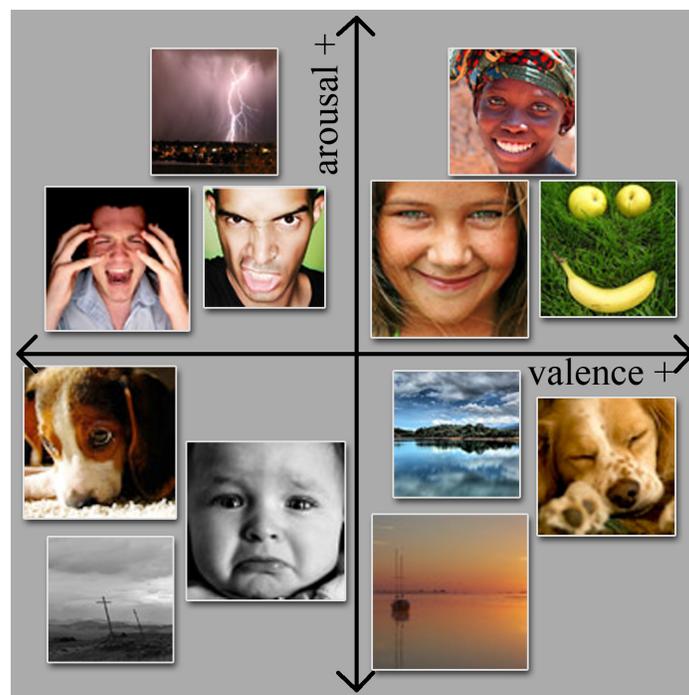


Figure 3.5, A sample of PAM images displayed in two-dimensional affective space.

### *PAM Scoring*

Scoring responses to the Photographic Affect Meter for the dimensions of arousal and valence is simple. Because the images are positioned in the scale in accordance with Russell's Circumplex Model of Affect (Russell, 1980) and Affect Grid (Russell et al.,

1989), the cell clicked dictates a value for each as per the table below (Table 3.6).

Arousal is scored from 1 to 4, starting from the bottom of the grid and increasing by one with each row toward the top. Valence is scored from -2 for the left most column to +2 for the right most column, with no score of 0 possible.

-2, 4	-1, 4	1, 4	2, 4
-2, 3	-1, 3	1, 3	2, 3
-2, 2	-1, 2	1, 2	2, 2
-2, 1	-1, 1	1, 1	2, 1

Figure 3.6, PAM Scoring, based on the cell location of the image clicked, displayed as Valence, Arousal.

Besides valence and arousal, many researchers are more interested in the constructs of Positive Affect (PA) and Negative Affect (NA) as reported by PANAS (Watson et al., 1988). Indeed, according to Google Scholar, PANAS is among the most widely cited measures of affect. As previously discussed, PANAS intentionally conflates valence and arousal to arrive at scores for PA and NA and as such this has become a common understanding of these concepts. Because PAM measures both valence and arousal, it should be possible to generate scores of both PA and NA. However, this scoring scheme was arrived at experimentally and will be discussed in Chapter 4.

### *The PAM System*

Ultimately, to ensure utility in Ecological Momentary Assessment and similar settings, PAM exists not only as a scale but also as a computer- and mobile phone-based system. Early EMA systems consisted of little more than an alarm or pager notifying the subject that it was time to take a pen-and-paper-based assessment (Csikszentmihalyi & Hunter, 2003; Larson & Csikszentmihalyi, 1983). These systems were constrained in a number of ways, chief among them the requirement that a subject carry with them at all times two extra items—a notification device and a journal. Further, because the resulting data was on pen and paper, the scope of studies was necessarily limited by the ability of researchers to digitize and code all of the journal entries. The advent of PDAs and handheld computers such as Palm Pilots improved the state of matters, reducing the number of extra devices a subject must carry to one and allowing for the collection of electronic data (L. Barrett & D. Barrett, 2001; Epstein et al., 2009; Stone & Shiffman, 2002). However, it is only now as smart phones rise to ubiquity that EMA methods will begin to realize their full potential.

Mobile phones, smart phones in particular, are always on, always with their owner, and possess a wealth of knowledge about their owner including where they have been, where they are going, with whom they are keeping company, and what they are doing (Fogg & Eckles, 2008). Because of this, not only do modern EMA methods not burden subjects with an extra device, but the devices the subjects are carrying contain a wealth of additional information that may be of relevance to the study. Current EMA systems such as the MyExperience platform (Froehlich et al., 2007)—employed by Mood Map (Morris et al., 2010) among others—are mobile phone-based applications that combine prompts and self-reports with the collection of a wealth of other data provided by the mobile phone.



Figure 3.7, PAM, running on a Google Android-based mobile phone.

Consistent with these modern EMA systems, PAM (Figure 3.7) has been designed to either fit within existing platforms such as MyExperience or operate as a stand-alone application. Primarily, PAM exists as a server-side, PHP-based API that allows for any device (or PC) to invoke an instance of PAM, carry out an assessment of affect, and recover the resulting PAM scores as data. For developers on Google's Android or Apple's iOS, this means the choice of either embedding PAM within existing applications to combine affective data with whatever else it is that they are collecting or using the stand-alone PAM app which simply allows for subjects to record

repeated measures of affect at prescribed times. For Blackberry users, a web app carries out a similar function. Another benefit to this approach is that PAM is also available as a widget to researchers wishing to use it in a computer-based lab setting or in web-based projects, not just those working with mobile phones.

From the user's perspective, PAM is simple. A text-based prompt (or instruction from a study administrator) instructs the user to select the photo which best describes how they feel "right now." The user can either tap or click the image of their choice, or tap or click the "More Photos" button. Requesting more photos refreshes the grid, randomly selecting a new set of photos. Once the user has selected a photo, a confirmation screen displays their choice and provides a confirmation button or the option to choose again. A single screen version of PAM is also available—in this version, the confirmation button appears at the bottom of the widget along with the button to request more photos. In this case, when the user taps or clicks the photo of their choice it becomes highlighted, and clicking the confirmation button submits their choice.

At this point, the PAM results are either stored locally on the device or sent to a secure server for storage. Each usage of PAM is associated with a study ID, a subject/user ID, a unique identifier for the usage, a time stamp, and the resulting PAM scores. Each of the first three unique identifiers can either be automatically generated by the system or supplied by the researcher if one of the embedded versions is used. This gives the researcher or developer several options for associating affect data with studies, subjects, and contexts, events, or conditions to be examined. At present, researchers wishing to use PAM must either have developmental resources at their disposal to embed PAM into existing applications, or they must use the stand-alone version. In the future, a complete and user-friendly web portal for setting up,

managing, and analyzing data from various kinds of experiments will be developed. At that point, PAM will be among the most easily incorporated digital scales available to researchers.

## CHAPTER 4

### VALIDATION OF PAM WITH THREE WIDELY USED MEASURES OF AFFECT

The previous chapter laid out an argument for the development of a novel measure of affect whose design is rooted in an affective computing literature calling for a more subjective view of emotion. The design and development of such a measure and system, PAM, was subsequently detailed. The next two chapters proceed to document the formal, quantitative evaluation of PAM, validating it as an objective measure of affect. In this chapter, PAM is compared with three of the most widely used measures of affect and found to be valid and reliable. Further, comparison with results from PANAS yields a means for developing Positive Affect and Negative Affect outputs for PAM. Finally, specific implications as well as limitations of PAM and these validation methods are discussed, but a broader arching discussion of implications and applications of PAM is reserved for a later chapter.

#### *Background: Validating a Measure of Affect*

Establishing the validity of any measure can be an arduous task involving the exhaustive assessment of various forms of validity (Campbell & Fiske, 1959; Cook & Campbell, 1979). Cronbach describes the process as iterative and nearly unending, but points out that a logical first step is to establish construct validity (Cronbach, 1988). In this case, the question of whether PAM is in fact measuring affect and doing so in a way that is meaningful and fits within expectations must be addressed. Campbell and Fiske classically identify two key components to establishing construct validity. First, a researcher must establish convergent validity, which is the extent to which the measure converges with other similar or theoretically correlated measures. Second, the researcher must establish discriminant validity, which is the extent to which the

measure does not converge with theoretically unrelated measures (Campbell & Fiske, 1959).

Along these lines, it is common practice when introducing a new measure of affect to validate the results of that measure against existing, previously validated measures. The most straightforward way to accomplish this is to assess subjects with two or more scales one immediately following the next. In some cases this is done in response to stimuli such as photographs or human facial expressions (Bradley & P. J. Lang, 1994; Russell et al., 1989) and in other cases subjects are simply asked about how they are feeling at a given moment (Russell et al., 1989; Watson et al., 1988). In either scenario, the premise is that if the outputs of the two scales converge significantly, then the argument can be made that the two scales are indeed reliably measuring the same phenomenon. As an added benefit, many of the previously established measures are comprised of various subscales, many of which behave completely independent of one another, providing opportunity to establish discriminant validity as well. An examination of the validation practices from existing key measures of affect should shed light on this subject.

Perhaps the clearest example of this process can be found in the validation of Russell's affect Grid. In the most illustrative study, subjects are asked to report on how they were feeling "right now" using Affect Grid, PANAS, and Semantic Differential scale. The results demonstrated significant correlations between the valence measures of Affect Grid and Semantic Differential (.77) and the arousal dimensions of Affect Grid and Semantic Differential (.80). Linear regression was used to estimate the correlation between the two-dimensional output of Affect Grid with the conflated PANAS Positive Affect score (.62) and PANAS Negative Affect score (.48). Russell reported that these findings were more than adequate to declare substantial convergent validity for Affect

Grid. Further, Russell found no or weak correlations between constituent subscales that should theoretically be independent, and reported this as evidence of discriminant validity.

Other researchers have taken on similar approaches. Watson et al (Watson et al., 1988) in part validated PANAS by assessing subjects with PANAS and five other scales representing various constructs of positive and negative affect. Bradley and Lang (Bradley & P. J. Lang, 1994) emphasized groups over individuals in their approach with SAM, asking a group of subjects to rate the emotional content in series of images from IAPS using SAM. A second group then rated the images using Semantic Differential and the results were compared and found to be convergent.

### *Choice of Existing Measures*

The first key decision in the validation process for PAM is which existing measures to use in establishing construct validity. Each of the measures of affect described thus far would offer something different to the validation of PAM as a new measure of affect. Ultimately, the selection of three existing measures—Affect Grid, the Self Assessment Manikin and PANAS—was based on the validity and high citation counts (cited 446, 1031, and 8697 times, respectively according to Google Scholar as of September 2, 2011) of the existing scales, the form factor of the scales themselves, and finally the primary research audience for PAM.

Affect Grid was chosen as the first measure of affect to be used in the validation of PAM in part because of an element of trustworthiness it owes to having been developed and validated by Russell himself explicitly around the Circumplex Model of Affect (Russell, 1980) and Semantic Differential Scales (Mehrabian & Russell, 1974) which are seen as benchmarks in the emotion literature. More pragmatically, Affect

Grid is attractive because it is similar in form factor to PAM, both in that it reports Valence and Arousal as the primary outputs but also in that it is grid based. When using Affect Grid, users complete instructions and examples training them to evaluate the valence and arousal of their current emotional state and identifying it on a grid. The result of a completed assessment using Affect Grid is two measures, one for valence and another for arousal. Further, Affect Grid was originally intended as a brief, momentary measure of affect as is PAM (Russell et al., 1989).

The Self Assessment Manikin, or SAM, was chosen for many of the same reasons noted above. With SAM, a subject is shown three series of five or ten cartoon characters and asked to choose the characters that best represent how they are currently feeling. The first series is intended to measure arousal, the second valence, and the third dominance. As a measure it has been extensively validated and likewise outputs Valence and Arousal (as well as dominance) as the primary measures of affect. SAM is also intended as a brief measure of affect and even exists in a computerized format. However, what differentiates SAM from Affect Grid and makes it a suitable additional choice for the validation of PAM is that rather than employing a grid or semantic approach, SAM relies on graphical representations of emotion. Like PAM, SAM doesn't require that the subject has a knowledge of emotion theory or think too heavily about the meaning of groupings of emotion words (Bradley & P. J. Lang, 1994).

PANAS was the final choice for a number of reasons. First, PANAS is one of the most widely used and extensively validated measures of affect, having been cited over 7,000 times according to Google Scholar. Second, PANAS is widely used in health and medical research, the domain in which Ecological Momentary Assessment methods play the largest role (Cohen & Pressman, 2006; Pressman & Cohen, 2005). For both of these reasons, demonstrating convergent validity with PANAS could in many ways be

the most important aspect in the validation of PAM. Further, given that Affect Grid and SAM are grid-based and graphical, the inclusion of PANAS adds a semantic angle to the validation. PANAS consists of 20 single-word items, each of which is an affect word denoting combinations of valence and arousal. Subjects rate the extent to which they are feeling each, and the items are summed to produce measures of positive and negative affect. As such, validation with PANAS presents an opportunity to assess the use of PAM for two slightly different conceptualizations of affect, Positive Affect and Negative Affect.

Finally, the question of how to actually administer the scales must be answered. On the one hand, assessing subjects for affect with two scales in response to a stimulus or asking, “how do you feel about X” provides an additional point of reference for validation, particularly if the subject matter has known emotional content, e.g. IAPS (P. J. Lang, 1995). However, given that the intent of PAM is to assess emotional states, it seemed more appropriate to simply assess the subject with two brief measures of affect, one immediately following the next, only asking that they document how they are feeling at that given moment.

### *Objectives*

At this point, the primary objective of this study should be clear: to establish convergent and discriminant validity for PAM through comparison with results from three established and widely accepted measures of affect. The approach for this is straightforward and was carried out over two studies. In the first study, subjects were assessed for their current emotional state using PAM and either Affect Grid or SAM. As the results will show, PAM Valence and Arousal outputs were found to correlate strongly with those from Affect Grid and SAM. In the second study, subjects were

assessed in the same way but with PANAS as the point of comparison. In this case, PAM was found to correlate strongly with PANAS Positive Affect scores and moderately with PANAS NA.

A secondary objective has been established as well: to assess and develop the ability of PAM to measure affect in terms of Positive and Negative Affect in addition to Valence and Arousal. As discussed in previous chapters, Positive Affect (PA) and Negative Affect (NA), at least as envisioned by Watson et al (Watson et al., 1988), are an intentional conflation of valence and arousal. PA increases as both valence increases to the positive and arousal increases, whereas NA increases as valence increases to the negative and arousal increases. This conflation makes PA and NA extremely attractive constructs for researchers in the health sciences, in which not only the valence of the emotion but often the arousal dictate the impact that it will have on the individual (Cohen & Pressman, 2006).

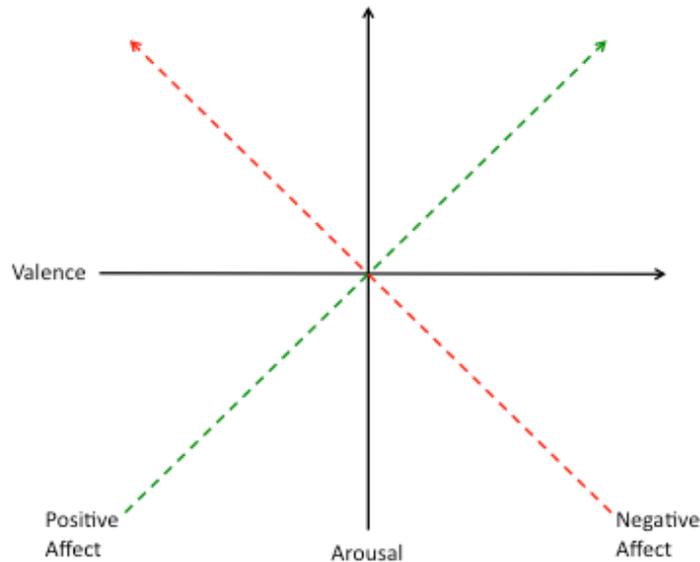


Figure 4.1. Two-dimensional affective space as defined by Valence and Arousal.

