BEYOND STATIC REPRESENTATIONS: DEVELOPMENT OF A MULTI-CHANNEL, DYNAMIC ASSESSMENT OF EMOTION PERCEPTION.

BY

SCOTT E. BEDWELL

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee

Professor Fritz Drasgow, Chair
Professor Howard Berenbaum
Associate Professor F. Chris Fraley
Assistant Professor Jesse Spencer-Smith
Mary L. Doherty, College of America Pathologists

ABSTRACT

Research on the perception of emotion in humans has a long and rich literature base. Ekman's (Ekman, 1972; Ekman & Friesen, 1971) research on the universal recognition of emotions represents some of the earliest work in the area of emotion recognition. However, with the exception of recent work on synthesized faces to examine facial features related to emotion recognition in humans (Spencer-Smith et al., 2001), the field has not progressed very far beyond Ekman's original stimuli of emotion expressions in terms of developing standardized assessments of emotion perception skills. Recently, several researchers have called for the field to move beyond the emotion stimuli based on posed expressions of high signal clarity (Elfenbein & Ambady, 2002; Ambadar, Schooler, & Cohn, 2005). This research seeks to expand the field of emotion perception by developing a performance-based measure of emotion perception that capitalizes on new technologies incorporating full multimedia stimuli.

To Julie and Thomas

ACKNOWLEDGMENTS

Successful research is rarely the result of a single individual's efforts and the current project greatly benefited from the support of a number of people. I would like to thank my dissertation chair, Fritz Drasgow for his guidance and insightful comments through-out the dissertation project. I would also like to thank the other members of the committee for their willingness to serve on the committee and for their suggestions over the course of the project. I would also like to thank Bill Lindemann for his help in automating the formatting process for the matrix calculations. Special thanks to Stacey Jefford and Robin Lindemann for spending long hours in front of the computer providing assistance with data entry. Finally, but certainly not least, I would like to thank my wife, Julie for her support and patience as she experienced many evenings and weekends as a single parent during the course of the dissertation.

TABLE OF CONTENTS

CHAPTER 1: STATEMENT OF THE RESEARCH PROBLEM 1

CHAPTER 2: REVIEW OF RELEVANT RESEARCH 4

Research on Emotion Perception 4

Nature of Emotion Perception 6

Criticisms of the Early Emotion Perception Research Paradigm 11

Individual Differences in Emotion Perception 13

Current Assessments of Nonverbal Sensitivity and Emotion Perception 15

An Initial Attempt to Develop an Assessment of Emotion Perception 17

CHAPTER 3: HYPOTHESES 22

Structure of Emotion Perception 22

Subgroup Differences 23

Current Mood State and Target Emotion 24

Convergent and Discriminant Validity 24

CHAPER 4: METHODS 29

Development of the Video Scenes 30

Measures 33

Participants 36

CHAPTER 5: ANALYSES AND RESULTS 38

Dealing with Missing Data 38

Scoring the Video Scenes 39

Gender Differences 51

Minority Status 52

Current Mood 53

Convergent and Discriminant Validity 55

CHAPTER 6: DISCUSSION 59

Implications for Research 66

Implications for Practice 67

Limitations and Future Research 68

Conclusion 70

REFERENCES 71

APPENDIX A: ITEM LEVEL STATISTICS 82

CHAPTER 1:

STATEMENT OF THE RESEARCH PROBLEM

Emotions affect our behavior in myriad ways. Our emotions influence what environmental stimuli we attend to, how we interpret and evaluate information, even what information we are most likely to recall. Furthermore, there is some evidence that emotions influence our judgments and decisions (Ashforth & Humphrey, 1995; DeStano, Dasgupta, Bartlett, & Cajdric, 2004; Izard, 1991; Palfai & Salovey, 1993; Zajonc, 1984). This study is concerned with one particular aspect of emotion, the perception of emotion, a common occurrence in our day to day lives. The shy flirtatious smile. The small disapproving frown. The gleam of pride and happiness in the eyes of a new parent. All are examples of emotional communication. Recognizing signals of this kind is important in order to be able to function within social environments (Arbib, & Fellous, 2004; Marsh, Ambady, & Kleck, 2005).

With the coining of the popular phrase emotional intelligence (Salovey & Mayer, 1990) interest in research on emotional phenomena has been renewed. Since the initial introduction of emotional intelligence as an overarching theory of how emotions interact with cognitive processes to influence behavior, much research has been conducted on the topic. The literature base on emotional intelligence has exploded, in part because there are a variety of theoretical frameworks and assessment tools with only partial overlap among them. In addition to peer reviewed articles, several internet listservs have been created as discussion forums on the topic of emotions. Furthermore, a new journal developed solely for publishing emotion related studies has been founded. These outlets for discussions and debate on emotions and their influence on our lives are but one example of the increased interest in emotional functioning.

While this resurgence of scientific interest is surely beneficial, much of the current research is limited by a mismatch between the theory and the assessment process. Although several models of emotional intelligence have been put forth (Mayer, Salovey, & Caruso, 2000), the ability model of emotional intelligence has been the most researched and thus, has received the most scrutiny. The ability based model defines emotional intelligence as an ability and is quite explicit in specifying how it should relate to other, more traditional assessments of intelligence (Mayer & Salovey, 1993; 1997; Salovey & Mayer, 1990), perhaps contributing to its position as the dominant model in the literature.

Unfortunately, most assessments of emotional intelligence rely on self-report measures. While self-report measures of emotional intelligence have demonstrated some level of criterion-related validity with important outcomes such as job performance ratings, sales revenue and ratings of leadership (Bar-On, 1997; Bedwell, 2003; Kostman & Bedwell, 2003; Martin, Easton, Wilson, Takemoto, & Sullivan, 2004), the problems with using self-report methods to assess an ability construct is well documented (Mabe & West, 1982). An additional criticism is that self-report measures of emotional intelligence tend to be highly related to personality measures, casting some doubt as to whether they are truly measuring emotional intelligence as opposed to facets or composites of personality constructs (Davies, Stankov, & Roberts, 1998)

Although studies have been conducted to examine many of the hypotheses and questions generated by Mayer and Salovey's (1998) theory, criticisms of the theory still exist and further research is needed (Gignac, 2005; Murphy, 2006; Roberts, Zeidner, & Matthews, 2001). For example, in their four branch model of emotional intelligence, Mayer and his colleagues (Mayer, Salovey, & Caruso, 2002) used a faces task, asking respondents to view a picture of a face and rate how well each of several emotion labels characterizes the emotion displayed. While the use

of static expressions has been useful for answering some questions regarding emotion perception (Ekman, 1994) there have been numerous criticisms when it is used as the only stimulus (Bruner & Taguiri, 1954, Izard, 1994). These criticisms focus on the lack of ecological validity with the daily process of recognizing emotions: other people do not hold a facial expression for an unlimited amount of time or even express their emotions with the same intensity across situations or individuals.

This study addressed a key criticism of current assessment methods related to emotional intelligence: using self-report and deficient performance based assessments (i.e. static facial images of exemplar emotion expressions) to assess the dynamic process of emotion perception limits our understanding of the process and how it relates to behavior. Dynamic stimuli were used to create an assessment tool more closely representing daily emotion perception activities. The stimuli in the dynamic assessment varied the diversity of the targets in terms of race and gender, duration of the emotion, and the intensity of the emotional display in order to create a more ecologically valid representation of emotion perception.

CHAPTER 2:

REVIEW OF RELEVANT RESEARCH

Research on Emotion Perception

Just how well to do we recognize emotions in other people? The literature on emotion perception has a long history. Darwin (1872/1965) is widely credited with initiating an empirical approach to addressing what facial expressions convey (Russell, 1994). Prior to his classic text, *The Expression of Emotions in Man and Animals*, the ability of individuals to decode emotions from facial expressions using a common interpretative system was largely attributed to divine power. Darwin's initial attempts to empirically catalog individual's interpretation of facial expression set the stage for most of the emotion perception field for more than 100 years.

Empirical research on emotion recognition via facial expressions experienced a resurgence during the 1960's and 1970's with the classic works of Ekman and his colleagues (Ekman & Friesen, 1971; Ekman, Friesen, & Ellsworth, 1972). During this time Ekman's research largely focused on the question of whether emotions were expressed universally across cultures using the same facial expression (1994). Put another way, do specific facial expressions indicate specific emotional experiences regardless of culture? The research base addressing this question has provided some evidence addressing the question of how well we recognize emotions in others through nonverbal signals.

In fact, there is quite good agreement that certain facial expressions do indicate specific emotions, even across cultures. Ekman (1994) concluded on the basis of his research on the issue that there are at least six basic emotions that are reliably recognized by all people, regardless of culture or language: happiness, surprise, sadness, fear, anger, and disgust.

However, some emotions are better recognized than others. For example, happiness is generally found to have the highest accuracy rates while fear and disgust generally have the lowest accuracy rates of the six emotions commonly studied (Elfenbein & Ambady, 2002). In addition, some emotion categories tend to be confused more often than others.

Examining the effects of target race, perceiver sex and mode of presentation, Gitter and his colleagues (Gitter, Kozel, & Mostofsky, 1972) observed that surprise was judged as happiness about 15 percent of the time and disgust was perceived as anger approximately 22 percent of the time. Looking at the data more specifically, they also found that when the target expresser was White, fear was most often confused with surprise. However, when the target expresser was Black, fear was most often confused with sadness. These authors did not report the ethnicity of the sample.

The research on emotion perception has historically assumed that facial expressions signal emotional experiences. However, some theorists have challenged this assumption and asked whether facial expressions truly convey signals revealing the subjective feelings an individual is experiencing, as emotion theorists assume, or whether they convey social communication signals designed to regulate social interactions as behavioral ecologists assert (Fridlund 1994). The behavioral ecology theory holds that communication signals via facial expression are relaying either behavioral intentions or action requests. Behavioral intentions are defined as signaling a behavior that is about to be initiated by the target. For example, "I am going to attack you." Similarly, an action request is a communicated message requesting an action from the perceiver. For example, "please leave now."

In order to address this question, Horstmann (2003) asked individuals to classify facial expressions as exhibiting feelings, behavioral intentions, or action requests by the target. The

term “feelings” was used so as not to unduly influence participants’ responses where the term “emotions” might have created demand characteristics. The results of the study were classified according to both emotion (six of Ekman & Friesen’s, 1976 facial expressions were used) and culture/language (US English vs German). Results indicated that with the exception of anger, all of the facial expressions were clearly classified as communicating a feeling state. The percentage of individuals classifying the facial expressions as indicating feeling states ranged from 44 percent for anger to 83 percent for surprise. The behavioral intention classification rate for anger was 32 percent in the US sample, suggesting that expressions of anger at least, may serve multiple communication functions. The percentage of the sample rating expressions of disgust as an emotion was also somewhat lower than the remaining emotion labels (by at least ten percentage points). This is not surprising given that anger and disgust have been observed to be more similar in previous studies using emotion recognition tasks (Gitter et al, 1972). Overall, the results of Horstmann’s study are more supportive of the emotion theorists’ perspective that the primary message conveyed by facial expression is emotional experiences or subjective feeling states.

Nature of Emotion Perception

Keltner and Ekman (2000) suggested that four lines of research provide evidence supporting a taxonomic structure of emotions as opposed to a set of underlying dimensions: categorical judgment studies, neuropsychological studies (e.g., *f*MRI), facial expressions and autonomic activity, and facial expressions and evoked reactions in observers. However, the authors acknowledged that while emotions may be discrete, the research studying emotions aggregated across may benefit from using dimensions to classify emotions. It is also possible to

reconcile the two approaches by looking at dimensions within discrete categories (e.g., arousal differences in irritation and fury), similar to the analogy of looking at the color spectrum. While the underlying structure of colors is continuous, we perceive colors categorically.

Further evidence that discrete emotion categories appear to show more correspondence to the way individuals typically think about emotions has been demonstrated by directly comparing a categorical structure to a dimensional structure from participant judgments. Etcoff and Magee (1992) asked respondents to make judgments on a series of morphed faces that ranged from an expression of a pure emotion (e.g, happiness) to another emotion such as sadness. The faces in between the pure anchor expressions consisted of morphed images of the two anchor expressions. The morphed images differed by a constant degree and followed a symmetrical distribution with ratios of 90:10, 80:20, 70:30, 60:40, 50:50. Their results indicated a clear boundary in the middle of distribution. Although the location of the boundary differed depending on the anchor emotions, the authors found that discrimination accuracy was more accurate for pairs of facial expressions crossing the boundary while paired expressions on the same side were discriminated at chance levels. The authors interpreted this finding as supporting a theory that individuals make categorical judgments of facial expressions displaying emotions, even if the underlying structure is not categorical.

Laukka (2005) conducted a similar set of studies using content masked emotional speech. Again constraining the separate stimuli to be equidistant in terms of emotional tone, she asked respondents to make judgments as to what emotion was communicated by the tone. The stimuli consisted of the same phrase spoken by a professional actor under four different emotion conditions; anger, happy, sadness, and fear. These were then blended in five different ratios (90:10, 70:30, 50:50, 30:70, 10:90) and participants were asked to identify the emotion. Similar

to Etcoff and Magee (1992), Laukka found that discrimination across emotion categories was better than discrimination within categories even when the stimuli were constrained to be equidistant on an underlying continuous distribution.

These two studies represent an alternative method of testing two competing theories on the structure of emotions. While both Ekman (Ekman & Friese, 1971) and Russell (Russell & Steiger, 1982) present empirical data supporting their own theories of categorical and dimensional structures of emotion, the Etcoff and Magee (1992) and Laukka (2005) studies were specifically designed to compare the two theories. Both studies, conducted independently but using similar methodology, came to the same conclusion that individuals make categorical judgments of emotion. Even more striking is that the results replicated across emotion cues.