Positive Affect conflates Valence and Arousal, and increases with each. Negative Affect increases as Arousal increases and Valence moves toward the negative.

Russell describes PA and NA as a “45 degree rotation” of the standard valence and arousal dimensions that captures much of the variability that makes up most measures of affect. This rotation (see Figure 4.1) simply means what has already been stated above—that as Valence increases toward the positive and Arousal increases, so does Positive Affect and that as Valence increases towards the negative and Arousal increases, so does Negative Affect. Further, Russell uses linear regression to model PA and NA on the Valence and Arousal scores produced by Affect Grid and finds good predictability (Russell, 1980). Employing a similar tactic, the findings from the second study presented here are used to generate PA and NA scores for PAM.

### *Method*

Two studies were conducted in a nearly identical matter. Study 1 was designed to evaluate the Valence and Arousal output of PAM and Study 2 was designed to evaluate and develop Positive Affect and Negative Affect scores for PAM.

#### *Participants*

Study 1 consisted of 81 individuals recruited through a combination of university departmental and student listservs, snowball sampling, flyers, and postings on a variety of websites. Subjects who participated were given the option of providing an email address to be entered into a drawing to win an iPod. Because Study 2 would be used to explore the relationship between PAM Valence and Arousal and Positive and Negative Affect, a substantially larger sample was collected. Without knowing a priori what

mathematical or computational model would be used to generate PAM PA and NA scores, it was deemed prudent to collect as large a data set as the sampling methods afforded. For study 2, 315 individuals were recruited using the exact same means as Study 1. Demographic data was collected during Study 2, finding the sample to include 45% male; Asian 21.3%, Black or African American 7.8%, Hispanic or Latino 8.5%, White 54.6%, and Other 7.8%.

### *Procedure*

#### Study 1: Affect Grid and SAM

Subjects ( $N=81$ ) logged into a study website and consented to participate. Subjects were randomly assigned into one of two primary groups. Subjects in the first group completed the PAM assessment with the prompt “Select the image that best captures how you feel right now” and were also asked to complete a web-based version of Russell’s Affect Grid. Subjects were asked to read through the instructions and examples before continuing to the page for the actual Affect Grid assessment. Subjects in this group were randomly assigned to either completing PAM then Affect Grid, or Affect Grid then PAM. The second group completed the PAM assessment with the same prompt, and was also asked to complete the SAM Arousal assessment, followed by the SAM Valence assessment, each with the prompt “Select the figure that best represents how you feel right now.” As with the first group, the order in which subjects completed the two assessments was randomized. Ultimately, 35 individuals completed the Affect Grid task and 46 completed the SAM task. Upon completing both assessments, subjects signed out of the website and were finished.

#### Study 2: PANAS

Subjects ( $N=315$ ) logged into a study website and consented to participate. Subjects completed the PAM assessment with the prompt “Select the image that best captures how you feel right now” and were also asked to complete a web-based version of PANAS, with the time scale of the PANAS prompt also changed to be “Indicate to what extent you feel this way right now, that is, at the present moment.” Subjects were randomly assigned an order of completion, either PAM first then PANAS or PANAS first then PAM. All subjects were asked to provide basic demographic information. This provided a data set with which to examine the relationship between each subject’s PAM and PANAS results, both theoretically representing measures of state affect. Upon completing both assessments, subjects signed out of the website and were finished.

### *Measures*

All subjects were assessed with PAM. In PAM, a subject is shown 16 images arranged in a four-by-four grid and must choose the one that best represents their current affective state. A “more photos” button refreshes the selection, randomly selecting a new batch of 16. Based on the location of the photo within the grid, scores for Valence and Arousal are assigned: PAM Valence (PAM Val) is computed as -2 to +2, moving left to right across the grid and PAM Arousal (PAM Ar) is scored 1 to 4 moving bottom to top up the grid, with higher scores representing higher arousal.

-2, 4	-1, 4	1, 4	2, 4
-2, 3	-1, 3	1, 3	2, 3
-2, 2	-1, 2	1, 2	2, 2
-2, 1	-1, 1	1, 1	2, 1

Figure 4.2. PAM Scoring, based on the cell location of the image clicked, reported as Valence, Arousal.

Subjects in Study 1 were assessed with either Russell’s Affect Grid (abbreviated RAG in variables for legibility) or the Self-Assessment Manikin (SAM). RAG, as described above, is a grid-based, single-item measure of affect in which subjects read through instructions and examples then identify the position in a nine-by-nine grid that best represents their current affective state. Based on their position they select in the grid, RAG Valence (RAG Val) and RAG Arousal (RAG Ar) scores are generated. In SAM, also described above, subjects are shown three series of nine cartoon characters and must select the character from each series that best represents their current affective state. The first series produces a SAM Arousal (SAM Ar) score, the second SAM Valence (SAM Val) and the third a score for dominance. Since PAM ignores dominance, subjects in this study were not asked to complete the dominance assessment.

Subjects in Study 2 were assessed with PANAS. PANAS consists of 20 items, each a single emotion word. Subjects are instructed to identify on a scale of 1 to 5 the extent to which they are feeling each word “right now.” The responses to ten of the words are summed to produce the PANAS Positive Affect (PANAS PA) score and the

remaining ten are summed to produce the PANAS Negative Affect (PANAS NA) score. Subjects in Study 2 were also asked simple demographic questions, as well as assessed with a brief five factor inventory of personality (Gosling, Rentfrow, & Swann, 2003).

## *Results*

The results of these two studies as reported below represent strong evidence of construct validity for PAM. First, in Study 1, comparative results with SAM and RAG provide evidence of the validity of the PAM Valence and Arousal scores. In Study 2, PAM scores are compared to PANAS scores to derive and validate a means of computing Positive Affect and Negative Affect scores for PAM. In this section, the following abbreviations may be used: Russell's Affect Grid will be referred to as RAG, Valence as Val, and Arousal as Ar, Positive Affect as PA, Negative Affect as NA. Table 4.1 displays the descriptive statistics for the four mood scales in these studies.

### *Study 1: Affect Grid and SAM*

Study 1 aims to validate the Valence and Arousal scores output by PAM by comparing it to the Valence and Arousal scores output by SAM and RAG when taken one after the next by the same subject who is presumably in the same affective state. The descriptive statistics for the three affect scales are presented in Table X.

<b>Scale</b>	<b>M</b>	<b>SD</b>
<i>PAM</i>		
Valence	0.37	1.48
Arousal	2.36	1.03
<i>Affect Grid</i>		
Valence	5.74	1.82
Arousal	5.00	1.80
<i>SAM</i>		
Valence	5.57	2.04
Arousal	4.04	2.30

Table 4.1, Descriptive statistics for responses to PAM ( $N=81$ ), Affect Grid ( $N=35$ ) and SAM ( $N=46$ ). Note that PAM Val ranges from -2 to +2, PAM Ar ranges from 1 to 4 and Affect Grid and SAM scores range from 1 to 9.

Following the work of Russell (Russell et al., 1989), Bradley and Lang (Bradley & P. J. Lang, 1994), and Watson et al (Watson et al., 1988), the most straightforward means of demonstrating convergent validity is through intercorrelations between the constituent measures. In each of the above referenced works, five- to nine-item affect scales (PAM is 16-item) are treated as continuous variables and assessed via linear regression or correlation. Table 4.2 presents correlations between each of the constituent scales for PAM, RAG, and SAM. By design, the same individual was never assessed with both SAM and RAG, so no correlations between them are available. Note that no order effect was seen between individuals taking either PAM or the second scale first versus second.

	<i>Valence Scales</i>			<i>Arousal Scales</i>	
	PAM Val	RAG Val	SAM Val	PAM Ar	RAG Ar
<i>Valence Scales</i>					
PAM Val					
RAG Val	<b>.71*</b>				
SAM Val	<b>.61*</b>	NA			
<i>Arousal Scales</i>					
PAM Ar	-.04	.10	.10		
RAG Ar	.16	.08	NA	<b>.67*</b>	
SAM Ar	-.21	NA	-.19	<b>.45*</b>	NA

Table 4.2, Correlations between affect scales, for PAM and RAG, N=35, for PAM and SAM, N=46. \* Denotes significance at  $p < .01$ . Key correlations are bolded.

For Valence, there are significant correlations between both PAM and RAG (.71) and between PAM and SAM (.61). Similarly for arousal, there are significant correlations between PAM and RAG (.67) and between PAM and SAM (.47). Taken together, these results point to the strong convergent validity of PAM. No other significant correlations were found among the other measures. The lack of a correlation between PAM Valence and any other measure of Arousal and between PAM Arousal and any other measure of Valence are suggestive of discriminant validity as these two dimensions are known to be independent (Russell, 1980; Russell et al., 1989). Note that a similar lack of correlation is seen between RAG Valence and RAG Arousal and SAM Valence and SAM Arousal.

### *Study 2: PANAS*

Study 2 aims to compare PAM scores with PANAS scores in order to validate the

use of PAM for the Positive Affect and Negative Affect constructs. Descriptive statistics for the two scales are presented in Table 4.3. Population means reflecting more positive than negative affect are consistent with previous findings (Crawford & Henry, 2004; Watson et al., 1988). It should also be noted that while the maximum range of PANAS is 10 to 50 for both PA and NA, the observed range was only 10 to 43 for PA and 10 to 39 for NA.

<b>Scale</b>	<b>M</b>	<b>SD</b>
<i>PAM</i>		
Valence	0.29	1.59
Arousal	2.56	0.93
<i>PANAS</i>		
Positive Affect	25.52	7.50
Negative Affect	15.64	6.25

Table 4.3, Descriptive statistics for responses to PAM and PANAS,  $N=315$ . Note that PAM Val ranges from -2 to 2, PAM Ar ranges from 1 to 4, and PANAS PA and NA range from 10 to 50.

For the purposes of this analysis, the PAM Valence score is recoded to a score of 1 to 4 counting from left to right for PA and recoded to 4 to 1 from left to right for NA. Linear regression found that PAM Val and PAM Ar were able to predict PANAS PA as detailed in Table 4.4.

	<i>b</i>	<i>SE b</i>	$\beta$
Constant	7.79	1.15	
PAM Valence	5.36	.32	.68*
PAM Arousal	1.16	.33	.14*

Table 4.4, Linear regression for PANAS PA on PAM Valence and PAM Arousal,  $R^2=.51$ .

\* Denotes  $p < .001$ .

Given the good fit of the linear model for these data and this sample, a simple linear combination of PAM Valence and PAM Arousal scores can be used to compute Positive Affect and Negative Affect. In the model, PANAS PA varies based on a 4.86:1 ( $b_{Valence}=5.36$ ;  $b_{Arousal}=1.16$ ) ratio of PAM Valence to PAM Arousal; for sake of simplicity, the ratio is rounded up to 5:1. This means that a basic formula can then be used to compute a PAM PA as function of five times Valence and the base Arousal score. This combination produces scores in a range of 1 to 19, which is counterintuitive given that the grid only contains 16 choices,, so the scores are compressed and rounded to generate a 1 to 16 score, with a single value represented by each cell in the grid. The formula to generate PAM PA from a subject's image selection for this sample is:

$$PAM\_PA = ROUND(16/19 * (5 * PAM\_VAL + PAM\_AR - 5))$$

Similarly, linear regression found that PAM Val and PAM Ar predicted PANAS NA, albeit with much less certainty (Table 4.5).

	<i>b</i>	<i>SE b</i>	$\beta$
Constant	11.51	1.34	
PAM Valence (NA)	2.32	.35	.35*
PAM Arousal	.425	.36	.06

Table 4.5, Linear regression for PANAS NA on PAM Valence and PAM Arousal,  $R^2=.18$ .

\* Denotes  $p < .001$ .

In the NA model, PANAS NA varies based on a 5.83:1 ratio of PAM Valence to Arousal. While the regression model is not as conclusive as that for PA, the ratio of  $\beta$  values for Valence to Arousal is similar. In the interest of simplicity and to maintain the same scale as the PA model, a 5:1 ratio is used to yield the formula:

$$PAM\_NA = ROUND(16/19 * (5 * PAM\_VAL\_NA + PAM\_AR - 5))$$

Because the formulae involve rounding rather than using the exact  $b$ -values, the  $R$ -values from the regression models are no longer appropriate estimates of the correlation between PAM scores and PANAS scores. So, for the sake of analysis, PAM PA and PAM NA scores were computed for each of the data subjects' responses. Correlations were computed between PANAS PA and NA and the newly derived PAM PA and NA scores. The rounding had a negligible impact on the resulting correlations; the results are presented in Table 4.6.

	<i>Positive Affect</i>		<i>Negative Affect</i>
	PAM PA	PANAS PA	PAM NA
<i>Positive Affect</i>			
PAM PA			
PANAS PA	<b>.72*</b>		
<i>Negative Affect</i>			
PAM NA	-.90*	-.66*	
PANAS NA	-.37*	-.30*	<b>.34*</b>

Table 4.6, Correlations between key aspects of PANAS and PAM,  $N=315$ . \* Denotes significance at  $p<.001$ .

PAM PA strongly correlates with PANAS PA (.72), indicating that PAM scores are in fact a good indicator of positive affect and continues with the theme of finding strong convergent validity for PAM. While it is difficult to compare correlations between studies, it is worth noting Russell’s Affect Grid produced a correlation of only .62 with PANAS PA (Russell et al., 1989), and that was considered to be sufficient in the validation of RAG. This makes a convincing argument that PAM is a valid measure of positive affect.

Further, note that there is a weak negative correlation (-.37) between PAM PA and PANAS NA, which is consistent with—and not significantly different from,  $p=.08$ —the observed correlation between PANAS PA and PANAS NA reported here (-.30). This negative correlation is also consistent with theoretical expectations. Revisiting PANAS ten years after it’s creation, Watson, et al (Watson et al., 1999) discuss oft repeated findings that PA and NA are not in fact polar opposites, and in a follow-up study once again find them to be negatively and weakly to moderately correlated. This

serves as further evidence that PAM PA is indeed mirroring PANAS PA.

However, while the PAM PA scores correlate strongly with PANAS PA, the correlation between PAM NA and PANAS NA is significant but not nearly as strong (.34 compared to .72). This is not surprising; as just discussed, theory predicts that NA and PA will be slightly negatively linked and that is what is found here. But, whereas PANAS has two separate items measuring PA and NA independently, PAM is a single, one-item measure used to derive both scores. For that reason, PAM PA and PAM NA can only be strongly and negatively correlated. Given the strong correlation between PAM PA and PANAS PA, it logically follows that the correlation between PAM NA and PANAS NA cannot be as strong. This matter is discussed further later in the discussion.

Finally, PAM PA did not correlate significantly with any of the reported items of personality (Extraversion, .03; Agreeableness, .02; Conscientiousness, -.06; Emotional Stability, -.04; Openness, .04), nor did PAM NA ((Extraversion, -.02; Agreeableness, -.01; Conscientiousness, .06; Emotional Stability, .04; Openness, -.04). This expected result—PAM Positive and Negative Affect scores do not correlate with measures of a concept with which there is no theoretical relation—is further evidence of discriminant validity for PAM.

### *Discussion*

The goal of these two studies was to first establish construct validity for PAM and second establish a means of generating PAM Positive Affect and Negative Affect scores. To the first goal, Study 1 found that when taken one after the next, PAM Valence and Arousal scores correlated with Valence and Arousal scores for both Affect Grid and SAM. Study 2 found that PAM could be used to predict PANAS scores, PA in particular, with much certainty. The procedures followed, and ultimately these

findings, are very much in line with the studies used to validate the other major scales described thus far (Bradley & P. J. Lang, 1994; Russell et al., 1989; Watson et al., 1988). PAM scores converge with existing, validated measures where they should and for measures that should be theoretically independent, no relationships are found. This is sufficient to establish both convergent and discriminant validity as discussed by Campbell and Fiske (Campbell & Fiske, 1959). Together, this is a strong argument for the validity of PAM. To the second goal, the findings from Study 2 were used to produce a simple linear combination of PAM Val and PAM Ar that yields PA and NA scores, with the PA score found to strongly correlate with PANAS PA.

#### *PAM Positive Affect*

The ability of PAM to produce a valid measure of Positive Affect is important for several reasons. First and foremost, it affords the use of PAM in numerous domains, particularly in health and medical research where PA is considered to be *the* key construct of affect, as mentioned above (Cohen & Pressman, 2006). Further, in these domains, PANAS is widely considered to be the gold standard, and there should be little doubt that PAM PA can be used as an acceptable surrogate for PANAS PA. PAM will likely be of great interest in these domains as adoption of Ecological Momentary Assessment in clinical trials increases (Mancuso & Charlson, 1995; Pressman & Cohen, 2005). For these purposes, only computer-based, brief measures are appropriate and, at this time, PAM is very likely the only validated measure of Positive Affect suitable for use in EMA.

Also worth noting is that the addition of a PA score to PAM might make it the only brief, validated scale that measures affect both in terms of Valence and Arousal and in terms of PA and NA. Russell reports on the ability of Affect Grid to predict

PANAS PA and NA scores, but doesn't take the next logical step of giving researchers a means for computing those scores from Affect Grid results (Russell et al., 1989). This is likely partly because Affect Grid is originally pen and paper based, so any such calculations would have to be carried out by the researcher after the fact. None of the other scales discussed attempt to report these different constructs of affect. This essentially means that if researchers are interested in both Valence and Arousal and PA and want to use only one scale, PAM is the only choice among brief measures.

#### *PAM Negative Affect*

While the PAM PA score correlates strongly with PANAS PA, the scores for PAM NA and PANAS PA demonstrate a weaker, but still significant relationship warranting further discussion. This is consistent with past findings and theoretical expectations. The Circumplex Model of Affect establishes the independence of the Valence and Arousal dimensions of affect (Russell, 1980) and results presented here and elsewhere continue to uphold this. For PA and NA, the conflation complicates matters slightly, as increases in Arousal tend to increase both the PA and NA scores. Watson et al in the original validation of PANAS (Watson et al., 1988) as well as much subsequent research has argued that only a weak, negative correlation exists between the NA and PA (E. Harmon-Jones, C. Harmon-Jones, Abramson, & Peterson, 2009; Russell et al., 1989; Watson et al., 1999). Again, the conflation complicates matters, and in some cases such as intense anger when arousal is at its highest, PA and NA actually become difficult to discern (E. Harmon-Jones et al., 2009).

Here lies the difficulty of reporting both PA and NA with a single-item measure. While in reality the relationship between the two constructs is complex and contextually dependent, if they are both derived from a single-item measure such as

PAM, the relationship is necessarily explicit and unchanging. Affect Grid suffers from the same limitation, and indeed, Russell reports substantially weaker estimated correlation with PANAS NA (.48) than with PANAS PA (.62) (Russell et al., 1989). Not surprisingly, the  $R$ -value reported by Russell is not substantially different from  $R$ -value output from PAM linear regression model for NA prior to rounding and scaling (.42, reported as  $R^2=.18$  above).

What then, would be required for PAM to generate NA scores of sufficient validity and reliability? PANAS accomplishes this by using two separate scales each with independent items. It should be clear from the discussion above that only by creating a second version of PAM geared specifically towards NA would this be possible; as long as NA and PA scores are derived from a single-item, one of the scores will always be unable to produce reliable results. A later discussion will focus on the possible creation of a NA-focused variant of PAM.

### *Other Limitations*

There are of course other minor limitations of PAM based on the scale itself and the means with which it has been validated to this point. The primary limiting factor of PAM itself is the compressed nature of the scale. While scales such as Affect Grid and SAM produce 1-9 scores of valence and arousal (there is also a 5-point SAM), PAM only produces a 1-4 score. For PA and NA, PANAS produces 40-point scores while PAM PA and NA are only 16-point. While the compressed scale might mask variability or limit the power of some analyses, it is unavoidable due to the form factor target of mobile phones. Further, because each “point” of PAM is so much more than just a grid cell or Likert-type point, it can be argued that the elements themselves are more distinct, each selection more meaningful, and effect sizes may be larger in spite of the compressed

scale. This true not only theoretically (i.e. richness and emotional meaning of imagery) but it is something that has been born of the actual development of the scale—each image was carefully selected through the iterative, data-driven process previously described. Further, each instance of the grid is generated at random with three possible images per cell, and individuals are free to request “more photos” as many times as they like, exploding the number of possible “items.” Finally, this is less of an issue for Ecological Momentary Assessment and similar use cases where significantly more data points are collected from each subject.

The validation procedures detailed in this study have left a few unanswered questions regarding PAM of which researchers should be aware. For example, to what extent is PAM susceptible to repeated measures effects? This does warrant further testing, however, given that every instance of PAM is different and there are 48 images to choose from, this is less likely as the same images and grid layout is not displayed every time. Further, because a PAM assessment is such a quick process, subjects have less incentive to simply go with the same response from their previous assessment.

Additionally, the fact that subjects were only asked to report on how they were feeling at the time of the study may have had an impact on the results. As such, it is less likely that extreme emotions were captured in the data. For example, the range of PANAS scores observed was somewhat less than the maximum possible range for both PA and NA, likely because no one would choose to participate in an online study when they are at their most excited or most upset. This is only a minor concern, but may be of interest to researchers interested in assessing subjects who are experiencing emotional extremes.

Finally, it should be noted that the current work on PAM has not yet investigated how cultural differences may shape the interpretation of emotional photos, or even

whether this model of affect would be appropriate to apply cross-culturally. While the sample used in both studies was relatively racially diverse, the participants were individuals living, working, or going to school in the United States. While previous work by Lang does suggest that the emotional meaning of photos can be culturally independent (P. J. Lang, 1995), this particular set of photos has not been validated in that way. Those looking to incorporate PAM into future projects should be mindful of this fact until more work along these lines is conducted.

### *Conclusions*

These two studies set out to establish the validity of PAM as a viable, brief measure of affect. Study 1 found that PAM Valence and Arousal scores converge with those obtained by Affect Grid and SAM, two widely accepted measures of affect. Study 2 found that PAM scores could be used to reliably predict Positive Affect as measured by PANAS and, to a lesser extent, Negative Affect. Using the results from Study 2, PAM Positive Affect and Negative Affect scores can now be computed using a linear combination of Valence and Arousal, and those scores (PA in particular) were found to correlate with PANAS scores. While there remains doubt about the utility of the PAM NA score, each of the remaining scores, Valence, Arousal, and PA all appear to be sufficiently valid and ready for use in future studies.

## CHAPTER 5

### PAM VALIDATION USING CLASSIC MOOD INDUCTION

Thus far, this dissertation has described the design, development, and classic validation of the Photographic Affect Meter, PAM. In the previous chapter, PAM results were found to correlate with the results of a selection of the most widely used measures of affect. This chapter seeks to demonstrate that in an experimental setting, subjects who are assessed with PAM report on their affective state in a manner consistent with expectation. To accomplish this, a variation on classic film-based mood induction is used to evoke a range of emotions in subjects who are subsequently assessed with PAM. The specific implications and limitations of these findings are discussed in detail here, with the broader discussion reserved for the final chapter.

#### *Background*

The primary means for establishing construct validity for a new scale is by providing evidence of convergence with other validated scales that are intended to measure the same construct (Campbell & Fiske, 1959) when other such scales are available. This has been the norm in emotion research, as this method has been employed by each of the prominent scales used in the measurement of affect, including PANAS (Watson et al., 1988), Affect Grid (Russell et al., 1989), and the Semantic Assessment Manikin SAM (Bradley & P. J. Lang, 1994). Interestingly, each of these scales has been validated against some combination of the others and perhaps semantic differential. While this may seem somewhat self-referential, the validity of these scales has not widely been questioned. This may be partially because each has been put to extensive use with more than adequate results in a wide variety of studies.

This assurance that when put to use in an experimental setting, the scale will

return the expected results is essential. While the assurance can be obtained through years of subsequent, confirmatory use, it can also be obtained preliminarily as part of the validation process. This tactic has been employed in the past by developers of emotion scales.

For example, the Self Assessment Manikin, or SAM—a three-item, graphical scale that measures affect in the dimensions of arousal, valence, and dominance—was validated in this fashion. Subjects assessed with SAM are asked to select one each from three sets of five or ten cartoon characters the image that best represent how they are currently feeling. One part of the SAM validation found the results to converge with those generated by other validated measures. In the second part of the validation process, subjects were shown images from a validated library of affect inducing images, asked to rate the emotion that the images convey with SAM, and were found to rate the images similarly and with expected results (Bradley & P. J. Lang, 1994; P. J. Lang, 1995) in each of the three measured dimensions.

Russell's Affect Grid—a grid-based, single-item scale that measures the valence and arousal dimensions of affect—was also validated in a similar fashion. As one step in the process, Affect Grid results were found to converge with those from other validated measures. For a second step, subjects were asked to rate facial expressions using the grid with similarly successful results (Russell et al., 1989).

In any case, whether as an explicit step in the validation process or as part of ongoing research utilizing a scale, the basic premise is that if subjects are experiencing a known or predictable affective response, measurement with that scale should reflect that. In a controlled setting, the best method for replicating this circumstance is by inducing specific forms of affect then assessing subjects with the scale in question. Evoking a targeted affective response—often referred to as *mood induction* or *mood*

*elicitation*—is common in experimental psychology and is supported by a substantial literature and history of practice.

### ***Mood Induction***

There are countless strategies for inducing affect, including directed imagination, reading of emotional statements, watching films, listening to music, social interaction with confederates, gifts, and so on. These strategies vary in effectiveness, and often the most successful approach is a combination of strategies, such as directed imagination while listening to music, or conducting a series of these tasks in serial (Westermann, Spies, Stahl, & Hesse, 1996). Velten (Velten, 1968) developed what is historically they most widely used approach, in which subjects are asked to read numerous self-referential statements, such as, “if your attitude is good, then things are good, and my attitude is good” or, “every now and then I feel so tired and gloomy that I’d rather just sit than do anything.” More recently, in part because of the ease of execution, film has become an increasingly popular means of mood induction. Further, an extensive review and evaluation of the reported effectiveness of the most prominent forms of mood induction found that film-based inductions were significantly more successful at inducing the target mood than any other approach individual approach (Westermann et al., 1996).

Ideally, there would be a published database of film clips of all lengths and genres that have been validated and shown to elicit a wide range of predictable affective responses, a la the International Affective Picture System for imagery (P. J. Lang, 1995). There have been efforts along these lines. Gross and Levenson (Gross & Levenson, 1995), evaluated 250 films and identified 16 that consistently and reliably elicited a predictable response across eight categories of emotion—amusement, anger,

contentment, disgust, fear, sadness, surprise, and neutral. The clips include scenes from a variety of genres of film and television, comedy routines, and documentaries.

Rottenburg, Ray, and Gross (Rottenberg, Ray, & Gross, 2007) expand on and validate this work, and go on to provide a series of recommendations for administering mood induction with film and a suggestion for a basic post-film questionnaire that can be used as either a form of manipulation check or in the selection of a new clip. What is missing from both of these analyses, however, is any discussion of what features of the clips themselves are linked with the various affective responses.

Lang (A. Lang, 1990) has worked along these lines, identifying key features of video that elicit emotional response, with a particular emphasis on the physiological response. Because of this emphasis on the physiological, generally measured in terms of Galvanic Skin Response and Heart Rate, most of the identified features are found to impact arousal. For example, this work has found that rapidly changing scenes and camera angles dramatically increases arousal, perhaps as a result of increased use of cognitive resources.

One can also look at the clips that have been used across a variety of studies to find common features. The most commonly used videos in mood induction are those identified by Gross and Levenson (Gross & Levenson, 1995). The videos intended to positive, high arousal emotions are all comedies, and comedy has been used successfully for this type of induction by other researchers as well (Hills, Hill, & Mamone, 2001). Videos intended to induce negative, high arousal emotions are from horror movies or are scenes in which someone is being abused or otherwise treated unfairly. Videos intended to elicit positive, low arousal emotions involve static nature scenes. Videos intended to elicit negative, low arousal emotions involve slow moving scenes involving the death of a loved one or animal.

## *Objectives*

At this point, it should be clear that the primary objective of this study is to demonstrate that when subjects are induced with affect, they will respond to PAM in such a way as to reflect the appropriate affective state. In looking to test PAM across a range of emotions for each of the subscales, the strategy was chosen to induce affect in three forms: 1) positive valence and high arousal, 2) neutral, and 3) negative valence and low arousal. This strategy should simultaneously test both the Valence and Arousal subscales of PAM. Further, this design should be a strong test of the Positive Affect subscale of PAM, given that these three inductions should produce a range from low Positive Affect in the negative, low arousal video to high Positive Affect in the positive, high arousal video. This design will be less of a test for the Negative Affect subscale. While differences should still be seen, the Negative Affect Scale is already of questionable utility as described in the last chapter.

## *Method*

This study was designed to assess PAM's capability of reporting changes in affect. The basic procedure was to induce positive, negative, or neutral mood through short video clips under the guise of a perception of Internet video study, then assess subjects using PAM.

### *Participants*

Participants ( $N=68$ , 33 Female, age 18-35) were recruited from public areas on a University campus and through departmental listservs. Participants were compensated through entry into a drawing to win an iPod. Five subjects were excluded from the

study because they indicated that the video had an emotional effect on them other than what was intended (e.g. they found the positive video to be unpleasant), reducing the effective sample size to  $N=63$ . Additionally, 15 individuals including students, co-workers, and friends were used in the piloting and selection of videos.

### *Video Selection*

The first step in this study was the selection of three videos—one to elicit low arousal, negative emotion, another to elicit high arousal, positive emotion, and a final clip to leave the subject in a neutral state. Ideally, the videos would have been drawn from a set of validated clips, such as those put forth by Gross and Levenson (Gross & Levenson, 1995). However, these clips are dated at this point, and many are lengthy. Further, because this study was to be conducted on computers and tablets, clips that were already available on Internet video services such as YouTube were required. In the end, members of the research lab were asked to search YouTube using positive and negative emotion terms as well as their own intuition to find clips that they found either highly negative, neutral, or highly positive. From this set of videos, the clips were then evaluated based on the criteria and examples described in the mood induction literature above. For example, videos for the negative condition would ideally include sad material, involve minimal cut scenes, use a single camera angle, and utilize appropriate music, if any. For the positive video, use of humor, action, multiple camera angles and many cut scenes, and energetic music would be ideal. Each of the final videos selected based on these criteria was pilot-tested by 15 undergraduates who verbally confirmed that the videos elicited the desired response.

For the positive clip, the official, extended trailer for the film *The Hangover 2* was selected (available at [youtube.com/embed/RYL\\_T7f59o8](https://youtube.com/embed/RYL_T7f59o8)). The sequel to a highly

successful comedy, the trailer is expected to resonate with and excite subjects recruited around a college campus. This clip shares many of the features that previous successful positive mood induction clips have employed, as described above, including comedy, high energy, and rapid scene changes. Based on this, the expected mood induction would be positive valence and high arousal. The clip is two minutes and 30 seconds long.

A video from BBC Science discussing Sir Isaac Newton's discovery of gravity was chosen as the neutral clip (available at [youtube.com/embed/D5BQkdyAw8A](https://www.youtube.com/embed/D5BQkdyAw8A)). The video is presented in classic documentary format, with monotonous narration, long still camera shots, and no emotional content. There is also precedence for using scientific documentaries for neutral videos, e.g. (Hills et al., 2001). It is expected emotion will not be induced by this clip, i.e. valence should be neutral and arousal should be neutral, or perhaps slightly low because of the slow-moving nature of the video. The clip is two minutes and 12 seconds long.

For the negative clip, a commercial featuring Sarah McLachlan advocating the SPCA was selected (available at [youtube.com/embed/9gspElv1yvc](https://www.youtube.com/embed/9gspElv1yvc)). This video depicts cute but ill or abandoned animals, uses very few scenes, and includes somber music from the artist. These features are consistent with those described above that are known to induce negative valence, low arousal emotions, and it is expected that this video will produce the same outcome. The clip is two minutes long.