Subgroup Differences in Emotion Recognition

Gender. In addition to studying the extent to which facial expressions are universally recognized as communicating a specific emotion, researchers have examined whether group membership status other than culture would have an effect on the ability to decode emotion signal conveyed by facial expressions. Early studies on gender and emotion recognition presented mixed results for the common belief in women's intuition (Gitter et al., 1972; Hall, 1978). However, in her empirical review of the early literature on emotion recognition, Hall concluded that females consistently performed better at recognizing emotions from nonverbal cues across a variety of experimental and observational methods. She demonstrated that on average women were more significantly more accurate in emotion recognition than males and the difference was approximately 0.4 standard deviations. More recent studies investigating

emotion perception continue to document this difference (Elfenbein & Ambady, 2002; Hall & Matsumoto, 2004; Kirouac & Dore, 1985).

In addition to the evidence that females are superior to males in emotion recognition, there is some evidence that perceptions of emotions expressed by females are influenced by Western gender stereotypes. Hess, Adams, and Kleck (2004) suggested that Western gender stereotypes portraying females as more affiliative and males as more dominant influence perceptions of emotions. Using morphed images to create an androgenous face to control for stereotypic markers of affiliation and dominance in male and female faces, the authors found that simply manipulating the hair style of the morphed image influenced perceptions of the emotion being expressed. Specifically, images of apparent females expressing anger were more likely to be judged as expressing anger than similar apparent male faces. In addition, ratings of emotion intensity for female faces were higher than similar apparent male faces. Likewise, apparent males faces expressing happiness were judged more often as happy compared to similar apparent females and apparent male faces expressing happiness were rated as more intense. Faces displaying sadness did not follow the same pattern. Hess et al. argue that due to stereotypical expectations for females to express affiliate emotions and males to express dominance emotions, judgments of emotions inconsistent with expectations are judged to be more intense.

Race. Research findings on the influence on race on emotion perception are not as clear as those for gender. However, there appear to be two common facets of study designs that may account for this lack of clarity. First, race or ethnicity has often been confounded with cultural differences and language barriers in emotion perception studies. Studies identifying multiple racial groups within a culture are rare. Second, when multiple racial groups are specified within a single culture, only the race the of the target expressing the emotion has been employed as an

independent variable (Gitter et al, 1972; Hugenberg, 2005). Studies focusing on the race of the perceiver are almost nonexistent.

Still, there are some inferences that can be drawn from the literature despite these limitations. In his critique of the emotion perception literature, Russell (1994) observed that Asian and African samples reported less agreement in identifying Ekman's six basic emotions from facial expressions than samples from Western cultures, suggesting there may be an influence of perceiver race on emotion perception. Furthermore, while Elfenbein and Ambady (2002) reported evidence supporting the universality hypothesis of emotion recognition in their meta-analysis of emotion perception research, they also observed a significant in-group advantage such that when the target and perceiver were members of the same group, emotion recognition rates were higher. Unfortunately in-group was broadly defined across racial, cultural and national group membership so that a clear picture of how ethnicity specifically influences emotion perception on the part of the perceiver independently of culture or nationality was not possible.

Minority Status. Elfenbein and Ambady (2002) also tested whether being a member of a minority group influences emotion recognition. For this analysis, minority status was defined a cultural subgroup living among members of the majority culture. While their sample size was limited to a small number of studies, they observed a trend suggesting that minority group members were better at perceiving emotions of majority group members than vice-versa. Even more striking was that the emotion recognition rates of minorities judging members of the majority group were on average higher than those of the majority group judging members of their own group.

Criticisms of the Early Emotion Perception Research Paradigm

Ekman's groundbreaking work on the universal recognition of emotions from facial expressions is important because it hints at a biological or neurological basis for emotion recognition. While this finding is important, the methodology of the research behind it has been criticized (Russell, 1994). For example, most research examining emotion recognition from facial expressions used static pictures or drawings of faces as the stimuli. In addition, the expressions displayed tend to be prototypical or exaggerated poses that capture only the upper range of intensity. Russell has argued that these research procedures, combined with a tendency to rely on forced choice response categories of emotion labels could inflate the overall agreement levels among participants.

Prototypical Expressions

Many of the stimuli used in the early cross-cultural research on emotion perception used faces expressing exaggerated or prototypical expressions. Russell (1994) refers to these as preselected faces. He argues that the process of iteratively selecting photographs for inclusion in research studies on the basis of how characteristic the expression is of the intended emotion will likely increase agreement among participants relative to using more natural expressions. Using images that include a range of intense and subtle expressions may result in lower levels of agreement.

Forced Choice Response Formats

Most studies of emotion perception follow Ekman's original research design, dating back to Darwin (1872/1965), asking participants to choose the emotion label that best describes the

emotion portrayed by the target facial expression. Indeed, in their meta-analysis Elfenbein and Ambady (2002) were only able to locate a handful of studies that did not include a forced choice response option. Russell (1994) contends that the use of the same choices for all trials result in demand characteristics such that participants are aware of the researcher's expectations. He notes that over repeated trials, "the set of category choices is primed and might influence subsequent perception". Russell further argues that allowing only one response choice per image can create an artificial pattern of mutually exclusive categories for emotion perception.

Russell (1994) advocates the use of ratings scales in emotion perception research to overcome this issue. Interestingly, this criticism is acknowledge by Ekman (1994) who agreed that including ratings scales in research on emotion perception is valuable. Izard (1992) goes further suggesting that there are occasions when allowing participants to choose or rate more than one emotion label is necessary, particularly when the image does not represent an exemplar of the intended emotion, but rather is more ambiguous or conveys a blend of emotions.

Russell (1994) also advocated the use of free responses to emotional stimuli. While this method has been used in previous studies, it is limited in that it is difficult to aggregate free response data in any standardized way. Thus, at some point researchers will need to collapse the data into meaningful categories or develop methodologies to analyze free responses in a systematic manner. If they choose the former, then they must deal with how much of the raw information to report in order to justify the decisions that were made during the process of developing the categories. Ekman (1994) argued that using freely chosen responses to emotional stimuli in research designs has largely been abandoned precisely due to these issues.

Static Images

Human perception processes inherently attend to change in our environment (Roseman & Smith, 2001) and so motion should be particularly relevant for assessments of emotion perception. Indeed, concerns about research designs relying on static images or still pictures are not new. Calls for the inclusion of movement in emotion perception research were made early on in the literature (Bruner & Taguiri, 1954; Jenness, 1932). Although it has long been acknowledged that perceptions of emotions occur in fleeting instances, emotion perception research did not begin to include dynamic stimuli until almost 40 years later, (Gitter, et al., 1972; Basili, 1979). These studies provided initial evidence that emotion recognition rates increase when dynamic images of emotion expressions are used relative to still images.

More recent evidence has confirmed these initial studies (Ambadar, et al., 2005; Elfenbein & Ambady, 2002). Ambadar et al. demonstrated that motion is particularly important in the detection of subtle expressions of emotion ubiquitous in day-to-day interactions. These authors compared real time video images of an emotional expression starting from a neutral baseline, which progressed to a low intensity expression with single images cropped from the last frame in the video, representing the peak of the expression. In addition they used multiple images from the video sequence that were displayed at a slower rate. The results of the study suggested that the more dynamic the display, the more accurate the emotion judgments.

Individual Differences in Emotion Perception

Although the criticisms leveled at the methodology of Ekman's research paradigm on cross-cultural emotion perception suggest that several aspects of the design could, in combination, artificially inflate agreement levels on emotion recognition, this does not

necessarily undermine his conclusions regarding the universality of recognition for some emotions (Ekman, 1994; Izard, 1994). While critics of the theory on universal recognition of emotions suggest that any deviation from complete agreement undermines the validity of the theory, proponents of theory argue that any agreement levels beyond chance are supportive (Ekman 1993; 1994). Indeed, several authors have suggested that both cultural and individual differences can affect emotion recognition agreement levels (Ekman 1993; 1994; Izard, 1994).

While much of the research on emotion perception has focused on the mean accuracy rates of emotion recognition, little attention has been paid to alternative explanations for the variance that almost always accompanies the experimental manipulation. That is, no emotion perception studies have reported 100 percent agreement in accuracy rates for any emotion, even when the emotion faces are prototypical or exaggerated versions of the expression that occurs in everyday encounters. The failure to obtain 100 percent agreement levels may be an indication of individual differences in emotion perception skill.

These stimuli tend to be highly exaggerated in intensity, leaving little room for reliable individual differences to emerge. The variance that has been used by some researchers to argue against universality may in fact represent an individual difference in the ability to identify emotional signals from other people. If individual differences in emotion perception accuracy are demonstrated to be reliable using standardized assessment procedures, then these individual differences may be useful in explaining the lack of perfect agreement between people, as well as help predict other important real world behaviors such as social skill, and influence and persuasion tactics.

Previously, several researchers have suggested that such individual differences in the accuracy in emotion perception exists (Mayer & Geher, 1996; Nowicki & Duke, 1994; Salovey

& Mayer, 1990). Empirical evidence suggests that there are differences in how well individuals can recognize emotions in other people. Forsyth, Kushner, and Forsyth (1981) used a hierarchical cluster analysis to identify groups of individuals who had rated photographs on four dimensions developed from free choice responses to the photographs. The dimensions were annoyance, interest, understanding, and spontaneity. Results from the cluster analysis revealed several homogenous groups with different profiles across the dimensions. Forsyth et al. concluded from these results that there are differences between individuals in terms of what cues they attend to when processing information from the face (e.g, curvature of the mouth, raise of cheeks or eyebrows, etc)..

Pessoa, Japee, and Ungerleider, (2005) also observed individual differences in the perception of fear. Facial expressions of fear, happiness or neutral expressions were displayed for 17, 33, or 83 ms. Participants were asked to respond to each expression as either "fear" or no fear". Results indicated individual differences in categorizing expressions as "fear or no fear" depending on the time display suggesting there was no universal cutoff for reliable detection.

Current Assessments of Nonverbal Sensitivity and Emotion Perception

Several assessments have been developed to assess various aspects of nonverbal communication. The Ekman and Friesen (1976) faces are perhaps the best well known emotion expression stimuli. Although these images are still widely used in research, given Russell's (1994) criticism, they may be of limited ecological validity. The Profile of Nonverbal Sensitivity (PONS; Hall, 2001; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979). is also a well known assessment of nonverbal decoding skill. The PONS assesses nonverbal sensitivity through several communication channels including face, body, and prosodic tone. These different

channels are assessed through separate stimuli and in various combinations. However, the PONS asks respondents to decide whether the emotional valence is positive or negative and whether the target is dominant or submissive. As such it is not, nor was it intended to be, an assessment of emotion perception.

The authors of the Diagnostic Analysis of Nonverbal Accuracy (DANVA; Nowicki & Duke, 1984; 2001) cite several of the reasons discussed above for the development of the DANVA. Because the original measure was developed primarily for use with children, the DANVA2 (Baum & Nowicki, 1998) was created for use with individuals across the developmental span. The DANVA2 measures emotion perception using static images of four of Ekman and Friesen's basic emotions: happiness, sadness, fear and anger. Emotion perception accuracy is assessed through two channels, facial and paralingual displays. The facial displays include both high and low intensity expressions for each emotion. The paralingual subtest consists of actors' speaking a neutral phrase with emotional tones for each of the emotions after reading vignettes designed to elicit emotions. However, the authors acknowledge that for questions requiring ecological relevance, an assessment of emotion perception using video would be more appropriate.

The Mayer-Salovey-Caruso-Emotional Intelligence Test (MSCEIT; Mayer et al., 2002) is the most recent assessment of emotion perception. The MSCEIT is a revised version of the authors' original performance assessment of emotional intelligence, the Multifactor Emotional Intelligence Scale (MEIS; Mayer, Salovey & Caruso,1997). The MSCEIT is based on a four branch model of emotional intelligence with the first branch comprising emotion perception. Two subtests on the MSCEIT assess emotion perception. The first is a task similar to Ekman and Frisen's cross-culture research stimuli and the DANVA. Participants are asked to view a static

image of a face and rate the extent to which several emotions are depicted. The second task requires individuals to rate the emotions they perceive in pictures of landscapes and abstract art. This latter task is based on research suggesting that perceptions of emotion in ambiguous stimuli is related to the overall ability to decode emotions (Mayer, DiPaolo, & Salovey, 1990).

Despite an empirical literature base suggesting that dynamic stimuli, a focus specifically on discrete emotions, and multiple channels of emotion communication are all important features, none of the assessments described above contain all of these feature. For this reason the current study is proposed as a next step in advancing the assessment of emotion perception as an individual difference construct.

An Initial Attempt to Develop an Assessment of Emotion Perception

In an initial effort to develop an assessment of emotion perception incorporating dynamic stimuli and multiple channels of information, Bedwell and Chuah (2007) reviewed movies for emotional expressions. The films were downloaded from an Internet site of archived movies for which the copyright had expired.

Identification of Emotion Clips

Scenes identified as suitable for inclusion in the stimulus materials were spliced together along with an instruction screen identifying the target in the video clip, followed by a screen displaying the response alternatives. The scenes were independently reviewed by a second researcher for emotion content prior to inclusion in the study. In order to obtain a judgment of the emotion portrayed in each video clip independent of researcher judgments, the film sequences were presented to research participants who were asked to identify the emotion

portrayed in the scenes using a free response format. The free response data were then combined to form emotion categories for subsequent trials by combining similar terms and taking the modal term as the label for the category.

The emotion recognition task in this study was intended to be more difficult than that employed by cross-cultural emotion researchers because we were interested in variance between participants rather than high levels of agreement within and across groups. Therefore, for several of the film clips, we included response categories that were gradations of the target emotion category, such as angry and frustrated, in order to capture information regarding the intensity of the emotion expression. However, after an initial data collection trial, it became clear that participants had difficulty discriminating among several of the gradations within emotion categories. Therefore, in subsequent data collection sessions the response alternatives were refined by limiting the response options so that only one emotion label from an emotion category was presented. For each video clip participants were asked to rate the extent to which the emotion was displayed in the scene using a scale from one to five. Three emotions were listed for each scene in order to avoid criticism of forced choice responses (Russell, 1994) as well as to allow multiple responses to scenes that might contain mixed or blended emotion expressions (Izard, 1994). In addition, we asked participants to rate the intensity of the emotion, regardless of content. Results indicated that there was substantial variance in labeling the emotion perceived in the video clip, even with this simplified response procedure.