### *Manipulations*

Participants were randomized into one of three conditions: Positive, Neutral, or Negative. Effectively, the only difference in each of the three conditions was the selection of video used in the inducement. Participants in the Positive condition ( $N=19$ )

were shown the positive clip—the trailer for the Hangover 2—intended to elicit positive, high arousal emotion (high Positive Affect). Participants in the Neutral condition ( $N=23$ ) were shown the neutral clip—a documentary about Sir Isaac Newton. Finally, participants in the Negative condition ( $N=21$ ) were shown the negative clip—the SPCA clip—intended to elicit negative, low arousal emotion (low Positive Affect). Participants in Negative condition were also shown the positive film clip at the conclusion of the study to counteract any potential negative effects of the negative clip.

### *Procedure*

Subjects were told that they were to participate in a study examining perceptions of Internet-based videos. Subjects were handed an iPad by a lab assistant and be instructed that they would watch an embedded YouTube video then answer a short series of questions about their perceptions of the clip. This deception was necessary to prevent demand characteristics from emerging related to the use of PAM. Following consent, the web application running on the iPad randomized the subject into one of the three conditions as described above. The experiment was conducted blind, with the lab assistant unaware of the video the subject was shown until after the fact. After the video was completed, subjects moved onto the next screen where they are assessed with PAM, given the prompt “Select the image that best represents how you feel *right now*.” After completing PAM, subjects answered a series of questions about the video. Two of the questions in particular, whether or not they have seen the video and “In one sentence, please explain how this video made you feel,” were used in analyses as a form of manipulation check. As indicated above, subjects who had been shown the negative clip were shown the positive clip to offset any impact of viewing the negative clip. Subjects were then debriefed and the lab assistant answered any questions that come

up.

### *Measures*

The primary measures are the four PAM subscales, Valence, Arousal, Positive Affect (PA), and Negative Affect (NA). As described earlier, Valence measures strictly the positivity or negativity of the emotion experienced and Arousal measures the level of energy or extent to which the emotion is felt. PA takes into account both Valence and Arousal to better describe the extent to which an individual is feeling positive emotion. NA takes into account both Valence and Arousal to describe the extent to which a subject is feeling negative emotion. Other measures were taken for use in analyses in the hopes of determining the effect the induction might have had on each subject. The actual duration of time the user spent watching the video was recorded; had subjects not completed the video their data would have been ignored, but this did not occur. The questions described above—has the subject seen the video and how did it make them feel—were recorded as well.

### *Results*

Analysis of variance (ANOVA) found that the effect of the video manipulation was significant for each of the four PAM scales, including Valence,  $F(2,60)=15.63$ ,  $p<.001$ ; Arousal,  $F(2,60)=14.79$ ,  $p<.001$ ; Positive Affect,  $F(2,60)=22.05$ ,  $p<.001$ ; and Negative Affect,  $F(2,60)=8.05$ ,  $p<0.005$ . Table 5.1 shows the means for each PAM scale by condition as well as the results of a post-hoc comparison using Tukey's HSD.

	<b>Video</b>		
	<b>Negative</b>	<b>Neutral</b>	<b>Positive</b>
<b>PAM Scale</b>			
Valence	-0.57 (1.34) <sup>a</sup>	0.48 (1.37) <sup>b</sup>	1.53 (0.77) <sup>c</sup>
Arousal	1.96 (0.77) <sup>a</sup>	2.24 (1.04) <sup>a</sup>	3.42 (0.90) <sup>b</sup>
PA	6.61 (3.92) <sup>a</sup>	9.43 (3.72) <sup>b</sup>	13.74 (2.45) <sup>c</sup>
NA	9.26 (3.77) <sup>a</sup>	7.00 (4.06) <sup>ab</sup>	4.74 (2.92) <sup>b</sup>

Table 5.1, Means of each PAM subscale for each video manipulation, reported as Mean (Standard Deviation). Means that have no superscript in common are significantly different from each other (Tukey's HSD,  $p < 0.05$ ).

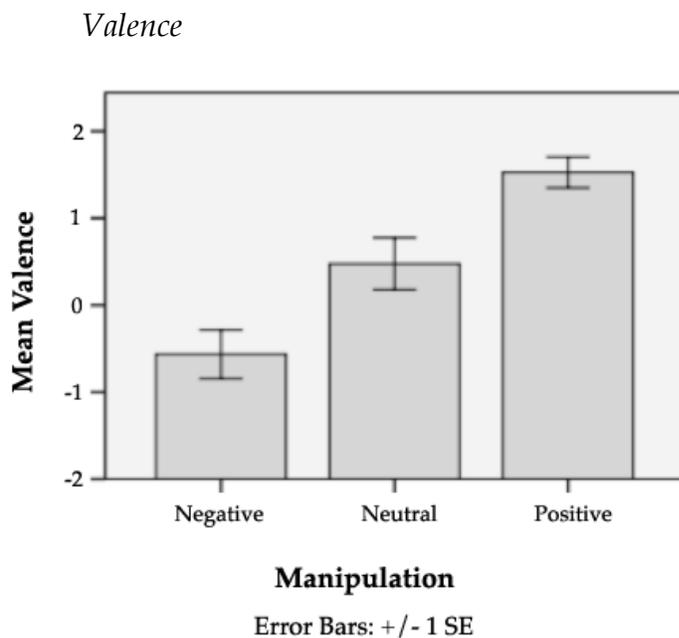


Figure 5.1 Plot of Mean Valence by condition.

Post-hoc testing using Tukey's HSD found that the effect of the video manipulation was significant across all levels for Valence,  $p < .05$ . As expected, subjects in the Negative video condition reported the lowest (most negative) valence,  $M = -0.57$

and subjects in the Positive video condition reported the highest (most positive) valence,  $M=1.53$ . Subjects in the Neutral condition reported very mild, positive valence affect on average,  $M=0.48$ . Valence is perhaps the most straightforward test of PAM in this study, i.e. does a more negative video result in a more negative response to PAM while a positive video results in a more positive response? One might expect the Valence scores to the negative video to be more negative for the Negative video condition, but this may have been a product of the video itself. Five of the subjects reported that the video either gave them hope for the animals or felt that the message was uplifting and more might have felt similarly if questioned specifically along these lines. Regardless, these results show strong evidence that PAM is responsive to changes in the valence of subjects' emotions.

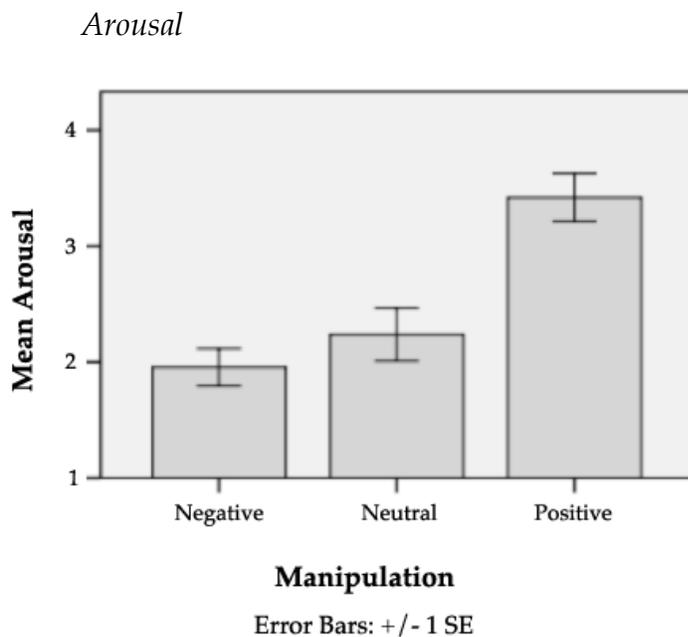


Figure 5.2 Mean Arousal by condition.

Subjects reported significantly higher Arousal after having been shown the

Positive video,  $M=3.42$  than those who were shown either the Negative,  $M=1.96$ , or Neutral,  $M=2.24$  video,  $p<.05$ . However, the difference in Arousal for those watching the Negative and Neutral videos was not significant. The higher Arousal found in the Positive video condition is exactly as expected. The Positive video contained significant energy, rapid scene changes, and humor, all elements of video known to increase arousal. Both the Negative and Neutral videos however were lower in energy, used fewer camera changes, and were fairly monotonous, resulting in similarly lower arousal scores. While less conclusive than the findings for Valence, these results do provide evidence that PAM is able to report changes in subjects' arousal.

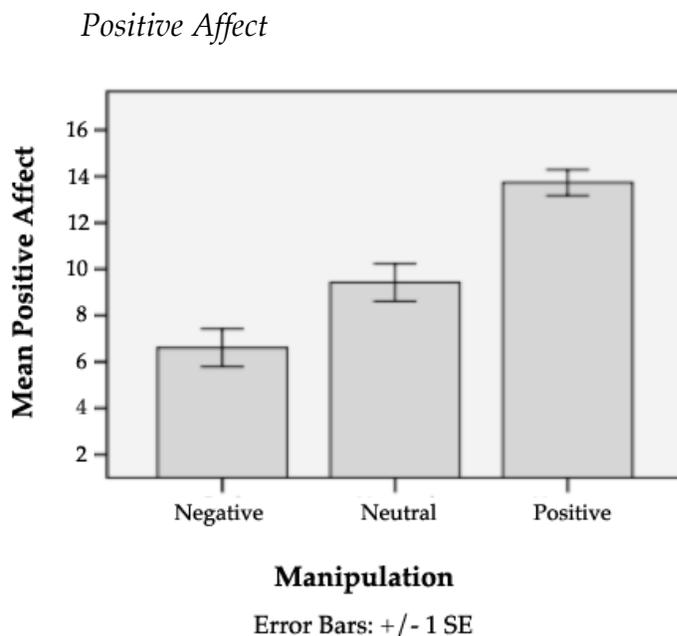


Figure 5.3 Mean Positive Affect by condition.

The effect of the video manipulation was significant across all levels for Positive Affect in the expected direction. Subjects in the Positive video condition exhibited the highest PA according to PAM,  $M=13.74$ , followed by the Neutral condition,  $M=9.43$ , then the Negative condition,  $M=6.61$ ,  $p<.05$ . Given that PA is a conflation of valence

and arousal, this is no surprise given the previous results for Valence and Arousal. The negativity combined with low arousal for the Negative video leads to low PA. The Positive video, which is both positive and high arousal, clearly induces high PA, while the Neutral video falls in the middle. These results are a strong indicator that PAM is capable of reporting changes in subjects' Positive Affect.

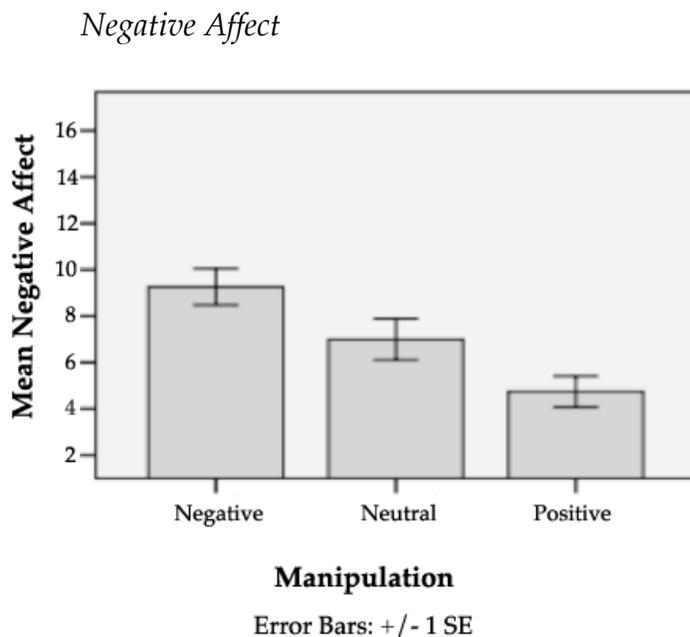


Figure 5.4 Mean Negative Affect by condition.

For Negative Affect, the effect of the video manipulation was significantly different for subjects in the Positive video condition and Negative video condition,  $p < .05$ , but neither differed significantly from the Neutral video condition. The highest NA was reported by subjects in the Negative video condition,  $M=9.30$ , followed by the Neutral video condition,  $M=7.00$ , and the lowest NA was reported by individuals in the Positive video condition,  $M=4.74$ . These results are less conclusive than those for PA, but they do nothing to suggest that PAM is not adequately recording NA. The

reason for this lies again in the fact that NA is a conflation of valence and arousal. Because the Negative video was by design low arousal, subjects who watched it should not exhibit high NA. Further considering that both the Neutral and Positive videos induced greater Arousal (albeit only significantly for the Positive video), the range of values for NA one would expect in this study is further compressed.

### *Discussion*

The primary objective of this study was to demonstrate that under experimental conditions, subjects' PAM scores are reflective of their actual affective state. To accomplish this, subjects were induced with positive, high arousal emotion, or neutral emotion, or negative, low arousal emotion and then assessed with PAM. The results of this study find that both PAM Valence and PAM Positive Affect differ significantly in the expected directions between the positive, neutral, and negative video conditions. This fully supports that the Valence and PA subscales of PAM are accurately reporting the affective states of subjects. PAM Arousal was significantly higher for the positive video condition than for the negative or neutral condition. Given the above discussion about the possibly low arousal-inducing nature of the neutral video, this finding provides strong support for the Arousal subscale of PAM as well. Finally, the PAM Negative Affect scale differed significantly between the positive and negative conditions as expected, but did neither differ from the neutral condition. In light of the fact that this study was intentionally designed to favor highlight differences in PA rather than NA, this is not a surprise, nor is it evidence against the validity of the NA scale.

Taken together, these findings provide compelling evidence that PAM is sufficiently capable of accurately measuring affect under experimental conditions.

Given this, researchers should feel comfortable using PAM in a number of experimental settings. Aside from the target uses already described to this point, the success of PAM in this setting indicates that it would be quite appropriate for use in a variety of experimental circumstances. Of course, the ability of PAM to differentiate and report both dimensions of affect as well as at least Positive Affect if not Negative Affect as well could make it quite useful in testing the emotional state of subjects in experiments. The brevity of PAM means that it could be assigned repeatedly at various stages of an experiment or make room for additional tasks or scales. Additionally, PAM is ideally suited for use as a manipulation check following mood induction, for example. Rather than asking subjects to spend a great deal of time on emotion assessment to determine the effectiveness of induction, the seconds required to complete PAM should suffice. Further uses of PAM will be discussed in the general discussion in the next chapter.

#### *YouTube Video for Mood Induction*

Also of note is that this may be among the first studies to rely entirely on YouTube for searching for, selecting, and embedding video as part of a mood induction. It would appear that this tactic was successful, in spite of the fact that it required considerably less effort than the watching of over 250 complete films documented by Gross and Levenson (Gross & Levenson, 1995). This could suggest that less rigor was used in the selection of these videos. On the other hand, it could be argued that because of the power to search and rapidly screen clips, this tactic is simply more effective.

By searching by emotion keywords (as described in Chapter 3 for the development of PAM itself), it could be argued that the videos that turn up in search have been communally determined to be 1) most relevant and 2) most popular and, hence, most engaging. Further, rather than watching entire films looking for key

scenes, YouTube, or more specifically other YouTube users, have done the difficult work already by selecting the most interesting material from film, television, and even home video, uploading it, and often describing it.

As just noted, portions of this process were similar to that of the development of PAM itself. It is possible that future research could take this one step further and employ the exact same methods used in the development of PAM and use them to create a new, very large corpus of YouTube videos suitable for inducing a wider variety of emotional states. In other words, search terms could be used to identify and categorize videos, and the number of views a video has received could be used as a sorting algorithm to bring the most popular videos to the top. Next subjects could be recruited to view videos and subsequently take PAM (or use Semantic Differential, or any other scale, for that matter) to validate and identify the specific emotional response elicited by each. Aside from the need to recruit large numbers of subjects, this may be an extremely simple process capable of generating a large, validated database of affect inducing video.

### *Limitations*

The primary limitation of this study is that only three conditions were tested. It is possible that greater clarity could be gained from testing each of the four quadrants of affect independently. In other words, future experimentation could induce each of the four combinations of positive and negative valence and high and low arousal, adding conditions for negative, high arousal emotion and positive, low arousal emotion. While both the valence and arousal dimensions of affect were varied with positive results, and valence and arousal are known to be independent (Russell, 1980), it is possible for scales to have more difficulty with certain combinations. For example, research has found

PANAS to have difficulty differentiating between positive and negatively valence emotions in circumstances where subjects are experiencing extremely high arousal (E. Harmon-Jones et al., 2009).

### *Conclusions*

As a follow-up the previous study demonstrating that PAM scores converge with scores from other validated measures of affect, this study aimed to demonstrate that PAM scores reflect the expected affective state of subjects in experimental conditions. In this study, a film-based mood induction was used to place subjects into either a positive, high arousal emotional state (high Positive Affect), or negative, low arousal emotional state (low Positive Affect), or a neutral state, at which point they were assessed with PAM. The results clearly found PAM scores to reflect the predicted affective states in each case. These findings are a strong endorsement for PAM, confirming it's viability and validity for use in experimental conditions.

## CHAPTER 6

### GENERAL DISCUSSION

This dissertation has presented the rationale, design, development, and validation of a brief, digital measure of affect. PAM, the Photographic Affect Meter, asks users to choose from a selection of emotionally charged and validated images the one that best represents their current emotional state. Based on the choice, PAM is able to report the affective state of the user in a manner consistent with the most widely used scales of affect. All of this can be accomplished in a matter of seconds with one to two clicks. As a result, PAM has utility wherever methodology calls for abbreviated, frequent, or timely measurement of affect and is appropriate for use across a variety of fields.

Given that emotion is generally believed to impact nearly every aspect of human behavior (Baumeister et al., 2007; R. S. Lazarus & B. N. Lazarus, 1994; Pressman & Cohen, 2005; Slovic et al., 2007), it is widely studied in the behavioral sciences (Bradley & P. J. Lang, 1994; P. J. Lang, 1995; Morris et al., 2010). For health-related behavior in particular, the ramifications of emotion-influenced decisions can be substantial. In a landmark review of the subject, Pressman and Cohen (Pressman & Cohen, 2005) document the important impact that affect can have on health, but also note that the complexity of the linkage between emotion and health-related behavior calls for improved measurement and assessment of emotion in health contexts. The field of Ecological Momentary Assessment (L. Barrett & D. Barrett, 2001) and an increasing prevalence of mobile phones have given researchers new methods that can begin to address the complexity of this linkage, but new tools for measuring affect are needed as well.

The overall objective of this work has been to produce a measure of affect that (1)

reliably measures affect, (2) is unobtrusive and pleasant enough to administer at least daily, and (3) can be administered in situ. While this has certainly been the subject of previous work, the novelty of the approach taken in this dissertation sets this work apart. Rather than approach the problem from a purely psychological perspective, this work set out to use knowledge and experience in affective computing to create a better interface for the measurement of affect. Further, this approach uncovered a new method for developing libraries of topical and encoded imagery.

Chapter 3 detailed the design and development process used in the construction of PAM. As a first step, thousands of images were downloaded from Flickr by using the Flickr API to automate a search for images tagged with emotion words. The resulting library consisted of images that, in theory, were identified as being associated with a certain emotion, and that association was maintained as metadata. In the next series of steps, subjects were employed to iteratively distill the library to smaller and smaller, but also more emotionally meaningful subsets. Eventually, 48 images remained—those that were quantitatively most representative of the ranges of valence and arousal intended to be measured. It is important to note that each step in the process, users/ subjects were responsible for selecting and assigning meaning to the various image rather than the designers of the system. This process was proven to be highly effective both in developing a library of emotionally charged imagery and in assigning emotional meaning to photographs.

As detailed in Chapter 4, the second phase of this work was to validate that results from PAM are consistent with those produced by other generally accepted measures of affect. Subjects were asked to nearly simultaneously rate their current emotional state with PAM in conjunction with either PANAS (Watson et al., 1988), Affect Grid (Russell et al., 1989), or the Semantic Assessment Manikin, SAM (Bradley &

P. J. Lang, 1994). Strong correlations were found between the outputs of PAM with each of the other scales. In other words, results from PAM were consistent with results from each of the widely used, generally accepted measures of affect, sufficient to establish convergent validity (Campbell & Fiske, 1959).

Finally, in Chapter 5, it was demonstrated that when the emotional states of subjects are manipulated, PAM results vary consistently with the manipulation. A classic film-based mood induction was used to place subjects into a negative, low arousal state, a neutral state, or a positive, high arousal state. When assessed with PAM immediately following the mood induction, individuals in the negative condition expressed the most negative valence, lowest arousal, lowest positive affect, and highest negative affect. Conversely, individuals in the positive condition expressed the most positive valence, highest arousal, highest positive affect, and lowest negative affect. Individuals viewing the neutral video reported middling scores in each of the dimensions of PAM. These results were completely consistent with expectations, demonstrating that PAM is indeed effective at detecting expected differences in affective states in an experimental setting.

These findings make important theoretical and practical contributions. First, to the affective computing community, this work is among the first to draw from the subjective, ambiguous nature of emotion in the design and development of a system that is decidedly objective and quantitative. This may serve as inspiration for future efforts along these lines. Second, this work has several implications for research in the measurement of affect—both methodological and in regards to the relationship between photographs and emotion. Finally, the most obvious contribution of this work is practical; it provides a new means for measuring affect in a wide variety of applications and domains. Researchers should find utility in PAM in any situation where brevity,

timeliness, and context are requirements when measuring affect, as it can easily be implemented on any computing platform, including mobile phones. Outside of the research community, PAM has possible applications in a wide variety applications ranging from marketing, to collecting user feedback, to recommender systems. The following sections walk through the key contributions to each field.

### *Affective Computing*

In the affective computing literature, a line has been drawn between more traditional work that views emotion as objective and quantifiable and work that views emotion as subjective and unquantifiable. The traditional view, originally put forth by Picard (Picard, 1997), posits that computers should be capable of measuring, emulating, and ultimately effectively communicating emotion with humans. Measurement tactics in affective computing have ranged from using linguistic markers (M. A. Cohn et al., 2004), to physiological markers (Nasoz et al., 2004), to self-report (Morris et al., 2010). Each of these tactics—and the systems that employ them—produce some quantitative description of a users affective state, generally in terms of valence, arousal, or specific emotional states such as anger or happiness. These works rely on two fundamental assumptions—first, that emotion is a construct that can be quantified and second, that computers are capable of carrying out the measurement. The opposing viewpoint takes issue with each of these assumptions.

Boehner et al (Boehner et al., 2005) argue that emotion is not something which should—let alone can—be measured. Essentially, the view is that context, culture, history, and sociality play such a strong role in emotion and the emotional experience and to measure it is to strip away each of these. Sundstrom et al (Sundström et al., 2007) agree, and in their mobile system eMoto simply provide a color-based construct in

which users can negotiate their own emotional meaning. Sengers (Sengers et al., 2008) takes issue with the second assumption above, arguing that computers are incapable of understanding the range of human emotion. In her experimental system *Affector*, affect is represented abstractly and arbitrarily, allowing users to assign any and all meaning to what they see in the system. In spite of this, Boehner et al (Boehner et al., 2007) ultimately concede a compromise:

Although our approach resonates with [the subjective] view, we argue that the way forward for affective computing and affective evaluation is not a debunking of objective approaches in total but a recognition of the limits and liabilities of both objective and subjective accounts of emotion (Boehner et al, 2007, p289).

For the objective approach, the limits and liabilities that the authors speak of are quite clear from the arguments leveled above. The complexity and context-driven nature of human emotion make it exceedingly difficult to isolate and measure. As such, any such measurement can only be considered an approximation and part of the complete story. Further, such measures often rely on arbitrary or biased decisions about how to represent and assess affect. Proponents of the subjective view would argue that this calls into question research that relies on such measurement. However, large segments of academic and commercial work rely on such measures, and this is where the limitations of the subjective view become apparent. Because the subjective view wholly discriminates against the measurement of affect, it offers little value to those working in quantitative fields that wish to assess affect.

This work in this dissertation takes an approach designed to borrow from the strengths and mitigate the limitations of both the objective and subjective view.

Specifically, the subjective view is adopted during the design of the system and objectivity is introduced by adding and rigorously validating the quantitative scoring used by the scale, and the findings lend credence to this approach. The subjective view was employed in the design of the system, by (1) using a representation of emotion that offered opportunity for interpretation and personal meaning and (2) allowing users (rather than the designers or the system) to determine the meaning of each representation of emotion. Both elements of this speak to strengths of a subjective view of emotion, but also to potentially eliminating some of the weaknesses of more typical objective approaches. The former, using a more interpretively flexible representation of emotion might allow for a broader range of emotional experience to be captured. The latter, allowing users to determine the meaning of each representation of emotion should do well to remove the bias of the designer from the scale.

The results clearly bore this out. Aside from the selection of photographs as the medium through which emotion would be represented and the overall format of the scale, the designer played little role. The meaning assigned to each photo was arrived at socially by hundreds of users. While die-hard supporters of the subjective view might take odds with the subsequent assignment of a quantitative output for each photo, they would certainly endorse the approach to this point as well as the notion of socially constructed meaning. And, while proponents of the objective view might not have seen the value of the design phase, it is difficult to argue with the findings that clearly demonstrate the viability of the scale as an objective measure of affect. In other words, it almost doesn't matter what happens in the design of the scale, how the photos were selected, or how the designer or even an expert in emotion might interpret the individual photos or their layout because the empirical evidence clearly shows that PAM produces expected output.

Framing this work as an intersection of these two diametrically opposing view points is not without risk, as it opens the door for criticism from both camps. However, in the spirit of the comments by Boehner et al (Boehner et al., 2007) above, this work has simply recognized the strengths and weaknesses of each approach and worked with those in the hopes of creating the best possible scale. Critics from the subjective view should recognize, as Boehner did, that in certain domains the objective approach is warranted and take solace in the fact that this approach embraces interpretive flexibility and social construction of emotional meaning. Critics from the objective view might criticize the selection of photos, the process used to collect them, or even the use of photos as a representation of emotion, but again, they need look no further than the results of the validation of the system.

### *Emotion, Photos, and Methodology*

This work has several implications for researchers in emotion, particularly those interested in the measurement of emotion. First, it takes the existing work linking photographs and imagery with emotion another logical step forward. Not only does it add further support for this theoretical linkage, but also provides evidence that images can be employed not only in the induction of affect but in the assessment of it. Further, the development process presented in this dissertation could serve as a template for future developers of new scales and assessment tools.

Chalfen (Chalfen, 1987) has described photographs as a means of communicating experiences by triggering emotional responses and memories. He goes on to describe a “Kodak Culture” in which photographs are among the most important artifacts for sharing experiences and feelings. Sondhi et al (Sondhi & Sloane, 2007) have further studied this in photo sharing among family and friends, finding that much emotional meaning is

conveyed in the act of sharing photographs. These works imply that there are two sides to the emotional story of photographs—the ability of the photo encode and convey emotional meaning as well as the ability to elicit an emotional response.

Lang (P. J. Lang, 1995) has made the most explicit link between photographs and emotions to date in the development and validation of the International Affective Picture System, IAPS, a massive library of emotionally charged imagery. Each photo in IAPS is empirically linked to the elicitation of a specific emotional state through extensive physiological testing, providing very concrete evidence. The second important component of Lang's work with IAPS is that it provides further evidence that emotional meaning in photographs is shared across many individuals. However, Lang's approach provides evidence that the photos *elicit* a similar emotional response, not necessarily that there is shared legibility. In the development of the Self Assessment Manikin, Bradley and Lang (Bradley & P. J. Lang, 1994) do use images from IAPS as reference points, asking subjects to rate each of the images using different scales for the purpose of validation. This comes closer to exploring the legibility of the emotional content of the photographs, but it ultimately wasn't the intention of the work and was not raised in the discussion.

Mayer et al (J. Mayer et al., 1990) get much closer to the notion of shared emotional legibility in imagery with an experimental approach, finding that visual stimuli such as photos do convey similar emotional content across large numbers of individuals. However, this work does not focus entirely on photos nor does it go to the lengths of the IAPS work in linking imagery with specific emotions. From this perspective, the work in this dissertation could be considered the next logical step in the line of work begun by Mayer, examining shared emotional meaning by employing larger numbers of subjects and focusing specifically on the emotional state represented

by each photo. Ultimately, the findings presented here do offer evidence of that next step—photos do indeed appear to carry enough shared emotional legibility that they can consistently and effectively be selected based on their emotional content. In short, the findings demonstrate that photos can not only be used as a means of eliciting an expected response, but that they can be used to do the opposite—to measure an existing state or response to other stimuli.

The methods employed to arrive at these conclusions also represent a significant contribution to the field. From start to finish, the process used in this work leveraged social and user-generated data to arrive at the artifacts used in the scale. The first step of this process used a combination of tags assigned to images and the socially constructed “interestingness” of photos on Flickr in an automated search process generated an initial corpus of images. This approach has several advantages. First, as discussed above, the fact that the images are socially tagged with emotion words and deemed interesting makes them much more likely to contain emotional content than starting with an arbitrary collection of photos. Second, as described above from the affective computing perspective, the photos are much more likely to be free from any biases of the designer of the scale. Finally, and more practically, the automated search process and availability of tens of thousands of freely available images is dramatically easier than alternative means of gathering and filtering photos. This process has broad applicability and could potentially be used to gather many other types of emotionally charged digital media. In addition to photos, music, video, and stories (all media used in mood induction, for example (Westermann et al., 1996)) are readily available and searchable on the Internet with all of the attending socially constructed meta-data needed to utilize the above approach.

The second step of the process involved using the images for a seemingly

unrelated task—an emotion sharing application—to distill the corpus into a body of photos that would be more interesting, emotionally charged, or emotionally legible. In the application, subjects shared images with one another that they felt represented how they were currently feeling. (Gay et al., 2011). While this specific application may not be transferrable to other development processes, the notion of using a social task to refine a corpus of emotional content could be employed in other ways. For example, researchers looking to distill a collection of videos gathered from YouTube could ask groups of individuals to nominate on their top ten happiest or saddest (or funniest, scariest, and so on) videos of all time.

The final step in the process was the rigorous quantitative evaluation of the photos and the emotional state that each represents. This part of the process is not substantially different than that employed in the other scales discussed to this point and likely no different than the process used by most doing work of this kind. The one point worth noting is that in the work presented here, the approach was iterative, repeating essentially the same quantitative evaluation over and over again, refining and improving the image set and scale with each iteration. Iterative design processes, e.g. (Nielsen, 1993), have been used in Human-Computer Interaction research and commercial applications for years and can significantly improve the utility and quality of systems.

### *PAM: A Novel Measure of Affect*

Of course the most significant contribution of this work is the practical offering of a novel measure of affect. As prescribed by the objectives of this work and upheld by the findings, PAM requires only seconds and one to two clicks to complete and is validated to output results that match expectations and converge with a selection of the

most widely used measures of affect. Furthermore, PAM is easily embedded or used as a standalone application on mobile devices, web applications, or desktop computers. While PAM is not intended as a substitute for PANAS or other more substantial measures of affect, this combination of features means that PAM can have significant utility in a broad array of contexts.

While there is evidence that shorter scales are not always necessarily worse than more complete ones (Burisch, 1984), PAM is designed for circumstances that necessitate a more efficient or portable instrument, particularly when assessment would be best carried out on a mobile device or computer. When possible, a most prudent approach would be to combine frequent measurement with PAM to capture variability around daily events or key happenings with a more substantial assessment of affect at the beginning and/ or end of the study. Of course in other settings where multiple forms of measurement are not possible but brevity is required, researchers should feel comfortable using PAM based on the evidence presented herein.