Scoring

Prior to examining the psychometric properties of the scenes, a scoring algorithm needed to be developed. There are a number of ways to score individual items. However, some of the

methods commonly discussed in the literature were unavailable to due to the method used to create the emotion stimulus scenes (e.g., correspondence of the rated emotion with the emotion intended by the target). A combination of expert ratings and consensus scoring was employed.

The authors served as the initial set of expert raters. During the course of selecting an appropriate scene from a film, a target emotion was identified by the primary reviewer. Then the second reviewer for that film would view the scene and make a judgment as to what emotion was being displayed. Scenes in which both reviewers did not agree on the emotion being displayed were not included in the final pool of video clips shown to research participants.

Each scene was considered a single measurement opportunity and was scored as correct if the target emotion identified earlier by the researchers received a higher rating than the other emotion categories presented for that scene. A consensus scoring procedure used by Mayer and Geher (1996) was also employed. For this procedure, each scene contained three measurement opportunities, one for each of the emotion categories presented. The modal response for each emotion was keyed as correct along with responses within one unit of the modal response. For example, if on the one to five rating scale, the modal response was three, responses of two, three, and four were keyed as correct.

Initial results from the pilot study suggest that the approaches to assessing emotion perception skill described above were fruitful. The expert scoring procedure resulted in a final measure consisting of 20 scenes and an internal consistency estimate of .63. The consensus scoring process resulted in a final scale consisting of 16 scenes with an internal consistency estimate of .65. Interestingly, only 12 scenes overlapped between the two scoring systems and the correlation between the scoring systems was fairly small (r=.22).

This latter result suggests that the consensus and expert scoring methods were not consistent with each other. Even correcting for unreliability in both scoring methods, the corrected correlation increased only to .34. Looking closer at the scoring routines, the expert scoring routine used more limited information than the consensus scoring routine. In the expert scoring, only one emotion category per video sequence was used as a correct score, so the number of measurement opportunities was less than the consensus scoring, which used all three emotion categories per video sequence. Clearly a more sophisticated scoring system is needed.

Keeping in mind the limitations just discussed, the authors found that the emotion perception scores using either expert scoring or consensus scoring were not related to personality or cognitive ability variables. In addition, both scoring routines demonstrated small to moderate correlations with perceptions of emotions from faces displayed on a static slide.

Limitations of the Pilot Study

The study was an initial foray into the use of existing film clips to demonstrate reliable individual differences in emotion perception, and like most first attempts, there were several limitations to the study. First, the authors used black and white films in which the picture quality was not as clear as would be desirable. Although there is a dearth research that has directly examined the effects of picture quality on the accuracy of judgments, Ambady, Hallahan, and Conner (1999) found that respondents could make consistently accurate judgments of sexual orientation from dynamic figural outlines created from previously filmed targets. However, the judgments were not as accurate as those made after viewing the actual video clips, suggesting that degradation of the video image affected the quality of the judgments.

Similar to the issue of video quality, the audio quality of the video used in the pilot study was somewhat poor. While this probably would not have been an issue for speech, cues such as a heavy sigh, gasp of breath, increased breathing, or rate of speech may have been more difficult to perceive than would be the case under more natural conditions.

Another limitation of the initial study is that while the authors attempted to include an approximately equal number of male and female targets in the video clips, ethnic diversity was not so easily obtained. Almost none of the films reviewed contained non-Caucasian actors and there were no usable scenes to include in the final pool of video clips in this research. Likewise, this study made no attempt to control the number video sequences that displayed a given emotion. Given the literature indicating that identification accuracy differs for specific emotions (Elfenbein & AMbady, 2002), this over sampling of some emotions may have led to an inaccurate picture of emotion perception skill.

CHAPTER 3:

HYPOTHESES

Structure of Emotion Perception

Several different structures of mood and emotion have been proposed. These included discrete emotion categories, Russell's (1980) circumplex model and Watson, Clark and Tellegan's (1988) variation from a rotation of Russell's model resulting in two factor model of positive and negative affect. While the current format for the item response options explicitly conforms to a discrete emotion structure on the basis of research supporting categories in perception, it is necessary to examine whether the pattern of responses also follows a discrete categorical structure. That is, should a score be developed for each individual emotion family targeted in the development of the film scenes? Or does a two factor model provide a better summary of the relationships among the responses. It is also possible the content of the emotion itself does not influence the perception skill, in which case a single factor model would be expected to demonstrate the best a better fit to the data. In order to investigate the structure of the current assessment of emotion perception, a series of confirmatory factor models were tested against each other to determine which model provides the best fit to the data. The models tested included a single factor model, a two factor model corresponding to positive and negative affect, and a six factor model corresponding to each of the distinct emotion portrayed in the video sequences. As there is little data on which to posit which of the hypothesized structures is most likely, these analyses were considered exploratory.

Subgroup Differences

As noted earlier, there is a large body of research suggesting that stereotypes of female and male emotionality influence judgments of perceived emotion (Brody & Hall, 2000; Hess et al., 2004). In addition, females have consistently demonstrated superiority in emotion recognition tasks relative to men (Elfenbein & Ambady, 2002; Hall, 1978). Considering these finding in combination, it is likely that the influence of gender on emotion recognition is not a simple main effect, but an interaction of both target and perceiver gender.

Hypothesis 1a: Females will be more accurate than males on the emotion perception tasks.

Hypothesis 1b: Females will make higher ratings for scenes with female targets experiencing emotions incongruent with western gender stereotypes .

While there have been fewer studies explicitly comparing the accuracy of judgments between race within the same culture, there is some evidence for an advantage for minority groups (Elfenbein & Ambady, 2002). Therefore, it is predicted that minority groups will be more accurate in recognizing emotions.

Hypothesis 2: Minority participants will score higher than Caucasian participants on the emotion perception task.

Current Mood State and Target Emotion

Moods and emotions influence what we attend to and the judgments people make. Mayer, Gashke, Braverman, and Evans (1992) demonstrated that mood congruent effects generalize to mood congruent memory tasks. In particular they found evidence that mood influenced subsequent perceptions of others. In addition, recent evidence that higher levels of negative affect, in particular anxiety, result in confusion for judgments of fear and sadness. McClure and Nowicki (2001) provide further support that current mood influences perceptions of others. Results from these studies suggests that there may be a specific effect of current mood on emotion perception judgments as well. One potential form of this effect could result in individuals who experience more negative moods will be more perceptive of negative emotion displays and individuals experiencing more positive moods will be more perceptive of positive emotion displays.

Hypothesis 3a: Individuals with higher scores on positive affect will rate scenes with positive emotions higher than individuals with low scores on positive affect.

Hypothesis 3b: Individuals with higher scores on negative affect will rate scenes with negative emotions higher than individuals with low scores on negative affect.

Convergent and Discriminant Validity

As part of the construct validity process, it is necessary to establish the position of emotion recognition in the nomological network of individual difference constructs. One method of doing so is to demonstrate a pattern of convergent and discriminant relationships with other

measures (Crohnbach & Meehl, 1955; Campbell & Fiske, 1959). Therefore, this study examined the relationships between the video-based measure of emotion perception and existing static images of facial expressions, empathy, personality, and cognitive ability.

Emotion Perception From Static Images

Comparisons of the video-based assessment with static images should result in moderate correlations. Although the addition of motion, context, and tone should provide the perceiver with additional information, enhancing recognition performance relative to static images, images have typically displayed exaggerated or exemplar expressions of emotions, which would make these static images more recognizable in terms of emotion displays. In addition, the video sequences were shown such that the intensity of the emotion varied from subtle to obvious displays. As a result, the video images varied along intensity, duration, and clarity (Izard, 1994) of the emotion expressed. Therefore, a ceiling effect on the means of the static images was expected to attenuate the correlation between static images and video sequences.

To control for the intensity of the static images, computer generated synthetic images with known valence and intensity characteristics were used rather than photographs. These "Poser" images were taken from the pool of images developed by Spencer-Smith and his colleagues (Spencer-Smith, Wild, Inners-Ker, Townsend, Duffy, Edwards, Ervin, Merritt, & Paik, 2001). In their study comparisons between the "poser" images and photographs of Ekman and Friesen's (1976) "universal emotions" (happiness, surprise, disgust, fear, anger, and sadness) suggested little differences in the perceived emotion by raters. A follow-up study demonstrated noticeable differences in ratings of intensity across synthesized expressions displaying subtle and intense expressions.

Hypothesis 4a: Scores on the video based emotion perception task will demonstrate moderate positive correlations with responses to static facial images posing emotional expressions.

In order to account for the additional information that movement in the expression creates (Ambadar et al, 2005; Bassili, 1979), pseudo dynamic images of facial expressions were also used. Computer generated faces were initially set at a neutral expression (no emotion) as well as an expression of one of the six basic emotions from Ekman's research (anger, fear, sadness, disgust, surprise and happiness). The neutral expression was separated from the emotion expression image by a 0.5 millisecond black screen. This effect creates the illusion of movement from a neutral to emotional expression. The dynamic facial expression can be considered a step closer toward fidelity in everyday activities, relative to the static facial images and should therefore be more similar to the video-based emotion perception task.

Hypothesis 4b: Accuracy of perception of the dynamic facial images will demonstrate higher correlations with the video-based emotion perception task than will responses to the static facial images.

Empathy

Empathy denotes a sharing of an emotional experience by two or more individuals (Mehrabian & Epstein, 1972). Empathy is not just an understanding of what another individual is feeling, it is the reproduction of that feeling by the observer. Empathetic individuals should be

adept at identifying the emotional state of others and previous research using several methods of assessing emotion perception have found some evidence supporting the relationship between emotional perception and empathy (Davies, Stankov & Roberts, 1998; Mayer et al., 1990; Mayer & Geher, 1996).

Hypothesis 5: Empathy will be moderately correlated with scores on the video-based emotion perception task.

Correlations with Personality

Previously, task based methods of assessing emotional intelligence have demonstrated low to zero correlations with dimensions of personality, typically using self-report measures of the five factor model (Digman, 1990). In addition, while the personality trait of Extraversion has been hypothesized to influence social skill, in which emotion perception should also play a role, previous research has found very little evidence that extraverted individuals are actually more socially adept in controlled experimental settings.

Recently, Lieberman and Rosenthal (2001) suggested a cognitive resources model to account for this lack of a relationship in previous studies. They suggested that extraverted individuals were better able to process the multiple tasks associated with social behavior, such as processing nonverbal signals, attending to content of conversations, and developing an appropriate response. Past research examining the relationship between extraversion and nonverbal communication have only required participants to focus on the nonverbal signals, so the remaining social interaction activities did not interfere with introverted individuals' ability to decode nonverbal signals. In a series of studies designed to separate out these task, Lieberman

and Rosenthal found support for this hypothesis. When asked to perform multiple social interaction tasks simultaneously, introverts were perceived as being less socially adept and could not accurately judge how partners viewed them, relative to extraverted individuals. However, in the current research, scores on the emotion perception task were not expected to have any relation with self-report personality measures because no multitasking activities was required.

Hypothesis 6: Scores on the personality factors of Extraversion, Neuroticism, Agreeableness, Openness, and Conscientiousness will not demonstrate meaningful correlations with the video-based emotion perception task.

Cognitive Ability

Previous research has found only small relationships between performance on emotion recognition tasks and scores on cognitive ability (Mayer et al., 2002; McClure & Nowicki, 2001). While there is not a clearly specified process for how cognitive ability should influence emotion perception ability, it was anticipated that the results from previous research using static images will be replicated in the current study, even though dynamic stimuli were used.

Hypothesis 7: Cognitive ability will demonstrate small positive correlations with scores on the video-based emotion perception task.

CHAPER 4:

METHODS

The current research seeks to extend previous work in the area of emotion recognition by taking advantage of technological advancements in multimedia presentations and develop a multimedia assessment of emotion recognition. Although there have been calls by researchers investigating emotion recognition to move beyond still photographs and stimuli limited to exemplars of emotion displays (Ambadar, et al. 2005; Bruner & Taguiri,1954), few studies have used a standardized assessment of emotion perception that includes motion. Those studies that have included some form of video have either not asked about discrete emotions, or are limited to assessing only the visual channel of communicating nonverbal information.

In order to address the paucity of research on individual differences in emotion perception ability, this study attempted to develop a measure of emotional perception using dynamic and multichannel stimuli. Participants viewed brief video clips of movie scenes and were asked to rate the extent to which of several emotions were present in the scene.

The development of high quality video scenes requires good writing, competent actors, sound and film recording capabilities, a soundstage, sets, costumes, and editing facilities. This expensive and time consuming endeavor is beyond the resources of all but the most well funded research labs. In order to examine emotion perception skills within an individual differences framework, I opted to use existing film clips as the basis for an assessment of emotion perception skill. I believe the current method demonstrates an interesting intersection of three areas of human endeavors: the arts, technology, and science.

Development of the Video Scenes

Relative to Bedwell and Chuah (2007), more modern movies with better film and audio quality, as well as better scripts/writing, were reviewed by the author for scenes depicting emotions via nonverbal and verbal cues. The overarching goal was to develop an item pool large enough to adequately capture a range of emotions as well as being sufficiently diverse in terms of gender and ethnicity.

Movies were converted to a digital format that video editing software could easily work with. The video editing software used in this study was version 11 of the Pinnacle Studio software (Copyright Pinnacle Systems Software). The editing software was chosen based on availability, ease of use, and features required for the study. After the movies were converted to a format readable by the editing software, they were captured by the software and stored as a series of chapters which could be cut and spliced together as needed.

Another feature of the software was the ability to create "title slides". Although useful in more straightforward film editing processes, this feature enabled the creation of several instructional slides. These slides included basic instructions for the participants prior to viewing the film sequences, a target slide instructing the participants which individual in the upcoming scene to focus on, and a standard response slide which listed the rating instructions and the emotion options for each video scene.

Films were editing by trimming a scene pre-selected during the review process. The trimming process included removing unwanted or distracting material before and after the content of interest. Care was taken during this process so that a scene did not start or end abruptly, thus distracting the participant from the emotional content. Scenes varied in duration from approximately two seconds to 10 seconds. In addition to emotional content, suitable film

sequences were also screened on the basis of the gender and race of the focal individual in the scene as well as sound and the presence of other actors, all of which could provide emotional cues to the participant.

During the development of the video scenes, it was necessary to use more than one scene per film. In some cases, an actor appeared in multiple scenes. As a result, it is possible that the order in which the scenes were presented could have an effect on the responses to subsequent video scenes. To investigate this possibly, two random numbers were generated for each scene and the order of presentation was dictated by the random numbers. The two film presentation sequences were then reviewed to ensure that video scenes using the same actors did not appear in sequence. This was the only altering of the random order.

Half of the participants were shown the films using the order determined by the first set of random numbers and the other half of the participants were shown the films in order using the second random number sequence. Data collection took place in groups of 15 to 30 participants at a time over several months. The presentation order was specified prior to scheduling the data collection.

Target Gender

Brody and Hall (2000) note that female targets are likely to be viewed as more emotional than males, reflecting a common Western gender stereotype. In their review, these authors also found that females have been found to refer to both positive and negative emotions more than males in verbal communications and to show more nonverbal cues of emotions than men, with the exception of anger. This latter finding has been replicated using external observers. In addition, Hess et al (2004) observed that the stereotype of females being more affiliative and

males being more dominant also influenced ratings of target emotions. As a result, I expected that these stereotypes would differentially influence the perception of emotions of female targets in the video sequences. Therefore, items with a mix of male and female targets for each emotion were necessary

Sound

Communication of emotion is a multichannel phenomenon (Borod, et al., 1990, Ekman, Friesen, O'Sullivan, & Scherer, 1980). Furthermore, experimental studies have found that recognition of expressed emotion increased when multiple communication channels (visual and auditory) were used relative to a single channel of communication (de Gelder & Vroomen, 2000; Gitter et al 1972). Therefore, to increase the variance of difficulty across items within the assessment, film sequences without sound were included. It was anticipated that emotion recognition from visual only film sequences would be more difficult. Attempts were made to locate suitable clips such that the artificial removal of verbal or prosodic cues was not required. Because prosodic cues in emotion expression are often confounded with verbal content, auditory only items were not included in the current version of the assessment.

To create the video based scenes, films were screened for emotional content. Films were chosen based on release date, genre, and, to a lesser degree, the popularity of the actors. Specifically, dramatic films released in the 1990's were targeted in order to minimize familiarity of the content by the sample (most of the sample would have been about 10 years old when these movies were first released), and to increase the probability of appropriate emotional content. In addition, due to the nature of the film genre, it was felt that these films would provide a greater range of intensity in the emotional expressions.

As the emotional content of the final set of scenes needed to cover all six of Ekman and Friesen's (1971) basic emotions, the film sequences were screened specifically for the following emotional expressions: happy, surprise, anger, disgust, fear, sadness. In addition, the duration of the scenes was also a concern. Both Ambady and Rosenthal (1992) and Pessoa et al. (2005) observed that the length of exposure to a emotion stimuli has some effect on the accuracy of judgments. Therefore scenes were chosen such that the length varied from as little as 2 seconds to as long as 10 seconds. In general this was a by-product of searching for emotional content in the films, so choosing scenes primarily on length was not necessary. A total of 92 video scenes were included in the study.

Response Format

Prior research on emotion perception tended to use forced choice response formats where participants responded by choosing one emotion label (Ambady, 2002), although other researchers called for less restrictive response alternatives (Izard, 1992; Russell, 1994). Recognizing that a forced choice response format is not particularly ecological valid when assessing emotion perception, the current employed a rating scale approach. Participants were asked to make ratings on a five-point scale for each of six discrete emotions for each video scene.

Measures

Participants were asked to complete a series of assessments during the data collection phase. This included a brief demographic form, test of general mental ability, a personality inventory, an assessment of current mood, and an empathy questionnaire. After the paper

questionnaires were completed, participants viewed a series of video clips depicting various emotional expressions followed by a series of facial images depicting facial expressions of emotions. A more detailed description of each measure is presented below.

Video Based Emotion Perception

The 92 video clips described above were used as the emotion-based perception task. Each scene was preceded by a "target" screen which instructed participated to focus on a specific individual portrayed in the upcoming scene. The video scene itself was followed by a response screen which listed the six emotions participants were asked to rate. The responses were always listed in the same order. The response screen also displayed a number which corresponded to item numbers on the response sheet to minimize confusion on the part of participants when making their responses.

Personality

Personality was assessed using the Big Five Inventory (BFI; John, Donahue, & Kentle 1991). The BFI uses 44 short phrases to assess Emotional Stability, Extraversion, Openness, Agreeableness and Conscientiousness. There are 8-10 items measuring each personality dimensions on the BFI.

Cognitive Ability

The Wonderlic Personnel Test, form A (WPT; Wonderlic, 2001) was used to measure general mental ability in the current study. The WPT is a highly speeded ability test, consisting of 50 items. Participants were instructed to answer as many the items as possible in 12 minutes.

A stop watch was used to track the time and the item booklets were collected at the end of the 12 minutes, prior to moving to the next task in the research protocol. The WPT was chosen for its brief format and broad content containing items assessing verbal, quantitative, and logical reasoning skills.

Empathy

In the current study, a self-report empathy questionnaire developed by Mehrabian and Epstein (1972) was administered along with the video scenes. This questionnaire contains 33 items and has been found to relate positively to lower levels of aggression and higher levels of helping behavior. The authors explicitly state that the questionnaire is designed around a definition of empathy emphasizing a "…heightened responsive to another's emotional experience" rather than a cognitive recognition of a target's feelings.

Facial Images

Two series of facial images were presented by a projector onto a screen as part of the convergent and discriminant validity for the video-based emotion perception task. The first set of faces consisted of 21 black and white static facial images. The images displayed one of the six emotions used to select the video scenes, or a neutral expression. The faces were computer generated three dimensional images that used Ekman and Frisesn's (1976) Facial Action Coding System (FACS) as a reference (Spencer-Smith, et al., 2001). A facial model is referred to as a poser image. Each of six basic emotions were presented three times using different poser images for each presentation and three neutral images were also presented.

The second series of facial images were used to add the illusion of movement. For these 12 facial images, a neutral face was presented first, followed by a black screen for 0.5 seconds, and finally the facial image displaying one of the six emotion expressions was presented. The actual intensity of the movement was relatively small. Due to the low intensity, combined with the simulated nature of movement, each of these "dynamic images" was repeated twice providing the participants a second perceptual opportunity. These facial images were created in a similar manner to the static images, although the computer software had improved so the poser images better represented human faces. In addition, these images were presented in color.

Mood

In order to collect information on participants baseline mood levels, the Positive Affect Negative Affect Schedule (PANAS; Watson et al, 1988) was administered to participants at the start of the data collection session. The PANAS is a short assessment of positive and negative affect. Watson and his colleagues have demonstrated that instructions asking the extent to which an individual is currently experiencing positive and negative mood versus asking participants about their moods over different time periods do not adversely affect the construct validity of the PANAS. Participants in the study were asked to respond to each of the adjectives on the PANAS based on how they were currently feeling.

Participants

Usable data was collected from 388 students from the undergraduate subject pool in the psychology department at a large Midwestern university. The sample was 50.5 percent male and

primarily Caucasian (65.7 percent). The mean age was 18.9 with a standard deviation of 1.07 .

Table 1 presents the complete demographic information for this sample.

Table 1

Demographic characteristics of the research sample

	Frequency	Percent
Sex		
Male	196	50.5
Female	192	49.5
Missing	0	0
Race		
African American	19	4.9
American Indian	1	0.3
Caucasian	255	65.7
Asian	54	13.9
Other	48	12.4
Missing	11	2.8
Age		
18	151	38.9
19	145	37.4
20	62	16.0
21	19	4.9
22	6	1.5
23	1	0.3
24	0	0.0
25	1	0.3
26	1	0.3
Missing	2	0.5

CHAPTER 5:

ANALYSES AND RESULTS

Dealing with Missing Data

A total of 397 cases were collected. Of these, three participants were observed during the data collection procedure to miss viewing several of the video scenes but completed the emotion perception rating exercise regardless. These three individuals were subsequently removed from further analyses. An additional six cases were deleted due to excessive missing data. Missing more than 10 percent of the responses across all 92 scenes was defined as warranting as deletion. There were six responses per scene, thus participants had to have missed over 50 responses in order to exceed this threshold. After removing these nine cases, data from 388 individuals were analyzed.

Another 167 cases were missing at least one response on the video scenes. The majority of these (122) missed no more than six response across the 552 potential response opportunities for the video sequences. For these remaining cases with missing data, a multiple imputation procedures was used to obtain a full response data matrix. Specifically, the Two-Way imputation process was performed using an SPSS routine developed by van Ginkel and van der Ark (2008). This routine estimates the missing response by averaging the responses across the person and summing this result with the average of the responses for the item. The resulting summation is then subtracted from the average observed response across all persons and all items. An error component can also be added to the scores. A normally distributed error term is the default error component for the routine and was used in estimating the missing responses. Additionally, the imputed scores were rounded to the nearest integer.

In a simulated data set the Two-Way multiple imputation performed substantially better than a listwise deletion strategy for data missing completely at random, missing at random and not missing at random assumptions (van Ginkel, van der Ark, & Sijtsma, 2007; van Ginkel, van der Ark, & Vermunt, 2010). Given the relatively small number of participants with a large amount of missing data, and the relatively large size of the data matrix (388 x 552) the listwise deletion strategy of a subset of missing data combined with a multiple imputation procedure for the remaining cases of missing data appears to be a reasonable strategy for dealing with nonresponse by participants. Few cases were deleted so there was little reduction in power when evaluating the substantive questions in the study and multiple imputation procedures have been shown to result in less biased statistical estimates than other methods of dealing with missing data (van Ginkel et al 2007).

Scoring the Video Scenes

Because the true correct response for what emotion is displayed in any of the video sequences was not known, developing a scoring key at this early stage of development could best be described as exploratory. Four scoring approaches were used in the current study. The four methods were also included in all further analyses. Specifically, a source by method matrix defined the four scoring routines. Two sources, a subject matter expert group and a consensus group (current sample) served as the two sources. Consensus scoring procedures have been used previously with other assessments when an objective correct answer is not available (Mayer, DiPaolo, & Salovey, 1990; Mayer & Geher, 1996; Mayer et. 2003). In addition, the high agreement levels observed with early emotion recognition tasks (Ekman, 1994) support a consensus approach to scoring emotion displays.

Attempts to recruit a subject matter expert group resulted in 6 experts. The group consisted of a practicing counselor with over 20 years of experience and 5 graduate students in clinical and counseling psychology. The graduate students were all senior students who had completed at least two semesters of practicum in their program. This group was chosen to serve as the expert group as there is some evidence that this group is more sensitive to nonverbal and emotional signals (Martin et al., 2004; Rosenthal et al, 1979) compared to the typical person in the US .

Both the student sample and the expert sample were asked to provide ratings on the extent to which six emotions were present in each scene and both groups observed the same video scenes under the same conditions. However, there were two differences to the conditions for the expert group. First, due to logistics of coordinating the expert group session, the presentation of the video scenes took place in a large conference room and the experts were seated around the conference table rather than a typical classroom layout consisting of rows of desks. Second the expert group did not complete the other measures used in the study prior to viewing the video. This latter exception to the goal of identical conditions across groups was necessary due to the limited time available with the expert group.

Limited Information Method

The first pair of scoring methods employed a limited information approach. The average rating of the subject matter experts for each emotion category in a scene was used to develop a scoring key for individual scenes. Specifically, an emotion was keyed as the correct response if the mean rating for that emotion was at least a half point higher than all of the other five potential emotions. This method was termed a limited information scoring function as it only

takes into account part of the information in the set of responses to the scene and the key obtained from the experts was termed Expert Limited Information.

The limited information method for the consensus group was somewhat more complicated. There was no reasonable expectation that the mean response from the participant sample should lead to the most correct answer. Indeed, if the assumption that emotion perception skill is normally distributed in the population, then simply using the mean of that sample will result in a upper limit on the distribution of scores. In order to counter this likely effect, a concentric scoring method was added to the limited information approach described above with the expert group. After the initial mean ratings were calculated from the entire group and the emotion keys identified for each scene, a score using this key was calculated for each participant. The frequency distribution of scores was then examined and the top 25 percent of the distribution was used to create the next key. This process was continued iteratively until the score key no longer changed over successive iterations. A total of five iterations was required to reach this stopping point.

As a check on the scoring method, the distribution of scores for successive iterations was approximately normal. In addition, 75 percent of the participants who determined the score key for the second through final iterations remained the same. That is, a core group of participants emerged consistently in the top 25 percent of the score distribution over successive iterations. The limited information key obtained by the concentric process from the student sample was called Consensus Limited Information.

Mahalanobis Distance

The second pair of scoring methods was based on the Mahalanobis Distance. In contrast to the limited information scoring method described above, the Mahalanobis Distance uses the responses to all of the emotion ratings in a scene resulting in a score that captures the full information for each video seen. In addition, this method takes into account the potential dependence between those emotion ratings within a single video clip by including the covariance among the responses.

The subject matter experts' mean ratings for each emotion within a scene served as a reference point in calculating Mahalanobis Distance. Specifically, the score for each scene was computed as the deviation of a respondent's vector of ratings, **x**, for the emotions in the scene from the subject matter experts' vector of mean ratings, **u**, for the emotions in the scene,

$$(x-\mu)'\Sigma^{-1}(x-\mu),$$

where Σ is the covariance matrix of the responses. This was called the Expert Mahalanobis Distance key.

Similar to the limited information scoring method, the mean vector for the Mahalanobis Distance scoring in the consensus group was calculated using the entire sample. Again, an iterative process was used where the highest scoring 25 percent of the sample was used to recalculate the mean vector and the scores for the entire sample was recomputed. Note that the term "highest scoring" is a misnomer as Mahalanobis Distance is, as the name suggests a distance measure. Thus, scores nearest to mean would in fact represent more accurate perceptions as defined by the current scoring procedure. None the less, the term highest scoring

is used in order to be consistent with earlier terminology and hopefully lead to less confusion on the part of the reader.