Two general cases comprised of classic examples of Ecological Momentary Assessment make up the ideal usage scenarios for PAM. First, PAM could be used in cases where researchers need to know about subjects' affective states during certain events or decision making processes, such as linking affect to dietary choices (Macht, 2008), exercise behavior (Charlson et al., 2007), lapses in substance abuse (Epstein et al., 2009), or around traumatic events (M. A. Cohn et al., 2004). Second, PAM would be a logical tool to incorporate into any experience sampling or other similar scenarios, such as examining the variability of affect over time (L. Barrett & D. Barrett, 2001; Csikszentmihalyi & Hunter, 2003), around use of an app or system (Isomursu et al., 2007), or around scheduled happenings like key political addresses or the FIFA World Cup. The majority of examples that fall into these two categories emphasize the

importance of measuring affect in situ. As such, they beg for the use of mobile devices, particularly a user's primary mobile phone that would be with them at all times and on hand whenever the key moment of assessment might take place.

Researchers and developers of mobile and web applications should be particularly interested in PAM. Android, iPhone, and web-based PAM widgets already exist and will be made available to researchers at <http://cornellhci.org/pam>. If these widgets are incorporated into existing applications as an additional single-click step of login, posting, or other common workflow processes, developers can immediately begin collecting contextually anchored data about the affective states of their users. PAM offers substantial value to researchers developing behavioral interventions that want to collect affect data for their analysis as highlighted by the VERA example above. Developers building experience sampling, diary, reflective, or other creative apps in which affect might play a role might also find value in a simple measure of affect. Even developers simply looking to better understand their users might find some value in the ability to quickly and easily measure their users' affective state, perhaps in response to various stimuli either in the app or their surroundings.

The use of PAM need not be restricted to mobile applications or EMA-like situations. Researchers conducting laboratory experiments might likewise benefit, particularly in cases where brevity or unobtrusiveness is favored. For example, PAM could serve as an ideal instrument for carrying out manipulation checks in studies relying on mood induction. Westermann et al (Westermann et al., 1996) argue that the use of traditional rating scales and self-report measures can lead to subjects guessing the true intention of a study, thereby potentially resulting in demand characteristics. Instead, they suggest the use of behavioral or physiological measures. However, Larsen and Sinnett (Larsen & Sinnett, 1991) found that effect sizes for mood inductions are

more pronounced and more easily understood through self-report. Perhaps PAM strikes a good balance between the two—brief, unobtrusive, and less obvious than the use of PANAS or other measures, but still fundamentally a self-report measure.

The brevity of PAM could also be of use in laboratory experiments working with social media or Computer-Mediated Communication (CMC)—arenas in which individuals' tasks are typically rapid and abbreviated. In these domains, the administration of typical measures of affect are likely to be extremely obtrusive—especially given that they take longer than the behaviors being studied, such as tweeting, posting on Facebook, or communicating via an instant messaging service. In response to this, researchers examining emotion in these spaces have often looked instead to behavioral indicators of affect. This has certainly been the case in work examining emotional contagion in CMC, where linguistic markers have successfully been employed to uncover evidence of contagion in online chat (Guillory et al., 2011; Hancock, Gee, Ciaccio, & Lin, 2008). PAM could be a benefit to work such as this, offering the ability to collect self-reported rather than inferential measures of affect without the need to administer lengthy and obtrusive measures. Further, the digital nature of PAM lends to seamless integration into study protocol and data collection in this type of work.

PAM is also a potential improvement on existing measures in situations where repeated measurement of affect is desired. Other measures of affect might suffer from something akin to a practice effect as subjects become bored or complacent with the same, possibly lengthy scale. This has led to a fairly common approach of using two or more separate scales to measure emotion at various points in lab studies. In PAM, given the combinatorial possibilities of randomly choosing from three images at each of 16 possible locations, there are  $3^{16}$ , or just over 43 million possible versions. In other

words, the probability of a subject being presented with an identical instance of PAM twice in a row is infinitesimally small. However, because there is reasonable likelihood of a subject seeing at least some of the same images in the scale again, it might be worth exploring a version of PAM which ensures the generation of an entirely grid for each of three uses for such cases. Further, this could be a case for the validation and inclusion of an expanded set of images. Regardless, the fact that PAM is substantively different each use offers could potentially alleviate some of the above problems.

The form factor of PAM offers several potential advantages over other measures that may be of use in certain research settings. For example, unlike the majority of affective measures, PAM does not rely on literacy or comprehension of any written language. This is particularly relevant to researchers involved in clinical trials, as the NIH and other foundations have put significant effort into funding work that targets or at the very least proportionally includes disparate populations. Reduced literacy rates have been a barrier to recruitment and compliance in clinical trials in the past, and some research has cited the need for inclusion of methods and measures that don't rely on reading or language-specific comprehension (Swanson & Ward, 1995). Also, given that PAM can be scaled for display on tablets or even larger screens and only requires a single click or tap of an image to complete, it could be of value in circumstances where vision or motor impairment is an issue, such as with the elderly or injured.

Finally, the use of PAM outside of academia should not be overlooked. While not the emphasis of this work, there are a few notable examples in which brief assessment of emotion could be carried out with PAM that could be of value. Market research and testing the feasibility of designs and new products is a classic example. PAM could easily be built into these web-based systems for this kind of work, or tablets or mobile devices could be given to testers and raters who are assessing new products.

Netflix or similar media providers wishing to connect with and better understand their customers could use PAM to assess the way people feel as they are making movie choices and in response to certain films, and incorporate such information into their recommender systems. As a last example, much like Myers-Briggs and other psychometrics are used to evaluate employees, PAM could easily be incorporated into office intranets and login systems to collect regular data about employee emotional state.

### *Limitations and Future Work*

The most significant overarching limitation of this work is related to the fact that PAM is a self-report measure. Aside from the obvious issues shared by nearly all self-report measures that won't be discussed here, the challenge with using PAM in certain studies will be ensuring that subjects are in fact reporting the right thing. In other words, one can be fairly confident that when a subject is assessed with PAM on its own with the prompt of "how are you feeling right now" that they are going to report on just that. But, for example, if PAM is combined with the reporting of other behaviors, subjects might be reporting on how they feel at that instant, how they felt when they conducted the behavior, how the behavior made them feel, or how they felt about the behavior. In different contexts, each of these responses might be perfectly appropriate, but it is important that subjects do this consistently across the study and that the researchers clearly understand what it is that subjects are reporting. Explicit instructions, obvious prompts, and re-evaluating this in post-tests (i.e. "when you completed the PAM assessment, how did you typically rate how you felt...") are suggested.

A primary limitation of the validation process is related to generating a

representative sample. In the validation component of this study, the sample was largely from a college campus or acquaintances of people associated with that campus. While the self-reported demographics of the participants showed quite a diverse sample, the population was still English speaking. The importance of representation in this sample is an interesting question however. On one hand, research surrounding Lang's International Affective Picture System suggests that many images contain universal emotional legibility that transcends demographic boundaries (P. J. Lang, 1995). On the other hand, research into interpretation of photography has found the opposite to be true; interpretation of imagery can be somewhat more culture and context dependent (Chalfen, 1987). For example, this could be an issue for use in the subset of clinical trials examining health disparities (a subset which is increasing in size under federal mandate) where populations are necessarily comprised of lower SES or under-represented minority groups. On the surface, this is an easy problem to solve—by simply selecting and validating a new set of images for each target culture—but the likelihood is that without more extensive research unforeseen problems might arise.

Along these lines, over time the set of images used by PAM should be expanded—whether to incorporate imagery that is more appropriate cross-culturally or simply to increase the number of available photos—provided that the same level of reliability can be maintained. The above discussion about the possible advantage of the use of PAM as a repeated measure has already highlighted a key benefit to expanding the number of photos. Another potential benefit of this has to do with the “More Photos” button in PAM and the approach that subjects can potentially refresh PAM over and over again until they find an image that best suits their mood. It's possible, perhaps even likely that certain subjects relate better to different types of imagery, i.e. human faces versus animals versus nature scenes (Gay et al., 2011). Increasing the

number of options for the subject could increase the likelihood of them selecting the most representative image.

The use of tagged photos on Flickr as the original source for PAM images presents not as much a limitation as an opportunity for future exploration in other sources. On Flickr, the density of tags and keywords placed on photos by the community is not particularly high compared to other sources. This does not raise issues for PAM given that the final set of images has ultimately gone through an extensive, multi-step selection and validation process, but it does beg the question of where else images might be found. Many photos posted on Facebook, for example, have been commented on many times, and these comments could easily be assessed using LIWC or simple word matching to identify emotional content. Of course, the use of Facebook photos raises a number of privacy and licensing issues that simply do not exist for free-to-use Creative Commons licensed photos like those used in PAM.

The tags themselves present another topic of discussion. Because of the way the images were derived, each is still associated with an emotion word tag. In other words, each image is actually coded with more information than a simple Valence/ Arousal/ PA/ NA score that is derived from its position in the grid. Further, there may be features or elements of the photos themselves that contain valuable information. A logical next step in this work might be to begin to examine these aspects of PAM. Perhaps a selection of certain images can indicate fear while another could indicate boredom. Or perhaps the selection of a face as opposed to a nature scene reveals something about the emotional state or even personality of the subject. Any of these possibilities could provide potentially useful data to the researcher employing PAM. This would of course require a substantial investment of time and resources to conduct this validation, given the need to evaluate each of the 48 images present in PAM.

Another possibility for future work with PAM is the development of other similar scales using the same method. The work discussed in Chapter 4 has raised a point about the possible need for better assessment of Negative Affect. One answer to this issue could be the development of two separate PAM scales, one for Positive Affect and one for Negative Affect. Posi-PAM would focus on the range of positive affect, differentiating perhaps not only on level of arousal but on individual emotions within the positive affect space such as happy versus excited or calm versus sleepy. Nega-PAM would of course focus on the range of negative affect and could potentially differentiate between such emotions as frustration versus anger or bored versus sad. In either case, it's plausible that Posi-PAM and Nega-PAM would individually be much more reliable measures of Positive Affect and Negative Affect respectively, given what is known about the weak association between the two (Watson et al., 1988; 1999).

Taking this notion one step further, it may be possible to tailor PAM to many specific environments and populations. Rather than presenting multiple scales, researchers could select an appropriate version of PAM for use in their population based on the expected affective experience of subjects. For example, a cancer patient population would likely be experiencing substantially more negative affect, and as such a PAM scale that has been optimized for a higher range of Negative Affect would be most appropriate. This version of PAM would be developed in a target population experiencing a significantly greater amount of Negative Affect than the sample used to generate the current version. Adopting such a strategy could in the end result in a flexible or even dynamic measure of affect that would be appropriate in many more situations than the current version or even the other measures discussed in this work.

Finally, emotion is not the only construct that could be measured using methods similar to those used in the development and administration of PAM. Hypothetically,

any concept that can be assessed with self-report and possibly mapped to imagery could be tested using these methods. Obvious choices include health-related measures such as stress, pain, or a wide variety of symptoms, both psychological and physiological. Take stress, for example, which is largely a one-dimensional construct. Flickr could be searched for images using keywords such as “stress”, or possibly words such as “nervous”, “irritated”, “problems”, “control”, or “overcome” pulled directly from a validated measure of stress such as Cohen’s Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983). The same iterative process could then be employed, repeatedly asking large numbers of subjects to select an image that best represents how they feel and at the same time assessing them with a validated measure of stress.

The discussion in this section has identified a number of limitations and possibilities for future work. The two next steps that will be taken with PAM will take place simultaneously. First a much larger population of users will go through a variation of the PAM plus other-accepted-measure validation process, with the goal being to identify more specifically what can be learned from the selection of each individual PAM photo. This will certainly help to improve the reliability of PAM if problem or rarely selected photos are identified and can be replaced, but it will also speak to the above discussion on what else might be encoded in the imagery that would be of value.

The second logical next step for PAM is simply to begin deploying it in the normal course of research and evaluating the effectiveness. Ideally these studies would represent a range of use cases. At the time of writing, PAM has been deployed in numerous mobile phone-based health behavior awareness studies using the application VERA (Baumer et al., n.d.), a study using mobile phones for momentary stress and affect assessment in medical residents at a major US teaching hospital, and an NIH

funded randomized clinical trial examining wait management in pregnant mothers. PAM will also be deployed in an NIH funded clinical trial examining weight loss and healthy eating in low-income populations and a health behavior study at a major US technology employer by the end of 2011. As this wealth of data comes in, the validity of PAM can be constantly revisited based on findings and data.

### *Conclusion*

This dissertation has presented the rationale, design, development, and validation of PAM, the Photographic Affect Meter. PAM is intended as a brief and reliable digital measure of affect that can be used in a wide variety of circumstances where timeliness, unobtrusiveness, and context are important to the researcher. Unlike other measures of affect, PAM was built from the ground up as a digital tool, incorporating lessons from the fields of Affective Computing, Design, and Ecological Momentary Assessment into the design process. The result is an elegant solution whose form and function are reflective of this design process. Ultimately, this design process was followed with rigorous evaluation and validation demonstrating that PAM is indeed as reliable and valid as many of the most commonly used measures of affect.

The success of the approach used in this work has implications for a number of areas. For researchers in Affective Computing, this work presents evidence that embracing both the objective/ quantitative view and the subjective/ qualitative view of emotion might best approach certain problems. To Emotion research, this work contributes further support to the notion that photos and emotions can be explicitly linked, arriving at similar conclusions to previous work but from a different approach. Lastly, this work introduces a new method for leveraging the social Internet to generate corpora of affective imagery, and this approach could potentially be co-opted to

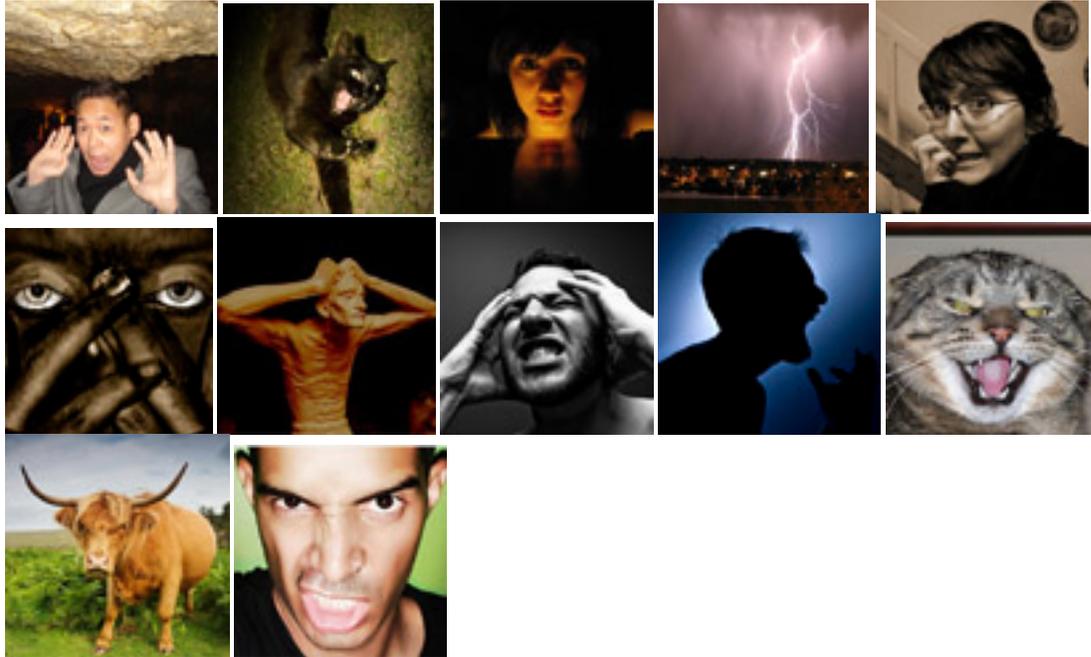
generate databases of media for a wide variety of purposes.