During the iterative process with the consensus group, the first iteration using the restricted sample (top 25 percent) resulted in nonpositive definite covariance matrices for the majority of scenes. Thus, the full sample covariance matrix was used to compute the Mahalanobis Distance scores for all subsequent iterations in the consensus sample. A similar problem was observed for the expert group and again, the covariance matrix from the unrestricted consensus group was used to estimate scores based on the expert group Manhalanobis Distance method. The key for the Mahalanobis Distance scoring based on the concentric process for the student sample was termed Consensus Mahalanaobis Distance.

Order of Presentation

As noted earlier, the video scenes were presented in two separate, randomly ordered, sequences in order to check for possible confounds due to presentation order. Mean scores on the four scoring methods were compared across the presentation order for the video scenes. Table 2 below presents the results of the t-test and Cohen's d statistic for standardized mean differences (Cohen, 1988). There were no significant differences for order of presentation for any of the scoring methods and the effect sizes were all near zero. Consequently, the two groups were combined and used in all subsequent analyses.

Item Selection

The different scoring methods resulted in different keys for computing a score for the emotion perception task. The expert limited information method resulted in a total of 63

scoreable video scenes while the consensus limited information method resulted in a total of 60 video scenes that met the criteria for scoring described above. For the Mahalanobis Distance method, there was a total of 11 video scenes for which the covariance matrix for the six responses within a scene was not full rank, despite using the full consensus sample as described above. This was the case for 10 of the 13 videos in which "happy" was the keyed response. The remaining video for which a Mahalanbois' Distance score could not be computed was keyed as "surprise".

Table 2

Mean differences in emotion perception by order of presentation

	Order 1	n=199	Order 2	n=189		
	Mean	SD	Mean	SD	*t*	*d*
Consensus Limited	38.90	9.11	40.24	9.50	-1.42	-0.14
Expert Limited	47.67	7.50	48.05	7.68	-0.50	-0.05
Consensus MD	165.93	63.06	162.13	63.28	0.59	0.06
Expert MD	183.69	58.36	182.37	59.96	0.22	0.02

In addition, by scoring the videos using the Mahalanobis Distance method, there was no way to attribute a substantive interpretation to the score for that item. That is, it was not possible to attach an emotion label to the video based on the consensus or expert group mean ratings. This was an issue for comparing the competing factor structure models of emotion perception, described in more detail below. In order resolve this issue, only video scenes which had been identified using the limited information scoring as clearly representing a single emotion were used to create the final scores for the Mahalanobis Distance method. While this eliminated a number of items, it was necessary in order to evaluate substantive question of whether the observed emotional content influences an individuals accuracy of emotion perception.

Upon further inspection of the mean ratings for emotions within the 92 video scenes, it became apparent that the mean rating for one or two emotions were substantially higher than the remaining emotion ratings for that scene. In fact, the only reason these videos did not meet the criteria for inclusion in the limited information scoring method was due to the fact that the two highest emotions were not sufficiently distant. However, as the Mahalanobis Distance method accounts for the patterns of ratings rather than relying on the singles highest rating for the video scene, it was decided to include these videos in this scoring method. For these videos, the emotion with the highest mean rating was used to assign the video an emotion label for the purposes of fitting the two and six factor models, described below. A total of 11 additional videos meeting this criteria were identified for the expert group, resulting in a total of 64 videos comprising the Mahalabosis Distance method for experts. For the consensus group, 16 additional video scenes were identified, resulting in 63 video scenes used to compute the consensus group Mahalabosis Distance score.

Structure of Emotion Perception Ability

The three alternative models of emotion perception were evaluated separately for each scoring method. Thus a total of twelve confirmatory factor analytic models (2 methods by 2 sources by three substantive models) were evaluated. Within each of the scoring methods, video scenes were assigned an emotion label based on the emotion with the highest mean rating for that scene. For the one factor model all videos were restricted to load on the first factor. For the two factor model, representing positive affect and negative affect, videos designated as Anger, Fear, Disgust and Sad were allowed to load only on the negative affect factor and videos designated as Surprise or Happy were only allowed to load on the positive affect factor. For the six factor

model, representing a discrete emotions model of emotion perception, video scenes were allowed to load only on the factor corresponding to the label designation for that video scene, based on the mean emotion ratings. The exception to this was the limited information method for the expert key in which case there were an insufficient number of indicators for one of the discrete emotions (disgust). Thus it was not possible to conduct the six factor solution for the Expert limited information scoring method.

The data were analyzed using LISREL version 8.3 (Jöreskog & Sörbom, 1999). A correlation matrix was analyzed using a generalized least squares estimation procedure. The scale of the latent variables was standardized by standardizing the phi matrix. In accordance with best practice, multiple fit indices were used to evaluate the models (Bollen, 1989, Bollen & Long, 1993, Jackson, Gillaspy, & Purc-Stephenson, 2009). Table 3 presents a summary of several of the overall fit statistics for each model and scoring method below.

Examining the overall fit indices below, the models appear to be well specified, although several of the overall fit indices did not quite reach the conventional rules of thumb that have been proposed for evaluating the fit of a model (Hu and Bentler, 1999). For example, the results of the Goodness of Fit Indicator (GFI) and the Adjusted Goodness of Fit Indicator (AGFI) for all of the models fell somewhat short of the recommended values for some rules of thumb (e.g., .90 or .95), while the RMSEA and SMR approach previously recommended acceptance cutoffs (Lance & Vanderberg, 2002). In addition, the chi square divided by the degrees of freedom for the model are all quite small. Indeed, these values are below the more rigorous rule of thumb value of 2.0 and are quite close to the expected value for the chi square. However, as other authors have noted, there have been multiple recommendations for evaluating the indices, and these have at times been contradictory. In addition, it is important to keep in mind that these

indices are descriptive of overall model fit and do not provide a way to statistically compare the fit across models (although see MacCallum, Browne & Cai, 2006 for proposed power analyses and effect sizes).

Table 3

Overall fit indices for the one, two and six factor confirmatory factor analyses

	χ2	df	χ^2/df	RMSEA	RMR	Std RMR	GFI	AGFI
Consensus Limited								
1 Factor	1824.52	1595	1.14	0.09	0.08	0.11	0.84	0.83
2 Factor	1802.85	1594	1.13	0.09	0.08	0.11	0.84	0.83
6 Factor	1750.91	1580	1.11	0.08	0.08	0.11	0.84	0.83
Expert Limited								
1 Factor	2356.09	1890	1.25	0.10	0.09	0.13	0.79	0.76
2 Factor	2348.47	1889	1.24	0.10	0.09	0.14	0.81	0.79
6 Factor	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Consensus MD								
1 Factor	2706.98	1890	1.43	0.13	0.07	0.1	0.78	0.73
2 Factor	2700.45	1889	1.43	0.13	0.07	0.1	0.78	0.76
6 Factor	2617.92	1875	1.40	0.12	0.07	0.11	0.79	0.77
Expert MD								
1 Factor	2753.21	1952	1.41	0.13	0.07	0.10	0.78	0.76
2 Factor	2538.96	1951	1.30	0.11	0.07	0.10	0.80	0.78
6 Factor	2485.89	1937	1.28	0.11	0.07	0.10	0.80	0.78

RMSEA = Root Mean Square of Approximation, RMR = Root Mean Residual, SRMR = Standardized Root Mean Residual, GFI = Goodness of Fit Indicator, AGFI = Adjusted Goodness of Fit Indicator.

More complex models will generally result in better estimates of fit, simply because of the additional number of parameters to be estimated. As result, both the degree of fit for the model, as well as its parsimony should be taken into account (Cudeck & Browne, 1993). This is particularly true when comparing different substantive models with the goal of identifying the best fitting model, as is the case here. For nested models, the difference in chi square between the more restricted model and the unrestricted model is itself distributed as a chi square with degrees of freedom equal to the difference in degrees of freedom between the two models. Table 4 presents these results for the models described earlier. At alpha equal to .05, the difference in the

chi squares for the one and two factor models is significant for all of the scoring methods. Likewise, the difference in the chi-squares for the two and six factor models also exceeds the critical chi square value at alpha = .05 with 14 degrees of freedom for all of the scoring methods (excepting the Expert limited information method). Comparing the overall fit of the models using statistical significance testing supports using the six factor discrete emotion model.

Table 4

Chi Square test for nested models.

	χ^2 difference	df difference	Critical χ^2
Consensus Limited			
1 Factor vs 2 Factor	21.67	1	3.84
2 factor vs 6 factor	51.34	14	23.68
Expert Limited			
1 Factor vs 2 Factor	7.62	1	3.84
2 factor vs 6 factor	N/A		
Consensus MD			
1 Factor vs 2 Factor	6.53	1	3.84
2 factor vs 6 factor	82.53	14	23.68
Expert MD			
1 Factor vs 2 Factor	214.25	1	3.84
2 factor vs 6 factor	53.07	14	23.68

Comparing alternative models from the overall fit indices results in some confusion as to which models are more appropriate. In the current study, the descriptive overall indices of model fit suggest little improvement in fit for the more complex models over simpler models. However, the statistical test of the chi square values suggests a significant difference in the fit of the models. For this reason, Bollen (1989), among others recommends evaluating the components of the models under consideration.

The squared multiple correlation for a variable in the model provides a measure of the strength of the linear relationship between the variable and the latent factor (Jöreskog & Sörbom,

1996). Given the complexity of the models and the number of models examining the current study, the means and standard deviations of squared multiple correlations were used to summarize the large number of correlations. The means and standard deviations for the models are presented in Table 5 below. As can be seen from the table, the mean of the squared multiple correlation increased very little for a given scoring method as the complexity of the models increased. This pattern of correlations, combined with the overall fit indices above and preference for parsimony, supports a single emotion perception ability underlying judgments of the six emotions. Therefore a single score for each scoring method was used in further analyses.

Table 5

Means and standard deviations for the squared multiple correlations of the confirmatory factor analyses.

	Consensus Limited		Expert Limited		Consensus MD		Expert MD	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	.175	.084	.136	.074	.479	.114	.432	.129
2	.175	.087	.134	.072	.481	.113	.434	.128
6	.200	.097	NA	NA	.488	.109	.444	.122

The means, standard deviations, and reliability estimates for all of the measures are presented in Table 6 below. As Table 6 shows, reliability estimates for all four scoring methods for the video based assessment were above .80. Item statistics for the video scenes are presented in Appendix A for all four scoring methods.

Correlations among the four scoring methods for the video-based emotion perception task are presented in Table 7. As can be seen from Table 7, the pattern of correlations demonstrates a method effect such that the two pairs of scoring methods were nearly perfectly related, regardless of whether the consensus or the expert group generated the scoring key. Across methods, the keys produce scores that were only modestly correlated. Other authors investigating the use of

consensus and expert groups for scoring have also observed relatively little difference in the resulting scores across the two groups (Mayer, Salovey &Caruso, 2002), although early attempts to use both consensus and expert groups were less successful (Matthew, Zeidner, & Roberts, 2002).

Table 6

Means and standard deviations for the variables

	Mean	Standard Deviation	Reliability
Consensus Limited Information	38.41	9.27	.88
Expert Limited Information	47.86	7.58	.83
Consensus MD	164.08	63.12	.97
Expert MD	183.05	59.07	.96
Wonderlic Personnel Test	25.53	5.05	n/a
Extraversion	27.30	6.22	.85
Agreeable	34.17	5.75	.79
Conscientious	31.46	6.07	.81
Neuroticism	21.62	6.25	.83
Openness	35.72	6.53	.79
Empathy	104.93	11.93	.76
PA	24.83	6.81	.84
NA	14.39	5.41	.86
Static Facial Images	14.04	2.20	.50
Dynamic Facial Images	6.24	1.58	.03

Table 7

Correlations among scoring methods for a single dimension of emotion perception

	Consensus Limited	Expert Limited	Consensus MD	Expert MD
Consensus Limited				
Experts Limited	.91**			
Consensus MD	.40**	.43**		
Expert MD	.42**	.43**	.99**	

Gender Differences

Hypothesis 1 a: Females will be more accurate than males on the emotion perception tasks

Partial support was observed for hypothesis one. Females scored higher on the limited information scoring across both the consensus and expert score keys at $p<.05$. No significant differences were observed for the consensus or expert key using the Mahalanobis Distance scores. Table 8 below presents the results of the t-test and Cohen's d statistic for standardized mean differences. Although the gender difference for emotion perception was significant for both the consensus and expert limited information scoring methods, the effect sizes (Cohen's *d*) were relatively modest. For the expert limited information scoring method, the effect size was below .30, which Cohen characterized as a small effect.

Table 8

Gender differences in emotion perception

	Male (n=196)		Female (n=192)			
	Mean	SD	Mean	SD	*t*	*d*
Consensus Limited	37.93	10.07	41.22	8.17	-3.53*	-0.36
Expert Limited	47.09	8.18	48.64	6.86	-2.03*	-0.21
Consensus MD	164.84	65.47	163.30	60.78	0.24	0.02
Expert MD	183.09	60.59	183.02	57.64	0.01	0.00

Hypothesis 1b: Females will make higher ratings for scenes with female targets experiencing emotions incongruent with western gender stereotypes .

There were a limited number of video sequences that contained a female target displaying anger, the emotion incongruent with western stereotypes of affiliation used in previous research. Table 9 below presents the results of an independent samples t-test and the *d* statistic comparing the ratings of anger in the four videos in which female targets displayed anger (according to the expert and consensus group means). The results supported hypothesis 1b. Females did in fact rate

these scenes higher, indicating they perceived anger in the video scene to a greater extent than did males.

Table 9

Gender Differences in Ratings of Western Stereotype Incongruent Emotions

	Male (n=196)		Female (n=192)			
	Mean	SD	Mean	SD	t	d
Raw Responses	16.68	1.88	17.34	1.83	-3.48*	-0.36

Minority Status

Hypothesis 2: Minority participants will score higher than Caucasian participants on the emotion perception task.

Hypothesis two was not supported. There were no significant differences between minority and nonminority participants on scores of emotional perception. This was true across all scoring methods (consensus vs expert and limited information vs. Mahalanobis Distance). Table 10 below presents the results of an independent samples t-test and the *d* statistic comparing the minority group and nonminority group for all four scoring methods. In this analysis, the minority group was comprised of individuals who responded that they were either not Caucasian or marked two ethnic categories. As can be seen from Table 10 the standardized mean difference index (Cohen's *d*) for the groups was near zero for all of the scoring methods.

As Table 1 above shows, of the minority participants in the research, only those participants who identified themselves as Asian comprised a sufficiently large sample to conduct a t-test for differences between emotion perception accuracy. Table 11 presents the results of a one tailed independent sample t-test. The results agree with the analysis of the combined

minority group showing no significant differences. The effect sizes associated with the analyses comparing Asians and Caucasians demonstrated small differences.

Table 10

Differences by minority status on emotion perception accuracy

	Minority (n=122)		Nonminority (n=255)			
	Mean	SD	Mean	SD	*t*	*d*
Consensus Limited	39.17	8.55	39.76	9.69	-0.58	-0.06
Expert Limited	47.66	7.39	47.99	7.69	-0.40	-0.04
Consensus MD	164.85	60.51	165.08	65.13	-0.03	0.00
Expert MD	182.87	56.44	184.47	60.99	-0.25	-0.03

Table 11

Mean differences in emotion perception accuracy between Caucasian and Asian participants

	Asian (n=54)		Caucasian (n=255)			
	Mean	SD	Mean	SD	*t*	*d*
Consensus Limited	37.72	8.88	39.76	9.69	-1.43	-0.22
Expert Limited	46.33	8.44	47.99	7.69	-1.41	-0.21
Consensus MD	155.65	46.64	165.08	65.13	-1.01	-0.16
Expert MD	171.01	43.12	184.47	60.99	-1.31	-0.24

Current Mood

Hypothesis 3 a: Individuals with higher scores on positive affect will rate scenes with positive emotions higher than individuals with low scores on positive affect.

Using the positive affect scores from the PANAS, the sample was split into the upper and bottom quartile for comparison. Ratings on positive emotions form the video scenes keyed as Happy or Surprise were summed and then compared for individuals scoring in the bottom and top quartile of the sample. This was done using the consensus group score key, the expert group score key and for only those videos in which both groups agreed in terms of the emotion with the highest mean rating ("universal key" in Table 12 below). There were 91 individuals in the

bottom quartile and 98 individuals in the upper quartile. A one-tailed independent samples t-test resulted in significant differences for the high and low positive affect groups at $p<.05$ in the predicted direction with a small effect size.

Table 12

Effect of positive mood on ratings of positive emotion scenes

	Low Positive Affect (n=91)		High Positive Affect (n=98)			
	Mean	SD	Mean	SD	*t*	*d*
Consensus Key	106.23	12.53	109.82	12.80	-1.94*	-0.28
Expert Key	103.19	11.81	106.49	12.25	-1.89*	-0.28
Universal Key	94.93	10.79	97.92	10.75	-1.9	-0.28

Hypothesis 3b: Individuals with higher scores on negative affect will rate scenes with negative emotions higher than individuals with low scores on negative affect.

Similar to the analysis for positive affect, negative affect scores on the PANAS were separated into upper and lower quartiles. There were 78 individuals in the bottom quartile and 105 individuals in the upper quartile. The ratings on video scenes keyed for negative emotions (Anger, fear, disgust, and sadness) were summed and the result was compared against the lower and upper quartiles. This was done using the consensus group score key, the expert group score key and for only those videos in which both groups agreed in terms of the emotion with the highest mean rating ("universal key" in Table 13 below). A one-tailed independent samples t-test resulted in significant differences for the high and low negative affect groups at $p<.05$ for the consensus and expert keys. An examination of Cohen's effect size for the standardized mean difference suggests that any practical differences were relatively small for both the consensus group and expert group scoring keys.

Table 13

Effect of negative mood on ratings of negative emotion scenes

	Low Negative Affect (n=73)		High Negative Affect (n=105)			
	Mean	SD	Mean	SD	*t*	*d*
Consensus Key	116.64	13.12	120.13	14.58	-1.64*	-0.25
Expert Key	125.44	14.29	129.80	15.64	-1.90*	-0.29
Universal Key	96.07	10.46	97.51	11.02	-0.88	-0.13

Convergent and Discriminant Validity

Table 14 presents the correlations among the criterion measures. Internal consistency reliability estimates are presented in Table 6 above, except for the WPT as internal consistency is not an appropriate measure for highly speeded tests. Test-retest reliabilities for the WPT are reported to have ranged from .82 to .94 (Wonderlic, 1992). Table 15 presents the correlations between the other individual difference constructs and the four scoring methods for the video-based emotion perception task.

Hypothesis 4 a: Scores on the video based emotion perception task will demonstrate moderate positive correlations with responses to static facial images posing emotional expressions.

Hypothesis four was partially supported. Responses to the static facial images were not significantly correlated with the either the consensus or expert limited information scoring methods. Responses to the static facial images were significantly correlated in the expected direction with the consensus and expert Mahalanobis Distance scoring methods, however, these relationships were fairly small.

Hypothesis 4b: Responses to the dynamic facial images will demonstrate higher correlations with the video-based emotion perception task than will responses to the static facial images.

This hypothesis was not supported. Responses to the dynamic facial images were not significantly correlated with the either the consensus or expert limited information scoring methods. While responses to the dynamic facial images were significantly correlated in the expected direction with the consensus and expert Mahalanobis Distance scoring methods, correlations of the same magnitude were obtained with the static faces measure.

Hypothesis 5: Empathy will be moderately correlated with scores on the video-based emotion perception task.

Hypothesis 5 was partially supported. Empathy was significantly related to the consensus and expert limited information scoring methods, but not with either of the Mahalanobis Distance scoring methods. The correlations with the limited information scoring methods were relatively small.

Hypothesis 6: Scores on the personality factors of Extraversion, Neuroticism, Agreeableness, Openness, and Conscientiousness will not demonstrate meaningful correlations with the video-based emotion perception task.

Hypothesis 6 was generally supported. The Big Five personality factors did not demonstrate a pattern of meaningful correlations. Extraversion was significantly related to the limited information scoring methods for both the consensus and expert group, however, these correlations were relatively small and Extraversion was not correlated with the Mahalanobis Distance scores. One correlation for Agreeableness and Neuroticism was significant with the expert Mahalanobis Distance. Openness and Conscientiousness were not significantly correlated with any of the emotion perception variables. The largest of the personality correlations with any

of the video based emotion perception task, was .20, which is small enough to indicate that the Big Five personality factors measure something different than emotion perception skill.

Hypothesis7: Cognitive ability will demonstrate small positive correlations with scores on the video-based emotion perception task and cognitive ability.

Hypothesis 7 was not supported. Cognitively ability was not significantly correlated with any of the video based emotion perception scoring methods.

Table 14

Correlations among criterion measures

	Static	Dyn	EMP	E	A	C	N	O	WPT	PA	NA
Static											
Dyn	.19*										
EMP	.13*	.08									
E	.09	.06	.23*								
A	.03	.14*	.30*	.23*							
C	-.03	.10	.02	.23*	.29*						
N	-.01	-.05	.31*	-.10*	-.33*	-.14*					
O	.10	-.03	.05	.14*	.10	.07	-.09				
WPT	.01	-.03	-.05	-.02	.03	-.01	-.07	-.11*			
PA	-.10	.09	.05	.20*	.25*	.30*	-.08	.19*	.02		
NA	-.09	-.07	.08	.01	-.18*	-.12*	.42*	-.05	.00	.09	

* p< .05; WPT=Wonderlic Personnel Test, E = Extraversion, A=Agreeableness, C=Conscientiousness, N=Neuroticism, O=Openness, PA = PANAS Positive Affect, NA=PANAS Negative Affect, EMP=Empathy, Static=Static Facial Images, Dyn=Dynamic Facial Images.

Table 15

Correlations between video based emotion perception and external variables

	Consensus Limited Information	Expert Limited Information	Consensus MD	Expert MD
Static Faces	.05	.04	-.13*	-.11*
Dynamic Faces	.05	.06	-.13*	-.14*
Empathy	.13*	.13*	.10	-.01
Extraversion	.13*	.13*	.07	.00
Agreeableness	.08	.02	-.02	-.12*
Conscientiousness	.08	.07	.03	-.03
Neuroticism	.08	.08	.09	.20*
Openness	.03	.04	-.00	-.05
WPT	-.02	-.02	-.05	.04

CHAPTER 6:

DISCUSSION

Salovey and Mayer's (1990; Mayer & Salovey,1997) theory of emotional intelligence asserts that individuals differ in their ability to process information about emotions. Specifically, their model suggests 4 factors: perceiving emotions, using emotions, understanding emotions, and regulating emotions. They further suggest that these four factors are hierarchically ordered such that perceiving emotions is the most basic process of emotional intelligence with emotion regulation being the most complex process in their model. These authors subsequently developed an assessment to assess the construct, and evaluated its validity, both in how it fits with the nomological network and also whether it has any use in applied contexts (Mayer, Salovey & Caruso, 2002). The resulting test, the Mayer, Salovey, and Caruso Emotional Intelligence Test (MSCEIT) has the distinction of being the only performance-based measure specifically developed to assess emotional intelligence. The current study sought to extend the assessment of the first branch in Mayer and Salovey's (1997) model, perceiving emotions, using video based emotion scenes and asking respondents to evaluate multiple emotions for each scene.

Early assessments of individuals emotion perception focused on whether certain emotional expressions were recognized across disparate cultures (Ekman & Friesen, 1971). The results were overwhelmingly positive. Individuals from vastly disparate cultures recognized the same emotional expressions as indicating the same emotions with agreements rates greater than 90 percent for some emotions (e.g., Happy). The current research suggests that although such high agreement rates are possible using prototypical examples of expressions, more complex

cues reveal somewhat large individual differences in emotion perception, thus providing support that emotion perception is a skill that distinguishes between individuals.

Four scoring methods were examined in the current study. First, two scoring keys were developed using a limited information procedure, which accounted for only the single emotion that best characterized the cues from the target in the scene. This method was used with both a group of individuals with 6 -20 years of training and experience in therapeutic counseling who served as the expert group, and the current sample of undergraduate students in psychology who comprised a consensus sample. For the consensus sample, an iterative method was applied in which the highest scoring 25 percent from the initial key were used to develop a subsequent key. The iterative process continued using individuals in the top 25 percent of the score distribution from the previous key, until no changes in the key occurred (i.e. the key converged).

The remaining scoring keys employed a Mahalanobis Distance approach in order to account for the ratings for each of the six emotions within a video. The expert group and the consensus group were used to develop separate keys using the distance methodology. Previous criticisms of the consensus approach centered around the observation that score patterns from consensus samples were dissimilar to those of expert scoring keys. This was not the case in the current study. Indeed, a method effect was observed in which the two limited information scores were highly related and the two Mahalanobis Distance scores were highly related. Relationships within the same source group (consensus and expert) were only moderated related across the limited information and Mahalanobis Distance methods.

The current method of using consensus to develop scoring keys for emotion perception differs from previous methods in two potentially important ways. Previous methods have used proportional scoring to develop the keys. In other words, a weight is applied to each response

option corresponding to the proportion of the scoring sample that chose the option. If, for example, 35 percent of the sample chose a particular response, then that response is worth .35 points in the scoring system. The rationale behind this approach is that emotion perception relies, in part at least, on social conventions for what a particular emotional expression is intended to convey. However, this rationale appears to either not hold up very well, or implies that individuals recognized as experts do not in fact know better than the lay person what emotional cues convey. If the latter implication is in fact the case, then an expert group would not be needed at all. Of course, it is possible that the definition of experts and selection into the expert group may be faulty

The current approach also differs from previous attempts at consensus scoring by using an iterative procedure. The rationale behind using iterations is that if there are indeed individual differences in emotion perception, then individuals who are truly more accurate in their emotional perceptions would be obfuscated by the larger proportion of individuals in the average ability range when developing the score key. This assumption was supported in the current sample by a core group of individuals who consistently appeared in the upper 25 percent of the score distributions of the limited information method across successive iterations.

In evaluating the factor structure of the scoring methods in the current study, three alternative models were posited and compared; a single factor model, a two factor model representing positive and negative affect, and six factor discrete emotion model. The latter two models reflected the possibility that the nature of the emotion under consideration may influence the rank order of people's accuracy of identification. That is, are some individuals better than the rest of the population at recognizing fear in others, but would not excel to such a degree in recognizing anger or happiness? Results from the confirmatory factor analyses suggest that there

is little information to be gained by a content driven theory of accuracy of emotion perception. The overall fit indices of the models do not demonstrate marked improved as the complexity of the model increases. Although the chi square test for nested models was significant, an examination of the squared multiple correlations for the items were consistent with the overall fit indices supporting parsimony over complexity.

In addition, an interesting pattern was observed when examining the squared multiple correlations presented in Table 5. Within a given scoring method, increases in model complexity did not improve the prediction of the indicator variables in the model (responses to the video scenes) as indicted by the squared multiple correlations. However, looking across the scoring methods for all three factor models reveals that the mean squared multiple correlations were substantially larger for the consensus and expert Mahalanobis Distance methods. While a general trend of this sort might be expected as the Mahalanobis Distance method accounts for more information in the response patterns than the limited information method, the size of the distance is striking, particularly for item level data. An interesting question then arises as to whether the Mahalanobis Distance scoring paradigm would result in superior prediction of individual behavior relative to the limited information scoring procedure.

The research on gender differences in emotion perception has generally reported that females are more accurate than males (Elfenbein & Ambady, 2002, Hall & Matsumoto, 2004). This finding was only partly replicated in the current study. Specifically, females scored higher on the emotion perception task for the limited information scoring method for both the consensus and expert keys. This was not the case for the Mahalanobis Distance method, for either the consensus or expert scoring keys. It is not immediately clear why this would the be the case and future research should examine why the longstanding observed female superiority for emotion

perception disappears when ratings on multiple emotions for a given scene were taken into account.