The main contribution of this work is of course the presentation and offering of PAM as a novel measure of affect, validated and ready to be used on a number of platforms including the web and mobile phones. The potential value of PAM is substantial when considering the confluence of the importance of measuring affect in research in the behavioral sciences, an increasing awareness of the importance of Ecological Momentary Assessment, and the rapidly growing prevalence of mobile phones. Taking the next steps toward developing PAM into a more flexible scale that can be tailored for a wide variety of different populations will only increase its potential utility. Ultimately, it will be the uptake and inclusion of PAM in future lines of research that will determine its overall success, validity, and contribution.

# APPENDICES

## Photographs Included in PAM

### High Arousal/ Negative Valence:

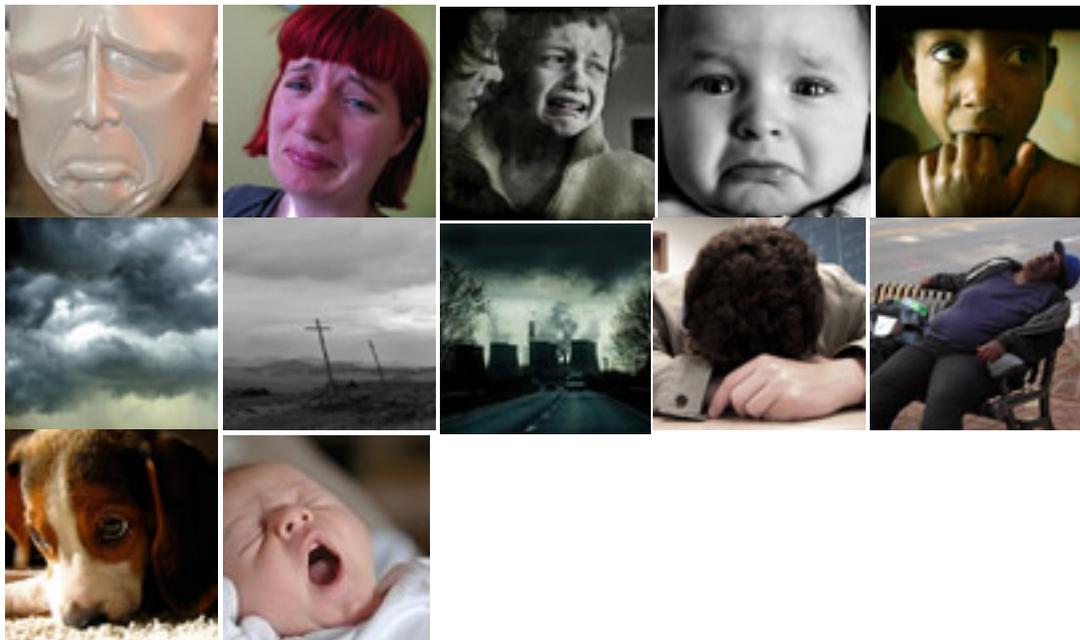


### High Arousal/ Positive Valence

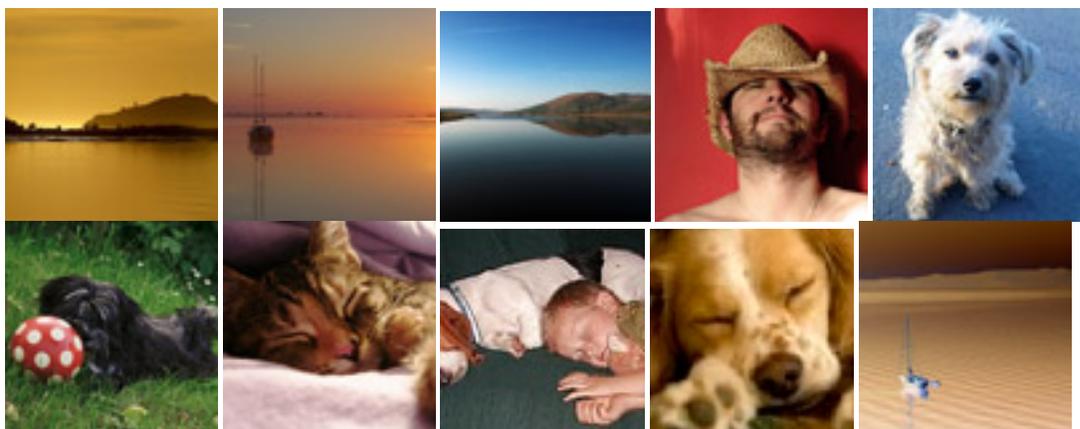




Low Arousal/ Negative Valence



Low Arousal/ Positive Valence





## PANAS: The Positive and Negative Affect Scale

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you feel each word at this moment, as in right now. Use the following scale to record your answers.

1 - very slightly or not at all   2 - a little   3 - moderately   4 - quite a bit   5 - extremely

1. interested
2. distressed
3. excited
4. upset
5. strong
6. guilty
7. scared
8. hostile
9. enthusiastic
10. proud
11. irritable
12. alert
13. ashamed
14. inspired
15. nervous
16. determined

17. attentive

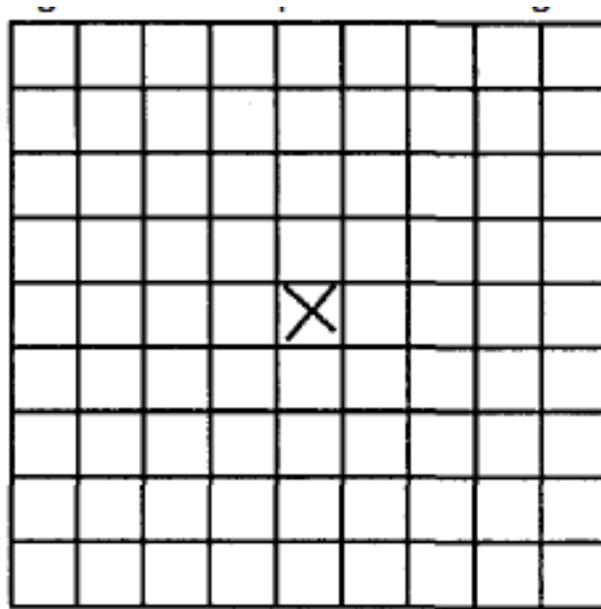
18. jittery

19. active

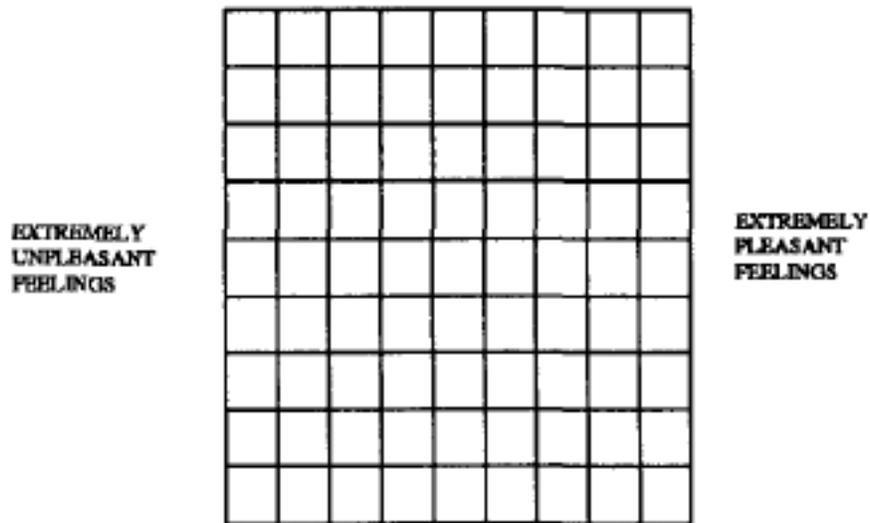
20. afraid

## Russell's Affect Grid

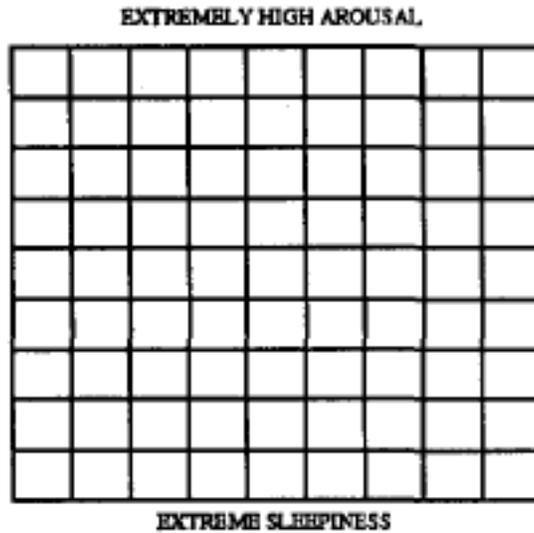
You use the "affect grid" to describe feelings. It is in the form of a square - a kind of map for feelings. The center of the square (marked by X in the grid below) represents a neutral, average, everyday feeling. It is neither positive nor negative.



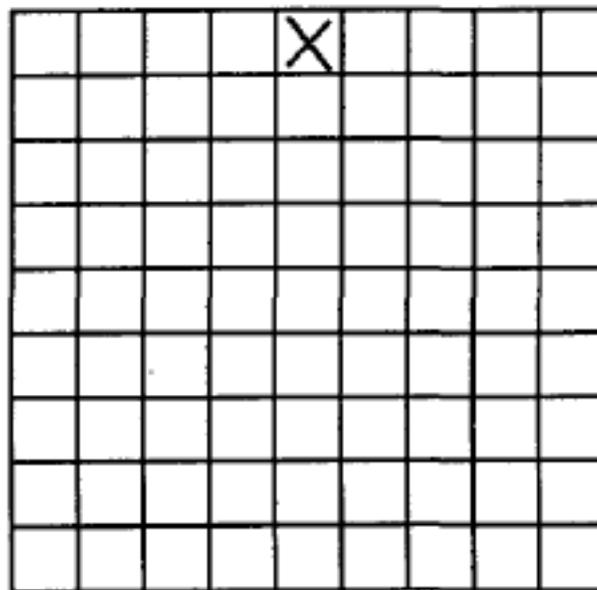
The right half of the grid represents pleasant feelings. The farther to the right the more pleasant. The left half represents unpleasant feelings. The farther to the left, the more unpleasant.



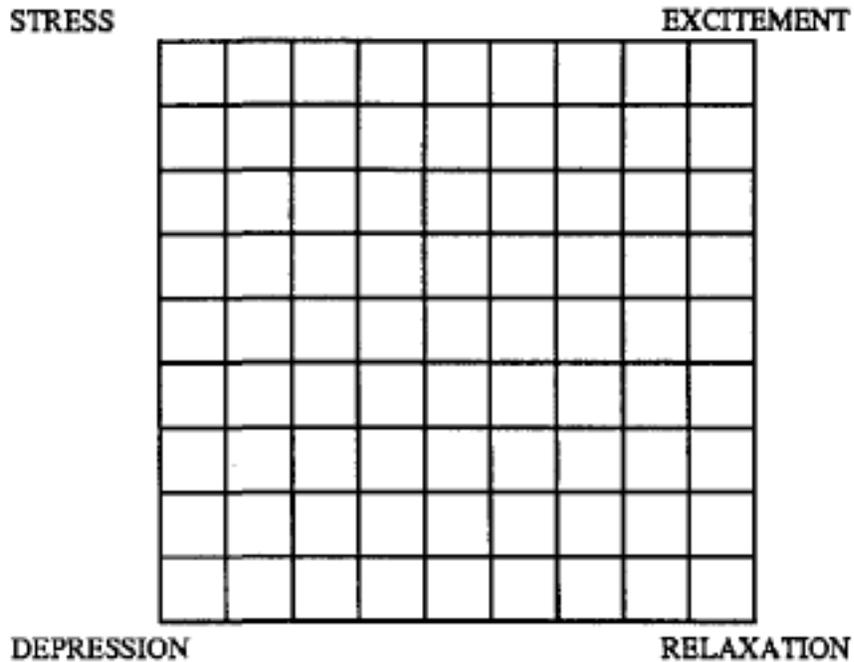
The vertical dimension of the map represents degree of arousal. Arousal has to do with how wide awake, alert, or activated a person feels - independent of whether the feeling is positive or negative. The top half is for feelings that are above average in arousal. The lower half for feelings below average. The bottom represents sleep, and the higher you go, the more awake a person feels. So, the next step up from the bottom would be half awake/half asleep. At the top of the square is maximum arousal. If you imagine a state we might call frantic excitement (remembering that it could be either positive or negative), then this feeling would define the top of the grid.



If the "frantic excitement" was positive it would, of course, fall on the right half of the grid. The more positive, the farther to the right. If the "frantic excitement" was negative, it would fall on the left half of the grid. The more negative, the farther to the left. If the "frantic excitement" was neither positive nor negative, then it would fall in the middle square of the top row, as shown below.

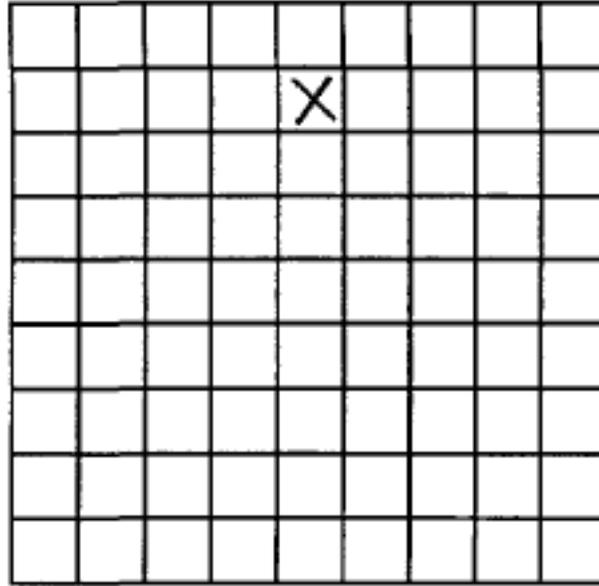


Other areas of the grid can be labeled as well. Up and to the right are feelings of ecstasy, excitement, joy. Opposite these, down and to the left, are feelings of depression, melancholy, sadness, and gloom. Up and to the left are feelings of stress and tension. Opposite these, down and to the right, are feelings of calm, relaxation, serenity.

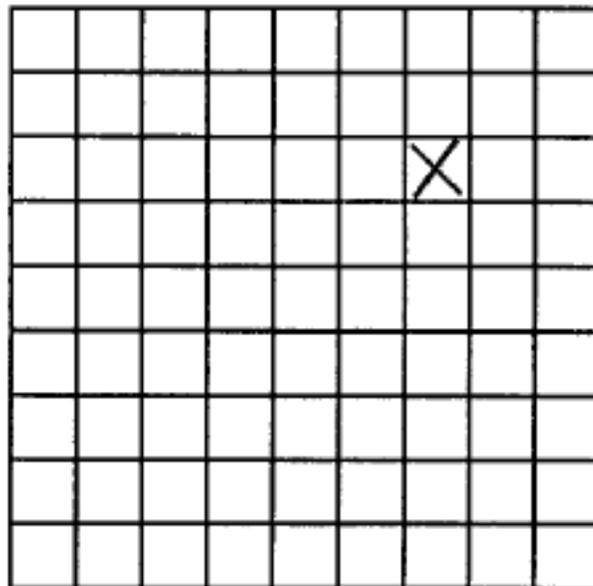


Feelings are complex. They come in all shades and degrees. The labels we have given are merely landmarks to help you understand the affect grid. When actually using the grid, put an X anywhere in the grid to indicate the exact shade and intensity of feeling. Please look over the entire grid to get a feel for the meaning of the various areas.