Results from previous research on race, minority status and ingroup/outgroup effects on emotion perception have not formed a clear picture. In the current study, there was an insufficient number of individuals from most of the non-Caucasian ethnic group to conduct analyses on specific groups. Furthermore, almost a third of participants who reported an ethnic category that was not Caucasian used the "other" category. Therefore the ethnic categories were collapsed into two groups, minority and nonminority. The results from the current study suggest that minority group status does not have an effect on accuracy of emotion perception. Although previous research has reported a slight trend for individuals in a minority cultural group to demonstrate better accuracy for emotion perception as compared to the majority group, this was not replicated in the current study. Not only were there no significant differences, but effect size estimates were very close to zero.

There are a number of possible explanations for the lack of clear cut findings on ingroup/outgroup effects in emotion perception. For this discussion, ingroup and outgroup encompass ethnic and cultural groups as well as the broader category of minority status. One notable issue that occurs throughout the literature on emotion perception is the lack of a careful definition of the outgroup. In the current study, this was comprised of all participants who chose an ethnic category other than Caucasian. Although this categorization was based on necessity due to small samples, this definition implicitly assumes that all non-Cauasian groups should be treated the same. Such an assumption may obscure results for specific minority groups within that definition. This interpretation is partially supported by the results examining differences between Caucasian and Asian samples. The differences were not statistically significant,

however, the effect size for all for scoring method displayed a small difference. Interestingly, the direction of the effect was opposite of previous findings. Caucasians tended to score higher on the video-based assessment of emotion perception. Future research examining ethnic or cultural differences on emotion perception should pay careful attention to the categorization scheme employed to ensure there are no confounding explanations for observed differences, or a lack thereof.

Previous research has observed a general mood congruent effect whereby an individual's current mood influences the judgments he or she subsequently makes on topics not related to current mood (Mayer et al., 1992). To ensure that respondents were not looking through rose-colored glasses as they made judgments regarding the emotions in the video scenes, a mood assessment was administered prior to viewing the video scenes. The mood congruent effect observed in earlier research was observed in the current study. Individuals scoring higher on positive affect rated positive emotion scenes higher than did individuals low on positive affect. Likewise, individuals who were higher in negative affect rated video scenes displaying negative emotions higher than individuals low in negative affect. The current results extend previous findings by examining the effect of current mood on ratings of observed emotion expressions. Taken together with previous research, these findings suggest future researchers examining emotion perception, emotion regulation, and emotional experiences are well advised to collect baseline mood data as control for potential confounds.

The emotion perception task was not strongly related to any of the Big Five personality traits or cognitive ability. This finding is consistent with prior research employing task based measures of emotion perception (Matthews et al (2002). While the lack of correlations with personality factors supports the assertion that emotion perception, as a component of emotional

intelligence has the potential to provide additional information to help understand, explain, and predict an individual's behavior, the lack of a correlation with cognitive ability does not align with the theory of emotional intelligence as a cognitive ability similar to traditional measures of intelligence. It is possible that the lack of a relationship in the current study may be result of using a speeded test which emphasizes crystallized intelligence over fluid intelligence. Additional research comparing the relationship between emotion perception and measures of both crystallized and fluid intelligence would further our understanding of this issue. Furthermore, the students at the large Midwestern university were highly selected, resulting in substantial restriction of range.

The video-based emotion perception task did not demonstrate any meaningful relationship to either the static facial images or the dynamic facial images. Given that the nature of the tasks are relatively similar, this result is puzzling. Examining the last column of Table 6 reveals that the internal consistency estimates are quite low for the static and dynamic faces tasks. The dynamic faces task in particular pose a problem as the reliability estimate for scores on this variable were essentially zero. The low reliability estimates prompted a re-examination of the scoring procedures for these two tasks. Although the scoring methods were found to be accurate, a close examination of the item properties for these stimuli revealed some interesting patterns. Examining the properties of the static facial images, the distribution of item means were severely negatively skewed. The process used to score responses to the facial images resulted in dichotomous item scores. For 11 of the 18 scored facial images (three facial images presented in the study were neutral), the proportion correct in the current sample exceeded .75., indicating that the emotions displayed in this set of static facial images were readily identifiable for the majority of participants in the current sample. With such a large proportion of the sample

correctly identifying the emotions in the static facial images, a ceiling effect results and the images do not discriminate well among the sample in terms of emotion perception skill.

Examining the item means for the 12 dynamic facial images indicated more spread in the proportion correct, relative to the static images. However, the intercorrelations among the facial images revealed very small relationships among the items. When focusing only on the facial images representing the same emotion, one of which was female and the other male, the correlations between the two images was still near zero, or even negative, for all six emotions. Given the lack of internal consistency in the current data set for these stimuli, interpretations for the lack of relationship with the video-based emotion perception task are difficult at best.

Every effort was made to include other emotion perception tests in the current research. The emotion perception branch of the MSCEIT would have been ideal. Unfortunately the individual branch tasks cannot be decoupled from the test as a whole. Further, the MSCEIT could not be administered in its entirety due to time limitations. Other published emotion perception measures were also sought but due to logistical concerns, could not be included in the current study. Future research using a paired down version of the video-based emotion perception task (63 items versus 92) could further our understanding of how high-fidelity assessments compare to more static assessments of emotion perception.

Implications for Research

A holistic approach to capturing emotion perception seeks to address a different set of questions relative to research framework grounded in a reductionist approach. Although a reductionist approach is important for determining the specific elements of how individuals recognize emotions and label them, the fundamental process in daily life is more than a simple

sum of its parts. For example, in order to isolate the components of the emotion perception process, certain variables must be controlled or restricted. However, in the process of recognizing, and subsequently reacting to, the emotional cues from our family, friends, colleagues and strangers, these restrictions are noticeably absent. We are typically not limited to certain subsets of emotions or emotion cues. Moreover the time frame in which all of these perceptual activities takes place is highly fluid. Individuals may have less than a second to recognize and react to the emotional cues from others. Finally, although knowing that the actual words used to convey the content of the message may impair the accuracy of emotion recognition is useful in understanding emotion perception (Ekman, 2008), this finding doesn't alter the reality that individuals do receive information from multiple channels and often need to attend to and integrate information communicated across all of the channels simultaneously.

Implications for Practice

Employee selection procedures have relied on paper and pencil assessments for the last 50 years, although new technologies are beginning to find their place in organizational selection methods (Lievens, Buyse, & Sacket, 2005). Recent research on video-based situational judgment tests (SJT's) suggests that, at least for interpersonal criteria, video based assessments demonstrate higher validity than equivalent written SJT's (Lievens & Sackett, 2006). Although criterion related research is still needed for the current assessment, this line of research suggests that video based assessments can improve an organization's ability to hire the more qualified candidates. Still, as Lievens and his colleagues demonstrate, video-based assessments are not a panacea and practitioners should pay careful attention to matching the predictor construct with the criterion of interest.

In practical contexts, the test taker reaction is often important, particularly in employee selection contexts. For highly skilled jobs, the applicant pool is often much smaller and selling the organization to the applicant can become a much bigger component of the job. At least in laboratory experiments, SJT's have been viewed more favorably than written tests (Chan & Schmitt, 1997; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000), which may help influence highly sought after candidates to join the organization. Finally, SJT's have been observed to result in less adverse impact relative to written tests (Chan & Schmitt, 1997). Given the increasing scrutiny on testing programs in employee selection procedures by enforcement agencies (Lundquist & Ashe, 2010), the combination of higher validity and less adverse impact would be extremely attractive to organizations.

Limitations and Future Research

As with any research project, the current study has a number of limitations. First, the sample is composed of undergraduate psychology students. The undergraduate survey course in psychology is one of the more popular courses at the university where this research was conducted. Despite such a diverse pool of potential research participants, it is unlikely that undergraduate students taking this course are representative of the population. Future research would benefit from using more diverse samples.

Second, the current research involved the use of existing film as source material for the film clips that comprised the emotion perception stimuli. Although efforts were made to choose scenes with which the participants would not be familiar, it is possible that some participants were "movie buffs" and were therefore more familiar with the clips. This familiarity with the scenes, and the larger story behind the scenes presented in this research may have given them an

advantage. However, a survey of the smaller group of experts revealed that they were not familiar with most of the scenes in the research, suggesting they the expert group ratings were not influenced by familiarity. The strong relationship between the consensus group scores and the scores from expert group key, suggest that familiarity with the film content is unlikely to have unduly influenced the results of the current study.

Three related issues also involve the use of the existing film as source material. First, using existing material results in a limitation over control of the stimuli. Although a variety of films and actors were included in the clips, it would have been desirable to develop scenes tailored to a given emotion and a context. Although limited financial resources precluded the development of such scenes, the results presented here support the viability of such an endeavor. Developing video scenes specifically for a test of emotion perception would also provide another benefit. The diversity of actors could be greatly expanded. Although minority actors appear in current films with some degree of regularity, locating scenes including minority actors under the criteria described above was a difficult task. This was particularly the case for Hispanic and Asian actors. Another potential benefit would be to film the same scene multiple times using diverse actors each time to investigate more fully the influence of target diversity as well as respondent by target diversity issues in emotion perception. Finally, using existing films necessarily meant that intended emotion for a specific scene could not be known a priori. Although a careful analysis of the larger context might grant some clues, without talking to the screenwriter, director, or actor directly, this information was simply not available. Developing the scenes specifically would grant control over the writing and production of the films thereby providing a third source of information for scoring video scenes.

Response Format

The current research incorporated only six emotions on which to make ratings. In part this was done to be consistent with previous research demonstrating the universal nature of this six basic emotions (Ekman & Friesen, 1976). Clearly, however, there are other emotions that could be included. In addition, if the video scenes were written specifically for the assessment, it would be possible to portray multiple emotions simultaneously. Commonly used terms in everyday conversations such as "conflicting emotions" and "bittersweet sorrow" provide a rationale that any given emotion is not mutually exclusive to all other emotional experiences. In particular, the Mahalanobis Distance scoring method would be especially appropriate as a scoring algorithm for scenes including more than one emotion as the target.

Conclusion

This research suggests that high fidelity video based tests of emotion perception have much to offer. Initial results suggest it would be worth the time, effort, and expense to develop realistic emotion vignettes for film in order to move the research further. As mentioned earlier, developing video scenes explicitly for such an assessment offer a significant advantage in terms of control and would allow more refined investigations of emotion perception.

REFERENCES

Adams, R.B., & Kleck, R.E. (2005). Effects of direct and averted eye gaze on the perception of facially communicated emotion. *Emotion, 5*, 3-11.

Ambadar, Z, Schooler, J.W., & Cohn, J.F. (2005). Deciphering the enigmatic face: The importance of dynamics in interpreting subtle facial expressions. *Psychological Science, 16,* 5, 403-410.

Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology, 77*, 538-547.

Ambady, N. & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*, 256-274.

Arbib, M.A., & Fellows, J. (2004). Emotions: from brain to robot. *Trends in Cognitive Science, 8,* 12, 554-561.

Ashforth, B. E., & Humphrey, R. H. (1995). Emotions in the workplace: A reappraisal. *Human Relations, 48*(2), 97-125.

Bar-On, R. (1997). *Bar-On Emotional Quotient Inventory: User's manual.* Toronto, Ontario, Canada: Multi-Health Systems.

Bassili, J. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology, 37*, 2049-2058.

Bedwell, S. (2003) *Emotional Judgment Inventory Manual*. Champaign, IL: Institute for Personality and Ability Testing.

Bedwell, S. & Chuah, S.C. (2007). Video-based assessment of emotion perception: toward high fidelity. In M.E. Bergman and J.L. Rasmussen (Chairs) *Situational Judgment Tests: Future Directions.* Symposium at the annual meeting of the Society of Industrial and Organizational Psychology, April, New York, NY.

Bernieri, F.J. (2001). Toward a taxonomy of interpersonal sensitivity. In J.A. Hall & F.J. Bernieri (eds.) *Interpersonal Sensitivity: Theory and Measurement.* Mahwah, NJ: Earlbaum.

Bollen, K.A. (1983). Structural Equations with Latent Variables. New York, NY : Wiley.

Bollen, K.A., & Long, J.S. (Eds, 1993). *Testing Structural Equation Models.* Newbury Park, CA: Sage.

Borod, J.C., Koff, E., Yecker, S., Santschi, C., Schmidt, J.M. (1998). Facial asymmetry during emotional expression: Gender, valance, and measurement technique. *Neuropsychologia, 36*, 1209-1215.

Borod, J.C., Pick, L.H., Hall, S.,, Slinwinski, Madigan, N., Obler, L.K., Welkowitz, J., Canino, E., Erhan, H.M., Goral, M., Morrison, C. & Tabert, M. (2000). Relationships among facial, prosodic, & lexical channels of emotional perceptual processing. *Cognition and Emotion, 14*, 193-211.

Brody, L.R., & Hall, J. (2000). Gender, Emotion, and Expression. In M. Lewis and J. M. Haviland-Jones (Eds.) *Handbook of Emotions, 2nd Edition.* New York: Guilford Press.

Browne, M.W. & Cudeck, R. (1993).Alternative ways of assessing model fit. In K.A. Bollen and J.S. Long (Eds.) *Testing Structural Equation Models*. Newbury Park, CA: Sage.

Bruner, J.S., & Taguiri, R. (1954). The perception of people. In G. Lindzey (ed.) *Handbook of Social Psychology (vol 2)*. Reading, MA: Addison-Wesley.

Byron, K., Terranova, S., & Nowicki, S. (2005). *Are more successful salespersons better able to "read" emotions?* Paper presented at the 20th Annual Meeting of the Society for Industrial and Organizational Psychology. April, Los Angeles CA.

Campbell, D.T. & Diske, D.W. (1959). Convergent and discriminant validation by the multitrait -multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 142-159.

Cohen, J, (1988). *Statistical Power Analyses for the Behavioral Sciences*, 2nd Edition. Hillsdale, NJ; Earlbaum.

Cronbach, L.J., Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Darwin, C. (1965). *The expression of emotions in man and animals.* Chicago, IL: University of Chicago Press (original work published in 1872).

Davies, M., Stankov, L., & Roberts, R. D. (1998) Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology, 75*, 989-1015.

de Gelder, B. & Vroomen, J. (2000). The perception of emotion by ear and by eye. Cognition and Emotion, 14(3), 289-311.