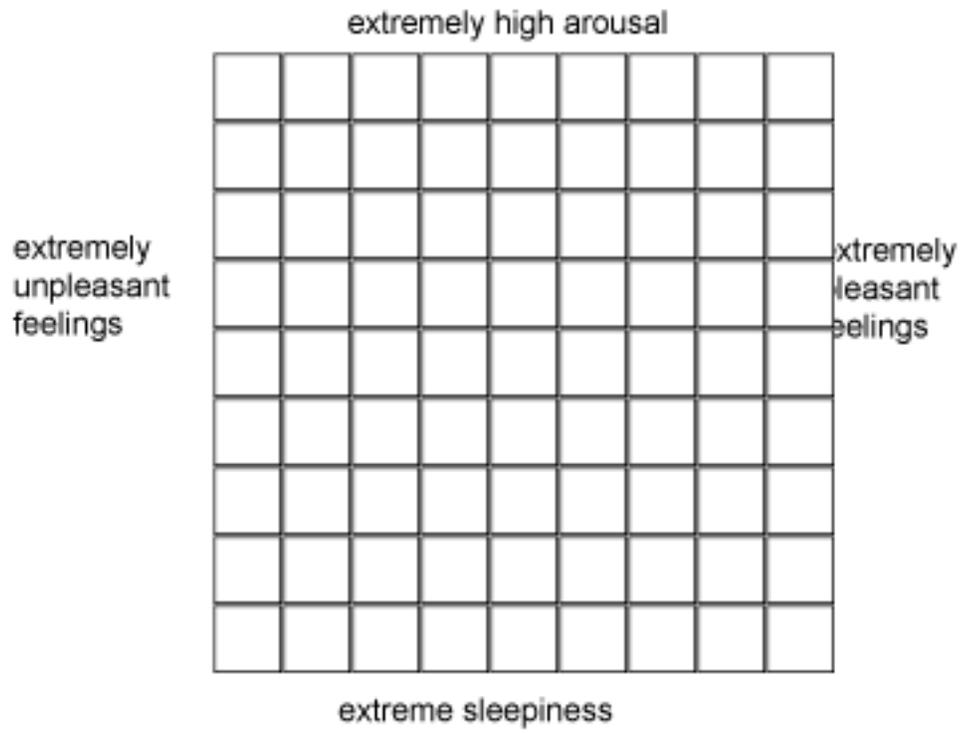
**Example:** Suppose that you were just surprised. Suppose further that the surprise was neither pleasant nor unpleasant. Probably you would feel more aroused than average. You might put your mark as shown.



**Example:** Suppose, instead, that you were only mildly surprised but that the surprise was a mildly pleasant one. You might put your mark as shown below.

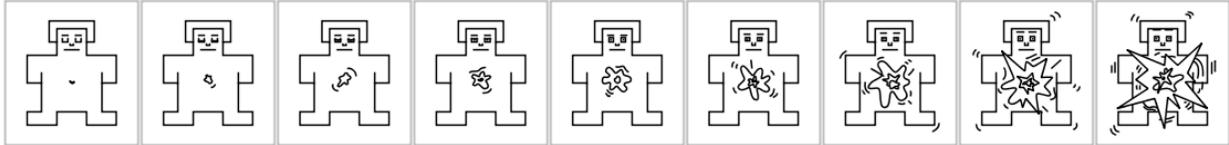


Now, click the square on the grid below that best captures how you feel right now:

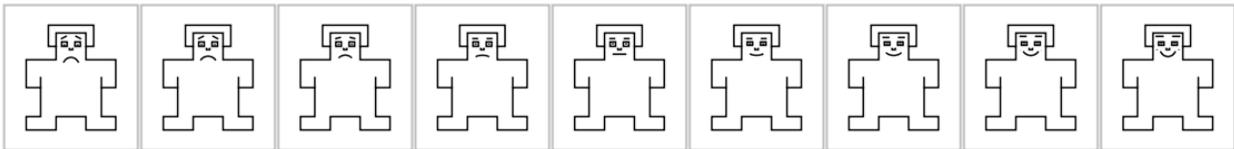


## SAM: The Self-Assessment Manikin

1. Select the image that best captures how you feel right now:



2. Select the image that best captures how you feel right now:



For this study, the ten item scales for Arousal (1) and Valence (2) were used.

## Perceptions of Internet Video Study

1. Have you seen this video clip before? (Yes/ No)

For the following questions, please select from a scale of 1 to 7 (with 1 being the least and 7 being the most) how much you agree with the following statements:

2. I found this video interesting

3. I enjoyed this video

4. This video will get me to [support the cause/ learn more/ go see the film].

5. In one sentence, please explain how this video made you feel.

## REFERENCES

- Barrett, L., & Barrett, D. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computer Review*, 19(2), 175–185.
- Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather Than Direct Causation. *Personality and Social Psychology Review*, 11(2), 167–203.
- Baumer, E. P. S., Katz, S. J., Freeman, J. E., Adams, P., Gonzales, A. L., Pollak, J. P., Retelny, D., et al. (n.d.). Prescriptive Persuasion and Open-Ended Social Awareness: Expanding the Design Space of Mobile Health. *Computer-Supported Cooperative Work 2012*.
- Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2005). Affect: from information to interaction. *Proceedings of AARHUIUS 2005*.
- Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, 65(4), 275–291.  
doi:10.1016/j.ijhcs.2006.11.016
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25(1), 49–59.
- Buck, R. W., Savin, V. J., Miller, R. E., & Caul, W. F. (1972). Communication of affect through facial expressions in humans *Journal of Personality and Social Psychology*, 23(3), 362–371. American Psychological Association.

- Burisch, M. (1984). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18*(1), 89–98.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.
- Cella, D. (1987). A brief poms measure of distress for cancer patients. *Journal of Chronic Diseases, 40*(10), 939–942.
- Chalfen, R. (1987). *Snapshot versions of life* (p. 213). Bowling Green, OH: Bowling Green.
- Charlson, M. E., Boutin-Foster, C., Mancuso, C. A., Peterson, J. C., Ogedegbe, G., Briggs, W. M., Robbins, L., et al. (2007). Randomized controlled trials of positive affect and self-affirmation to facilitate healthy behaviors in patients with cardiopulmonary diseases: rationale, trial design, and methods. *Contemporary clinical trials, 28*(6), 748–762.
- Cohen, S., & Pressman, S. (2006). Positive affect and health. *Current Directions in Psychological Science, 15*(3), 122–125.
- Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(4), 385–396.
- Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science : a journal of the American Psychological Society / APS, 15*(10), 687–693.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design & analysis issues for*

- field settings* (p. 405). Boston, MA: Houghton Mifflin.
- Crawford, J., & Henry, J. (2004). Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43*, 245–265.
- Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 3–17). Educational Testing Service.
- Csikszentmihalyi, M., & Hunter, J. (2003). Happiness in everyday life: The uses of experience sampling. *Journal of Happiness Studies, 4*, 185–199.
- D'Andrade, R., & Egan, M. (1974). The Colors of Emotion. *American Ethnologist, 1*(1), 49–63.
- Epstein, D. H., Willner-Reid, J., Vahabzadeh, M., Mezghanni, M., Lin, J.-L., & Preston, K. L. (2009). Real-time electronic diary reports of cue exposure and mood in the hours before cocaine and heroin craving and use. *Archives of general psychiatry, 66*(1), 88–94.
- Fogg, B., & Eckles, D. (2008). *Mobile Persuasion: 20 Perspectives on the Future of Behavior Change* (p. 5). Palo Alto, CA: Stanford University Press.
- Froehlich, J., Chen, M., & Consolvo, S. (2007). MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. *Proceedings of Mobysis 2007*, 57–70.
- Gaver, W., Beaver, J., & Benford, S. (2003). Ambiguity as a resource for design. *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*.

- Gay, G., & Hembrooke, H. (2004). *Activity-Centered Design: An Ecological Approach to Designing Smart Tools and Usable Systems*. Boston, MA: MIT Press.
- Gay, G., Pollak, J. P., Adams, P., & Leonard, J. P. (2011). Pilot Study of Aurora, a Social, Mobile-Phone-Based Emotion Sharing and Recording System. *Journal of Diabetes Science and Technology*, 5(2), 325–332.
- George, J. M. (1996). State and Trait Affect. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 145–171). San Francisco, CA: Jossey-Bass.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504–528.
- Gross, J. J., & Levenson, R. W. (1995). Emotion Elicitation Using Films. *Cognition and Emotion*, 9(1), 87–108.
- Gross, J. J., Richards, J., & John, O. (2006). Emotion regulation in everyday life. *Emotion regulation in couples and families: Pathways to Dysfunction and Health*. Washington, D.C.: American Psychological Association.
- Guillory, J., Spiegel, J., Drislane, M., Weiss, B., Donner, W., & Hancock, J. (2011). Upset now?: emotion contagion in distributed groups. In *CHI '11: Proceedings of the 2011 annual conference on Human factors in computing systems*.
- Hancock, J., Gee, K., Ciaccio, K., & Lin, J. (2008). I'm sad you're sad: emotional contagion in CMC. *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*.
- Harmon-Jones, E., Harmon-Jones, C., Abramson, L., & Peterson, C. K. (2009). PANAS

- Positive Activation Is Associated With Anger. *Emotion*, 9(2), 183–196.
- Hills, A., Hill, S., & Mamone, N. (2001). Induced mood and persistence at gaming - Hills - 2002 - Addiction - Wiley Online Library. *Addiction (Abingdon, England)*.
- Isomursu, M., Tähti, M., Väinämö, S., & Kuutti, K. (2007). Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies*, 65, 404–418.
- Kahneman, D. (2003). Objective Happiness. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-Being: The Foundations of Hedonic Psychology* (pp. 3–25). New York: Russell Sage Foundation.
- Kanade, T., Cohn, J. F., & Yingli Tian. (2000). Comprehensive database for facial expression analysis. *Fourth International Conference on Automatic Face and Gesture Recognition* (pp. 46–53).
- Kun, L., & Marsden, G. (2007). Co-present photo sharing on mobile devices. *Proceedings of Mobile HCI 2007*.
- Lang, A. (1990). Involuntary Attention and Physiological Arousal Evoked by Structural Features and Emotional Content in TV Commercials. *Communication Research*, 17(3), 275–299.
- Lang, P. J. (1995). The emotion probe. Studies of motivation and attention. *The American psychologist*, 50(5), 372–385.
- Larsen, R. J., & Sinnott, L. M. (1991). Meta-Analysis of Experimental Manipulations: Some Factors Affecting the Velten Mood Induction Procedure. *Personality and Social*

*Psychology Bulletin*, 17(3), 323–334.

Larson, R., & Csikszentmihalyi, M. (1983). The Experience Sampling Method. *New Directions for Methodology of Social and Behavioral Science*, 15, 41–56.

Lazarus, R. S., & Lazarus, B. N. (1994). *Passion and Reason: Making Sense of Our Emotions*. New York, NY: Oxford University Press.

Leahu, L., Schwenk, S., & Sengers, P. (2008). Subjective objectivity: negotiating emotional meaning. *Proceedings of DIS 2008*.

Macht, M. (2008). How emotions affect eating: A five-way model. *Appetite*, 50, 1-11.

Malouff, J., & Schutte, N. (1985). Evaluation of a short form of the POMS-Depression scale. *Journal of Clinical Psychology*, 41(3), 380-391.

Mancuso, C., & Charlson, M. E. (1995). Does Recollection Error Threaten the Validity of Cross-Sectional Studies of Effectiveness *Medical Care*, 33(4), AS77–AS88.

Mateas, M. (2001). Expressive AI: A hybrid art and science practice. *Leonardo*, 34(2), 147-153.

Mayer, E. A., Bradesi, S., Chang, L., Spiegel, B. M. R., Bueller, J. A., & Naliboff, B. D. (2008). Functional GI disorders: from animal models to drug development. *Gut*, 57(3), 384–404.

Mayer, J., DiPaolo, M., & Salovey, P. (1990). Perceiving Affective Content in Ambiguous Visual Stimuli: A Component of Emotional Intelligence. *Journal of Personality Assessment*, 54(3&4), 772–781.

- McNair, D., & Lorr, M. (1992). *Revised Manual for the Profile of Mood States* (p. 40). San Diego, CA: Educational and Industrial Testing Service.
- Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. MIT Press. Boston, MA: MIT Press.
- Meschtscherjakov, A., Weiss, A, & Scherndl, T. (2009). Utilizing Emoticons on mobile devices within ESM studies to measure emotions in the field. *Proceedings of MobileHCI 2009*.
- Morris, M. E., Kathawala, Q., Leen, T. K., Gorenstein, E. E., Guilak, F., Labhard, M., & Deleeuw, W. (2010). Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. *Journal of Medical Internet Research*, 12(2), e10.
- Moskowitz, D. S., & Young, S. N. (2006). Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of psychiatry & neuroscience : JPN*, 31(1), 13–20.
- Nasoz, F., Alvarez, K., Lisetti, C. L., & Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1), 4–14.
- Naz, K., & Helen, H. (2004). Color-emotion associations: Past experience and personal preference. *Proc. AIC 2004, The Interim Meeting of the International Color Association*.
- Nielsen, J. (1993). Iterative user-interface design. *IEEE Computer*, 26(11), 32–41.
- Pennebaker, J. W., Zech, E., & Rime, B. (2001). Disclosing and sharing emotion:

- Psychological, social, and health consequences. In M. S. Stroebe, W. Stroebe, R. O. Hansson, & H. Schut (Eds.), *Handbook of bereavement research: Consequences, coping, and care* (pp. 517–539). Washington, DC: American Psychological Association.
- Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.
- Pressman, S., & Cohen, S. (2005). Does Positive Affect Influence Health *Psychological Bulletin*, 131(6), 925–971.
- Reis, H. T., & Gable, S. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 190–217). Cambridge: Cambridge University Press.
- Rottenberg, J., Ray, R., & Gross, J. J. (2007). Emotion elicitation using films. In J. A. Coan & J. B. Allen (Eds.), *The Handbook of Emotion Elicitation and Assessment* (pp. 9–28). New York, NY: Oxford University Press.
- Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.
- Russell, J. A., Weiss, Anna, & Mendelsohn, G. A. (1989). Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology*, 57(3), 493–502.
- Salovey, P., Sieber, W., & Jobe, J. (1994). The recall of physical pain. In N. Schwartz & S. Sudman (Eds.), *Autobiographical Memory and the Validity of Retrospective Reports*. New York, NY: Springer-Verlag.
- Scherer, K. (2005). What are emotions? And how can they be measured *Social Science*

*Information*, 44(4), 695–729.

- Sengers, P., Kaye, J., Boehner, K., Fairbank, J., Gay, G., Medynskiy, Y., & Wyche, S. (2004). Culturally embedded computing. *Pervasive Computing, IEEE*, 3(1), 14–21.
- Sengers, P., Boehner, K., Mateas, M., & Gay, G. (2008). The disenchantment of affect. *Personal and Ubiquitous Computing*, 12(5).
- Slovic, P., Finucane, M., & Peters, E. (2007). The affect heuristic. *European Journal of Operational Research*, 177, 1333–1352.
- Sondhi, G., & Sloane, A. (2007). Digital Photo Sharing and Emotions In A Ubiquitous Smart Home. *IFIP International Federation for Information Processing, Volume 241, Home Informatics and Telematics: ICT for the Next Billion*, eds. Venkatesh, A., Gonsalves, T., Monk, A., Buckner, K. (Boston: Springer), 241, 185–200.
- Stone, A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, 24(3), 236–243.
- Sundström, P., Ståhl, A., & Höök, K. (2007). In situ informants exploring an emotional mobile messaging system in their everyday practice. *International Journal of Human-Computer Studies*, 65(4), 388–403.
- Swanson, G. M., & Ward, A. J. (1995). Recruiting Minorities Into Clinical Trials Toward a Participant-Friendly System. *JNCI Journal of the National Cancer Institute*, 87(23), 1747–1759.
- Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour research and therapy*, 6, 473–482.

- Watson, D., Clark, L., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070.
- Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The Two General Activation Systems of Affect: Structural Findings, Evolutionary Considerations, and Psychobiological Evidence. *Journal of Personality and Social Psychology*, 76(5), 820–838.
- 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, 557–580.
- Wright, P., & McCarthy, J. (2008). Empathy and Experience in HCI. *Proc. CHI 2008*, 10.
- Wundt, W. (1904). *Principles of Physiological Psychology*.