DeStano, D., Dasgupta, N., Bartlett, M.Y., & Cajdric, A. (2004). Prejudice from thin air: The effect of emotion on automatic intergroup attitudes. *Psychological Science,* 15, 319-324.

Digman, J.M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Ekman, P. (1972). Universals and cultural difference in facial expressions of emotion. In J. Cole (Ed.) Nebraska Symposium on Motivation, 207-283. Lincoln, NE: University of Nebraska Press.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*, 384-392.

Ekman, P. (1994). Strong evidence for universal in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin, 115*, 268-287.

Ekman, P. (2008). *Emotional Skills.* Invited address at the annual meeting of the Society of Industrial and Organizational Psychology, April, San Francisco , CA.

Ekman, P. and Friesen, W.V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*, 124-129.

Ekman, P., & Friesen, W.V. (1976). *Pictures of Facial Affect.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W.V., O'Sullivan, M., & Scherer, K. (1980). Relative importance of face, body, and speech in judgments of personality and affect. *Journal of Personality and Social Psychology, 38*, 207-277.

Elfenbein, H.A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*, 203-235.

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists.* Mawah, NJ: Earlbaum.

Etcoff, N.L. & Magee, J.J. (1992). Categorical perception of facial expressions. *Cognition, 44,* 227-240.

Fridlund, A.J. (1994). *Human facial expression: An evolutionary view.* San Diego, CA: Academic Press.

Gignac, G.E. (2005). Evaluating the MSCEIT V2.0 via CFA: Comment on Mayer et al. (2003). *Emotion, 5*, 233-235.

Gitter, G.A., Kozel, N.J., & Mostofsky, D.I (1972). Perception of emotion: The role of race, sex and presentation mode. *Journal of Social Psychology, 88*, 213-222.

Hall, J. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*, 845-857.

Hall, J.A. (2001). The PONS test and the psychometric approach to measuring interpersonal sensitivity. In J.A. Hall & F.J. Bernieri (eds.) *Interpersonal Sensitivity: Theory and Measurement.* Mahwah, NJ: Earlbaum.

Hall, J. & Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion, 4*, 201-206.

Hess, U., Adams, R., & Kleck, R. (2004). Facial appearance, gender, and emotion expression. *Emotion, 4*, 378-388.

Horstmann, G. (2003). What do facial expression convey: Feeling states, behavioral intentions, or action requests? *Emotion, 3,* 150-166.

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Hugenberg, K. (2005). Social categorization and the perception of facial affect: Target race moderates the response latency advantage for happy faces. Emotion, 5, 267-276.

Hulin, C.L., Drasgow, D., & Parsons, C.K. (1983). *Item Response Theory: Application to Psychological Measurement.* Homewood, IL: Dow Jones-Irwin.

Izard, C. E. (1991). *The Psychology of Emotions.* New York, NY: Plenum Press.

Izard, C.E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*, 288-299.

Jackson, D.L. Gillaspy, J. Arthur, & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*, 6-23.

Jenness, A. (1932). The recognition of facial expressions of emotion. *Psychological Bulletin, 29*, 324-350.

John, O.P., Donahue, E.M., & Kentle, R. (1991). The "Big Five Inventory": Versions 4a and 54. Berkeley: University of California, Berkeley, Institute of Personality and Social Research.

Jöreskog K.& Sörbom, D.(1999). LISREL 8: User's Reference Guide. Chicago, IL: Scientific Software International.

Jöreskog K.& Sörbom, D.(1999). LISREL 8.30. Chicago, IL: Scientific Software International.

Keltner, D. & Ekman, P. (2000). Facial expression of emotion. In M. Lewis and J. M. Haviland-Jones (Eds.) *Handbook of Emotions, 2nd Edition.* New York: Guilford Press.

Kirouac, G. & Dore, F.Y. (1985). Accuracy of the judgment of facial expression of emotion as a function of sex and level of education. *Journal of Nonverbal Behavior, 9*, 3-7.

Kostman, I. & Bedwell, S. (2003). Predicting multidimensional performance using cognitive ability, personality, and emotional intelligence. In C. Miller (Chair), *Researching emotional intelligence in a personnel psychology context.* Symposium conducted at the annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

Kuncel, N.R., Credé, M., & Thomas, L.L. (2005). The validity of self-reported grade point average, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*, 63-82.

Lance, C.E. & Vanderberg, R.J. (2002). Confirmatory factor analyis. In F. Drasgow and N. Schmitt (eds.) Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis. San Francisco, CA: Jossey-Bass.

Laukka, P. (2005). Categorical perception of vocal emotion expressions. Emotion, 5, 3, 277-295.

Lieberman, M.D., & Rosenthal, R. (2001). Why introverts can't always tell who likes them: Multitasking and nonverbal decoding. *Journal of Personality and Social Psychology, 80*, 294-310.

Lievens, F., Buyse, T. & Sacket, P.R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442-452.

Lievens, F., & Sacket, P.R. (2006). Video-based versus written situational judgment tests: a comparison in terms of predictive validity. *Journal of Applied Psychology, 91*, 1181-1188.

Lundquist, K.K., & Ashe, R.L. (2010) Trends in Employment Law: Ricci and Beyond. Workshop presented at the annual meeting of the Society of Industrial and Organizational Psychology, Atlanta, GA.

Mabe, P.A., & West, S.G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280-296.

MacCallum, R.C. Browne, M.W.,& Cai, L/ (2006). Testing differences between nested covariance structure models: Power analysis and null hypothesis. *Psychological Methods, 11*, 1, 19-35.

Marsh, A., Ambady, N., & Kleck, R. (2005). The effects of fear and anger facial expression on approach- and avoidance-related behaviors. *Emotion, 5*, 119-124.

Martin, W.E., Jr., Easton, C., Wilson, S., Takemoto, M., & Sullivan, S. (2004). Salience of emotional intelligence as a core characteristic of being a counselor. *Counselor Education and Supervision, 44*, 17-30.

Matthews, G. Zeidner, M., & Roberts, R.D. (2002). *Emotional Intelligence: Science and Myth.* Cambridge, MA: MIT Press.

Mayer, J.D., Gaschke, Y.N., Braverman, D.L., & Evans, T.W. (1992). Mood-congruent judgment is a general effect. *Journal of Personality and Social Psychology, 63*,1, 119-132.

Mayer, J.D., DiPaolo, M., & Salovey, P. (1993). Perceiving affective content in ambiguous visual stimuli. *Journal of Personality Assessment, 54*, 772-781.

Mayer, J.D., & Geher, G. (1996). Emotional intelligence and the identification of emotion. *Intelligence, 22*, 89-113.

Mayer, J.D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence, 17*, 433-442.

Mayer, J.D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D.J. Sluyter (eds.) *Emotion Development and Emotional Intelligence.* New York, NY: BasicBooks

Mayer, J.D., Salovey, P., & Caruso, D. R. (2000). Models of emotional intelligence. In R.J. Sternberg (Ed.) *Handbook of Human Intelligence.* New York, NY: Cambridge.

Mayer, J.D., Salovey, P., & Caruso, D. R. (1997). Emotional IQ test. Needham, MA: Virtual Knowledge.

Mayer, J.D., Salovey, P., & Caruso, D. R. (2002). *Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) user's manual.* Multihealth Systems; Toronto, Ontario, Canada.

McClure, E.B., & Nowicki, S., Jr. (2001). Associations between social anxiety and nonverbal processing in preadolescent boys and girls. Journal of Nonverbal Behavior, 25, 3-19.

Mehrabian, A. & Epstein, N. (1972). A measure of emotional empathy. J*ournal of Personality, 40*, 525-543.

Murphy, K. (2006). *A Critique of Emotional Intelligence: What Are the Problems and How Can They be Fixed?* Earlbaum; New York, NY.

Nowicki, S., Jr. & Duke, M.P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18*, 9-35.

Nowicki, S., Jr. & Duke, M.P. (2001). Nonverbal receptivity: The Diagnostic Analysis of Nonverbal Accuracy (DANVA). In J.A. Hall & F.J. Bernieri (eds.) *Interpersonal Sensitivity: Theory and Measurement.* Mahwah, NJ: Earlbaum.

Palfai, T. P., & Salovey, P. (1993). The influence of depressed and elated mood on deductive and inductive reasoning. *Imagination, Cognition, and Personality, 13*, 57-71.

Pessoa, L., Japee, S.,, & Ungerleider, L.G. (2005). Visual awareness and the detection of fearful faces. *Emotion, 5*, 243-247.

Richman-Hirsch, W.L., Olson-Buchanan, J.B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880-887.

Roberts, R. D., Zeidner, M., & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusions. *Emotion, 1*(3), 193-231.

Roseman, I.J., & Smith, C.A. (2001). Appraisal theory: Overview, assumptions, varieties, controversies. In K.R. Scherer, A. Schorr, & T. Johnstone (eds.) *Appraisal Processes in Emotion: Theory, Methods, Research.* Oxford, UK: Oxford University Press.

Rosenthal, R., Hall, J.A., DiMatteo, M.R., Rogers, P.L., & Archer, D. (1979). *Sensitivity to Nonverbal Communication: The PONS Test.* Baltimore, MD: Johns Hopkins University Press.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161-1178.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. Psychological Bulletin, 115, 102-141.

Russell, J.A. & Steiger, J.H. (1982). The structure in persons' implicit taxonomy of emotions. *Journal of Research in Personality, 16*, 447-469.

Salovey, P., Bedell, B.T., Detweiler, J.B., & Mayer, J.D. (2000). Current directions in emotional intelligence research. In M. Lewis and J. M. Haviland-Jones (Eds.) *Handbook of Emotions, 2nd Edition.* New York: Guilford Press.

Salovey P., & Mayer J. D. (1990). Emotional intelligence. *Imagination, Cognition, and Personality, 9*, 185-211.

Spencer-Smith, J., Wild, H., Innnes-Ker, A., Townsend, J., Duffy, C., Edwards, C., Ervin, K.,, Merritt, N., Won Paik , J. (2001). Making faces: Creating three-dimensional parameterized models of facial expression. *Behavioral Research Methods, Instruments, & Computers, 33*, 115-123.

van Ginkel, J.R., Sijtsma, K., van der Ark, A., & Vermunt, J.K (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. Methodology, 6(1), 17-30.

van Ginkel, J.R., van der Ark, A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42* (2), 387-414.

van Ginkel, J.R., van der Ark, A. (2008). SPSS syntax for the imputation of missing test data. Downloaded from http://www.datatheory.nl/pages/ginkel.html on October 25, 2008.

Wonderlic Inc.(1992). *Wonderlic Personnel Test & Scholastic Level Exam User's Manual.* Libertyville, IL; Author.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist, 39*(2), 117-123.

APPENDIX A:

ITEM LEVEL STATISTICS

Table 16

Consensus Limited Information

	Item Mean	Item Standard Deviation	Corrected Item-Total Correlation
Scene01	.78	.41	0.13
Scene02	.73	.45	0.25
Scene04	.78	.41	0.35
Scene05	.85	.36	0.18
Scene06	.55	.50	0.30
Scene07	.74	.44	0.32
Scene08	.58	.50	0.14
Scene10	.72	.45	0.38
Scene11	.37	.48	0.39
Scene12	.25	.44	0.27
Scene15	.80	.40	0.29
Scene18	.60	.49	0.33
Scene21	.72	.45	0.47
Scene24	.84	.37	0.13
Scene25	.69	.47	0.29
Scene27	.74	.44	0.36
Scene28	.81	.40	0.21
Scene30	.63	.48	0.44
Scene31	.63	.48	0.39
Scene34	.69	.46	0.36
Scene35	.87	.34	0.37
Scene39	.69	.47	0.08
Scene40	.71	.46	0.28
Scene41	.59	.49	0.29
Scene44	.79	.41	0.41
Scene45	.65	.48	0.34
Scene49	.51	.50	0.41
Scene51	.95	.21	0.21
Scene52	.76	.43	0.29
Scene53	.88	.33	0.33
Scene54	.49	.50	0.32
Scene56	.56	.50	0.38
Scene58	.73	.45	0.25
Scene60	.71	.45	0.25
Scene61	.81	.39	0.14
Scene62	.63	.48	0.33
Scene63	.77	.42	0.43
Scene64	.73	.44	0.25
Scene67	.46	.50	0.39

Table 16 (cont.)

Scene68	.58	.49	0.23
Scene69	.83	.37	0.48
Scene70	.77	.42	0.36
Scene71	.46	.50	0.29
Scene72	.57	.50	0.37
Scene73	.98	.12	0.23
Scene74	.92	.27	0.28
Scene75	.46	.50	0.29
Scene76	.86	.35	0.27
Scene77	.47	.50	0.27
Scene79	.52	.50	0.33
Scene80	.70	.46	0.28
Scene81	.83	.38	0.35
Scene83	.60	.49	0.39
Scene84	.77	.42	0.37
Scene87	.40	.49	0.28
Scene88	.72	.45	0.23
Scene89	.48	.50	0.31
Scene90	.87	.34	0.30
Scene91	.66	.48	0.44
Scene92	.35	.48	0.41

Table 17

Expert Limited Information

	Item Mean	Item Standard Deviation	Corrected Item-Total Correlation
Scene02	0.72	0.45	0.23
Scene04	0.79	0.41	0.35
Scene05	0.85	0.36	0.17
Scene06	0.82	0.39	0.19
Scene07	0.98	0.15	0.03
Scene08	0.59	0.49	0.19
Scene09	0.52	0.50	0.37
Scene10	0.71	0.45	0.30
Scene11	0.37	0.48	0.17
Scene12	0.82	0.38	0.20
Scene14	0.60	0.49	0.26
Scene15	0.80	0.40	0.35
Scene19	0.58	0.49	0.28
Scene21	0.97	0.18	0.15
Scene22	0.73	0.44	0.06
Scene24	0.94	0.24	0.03
Scene27	0.75	0.44	0.36
Scene28	0.81	0.40	0.17
Scene30	0.85	0.36	0.29
Scene31	0.90	0.31	0.29
Scene33	0.75	0.43	0.37
Scene34	0.88	0.32	0.13
Scene35	0.87	0.34	0.33
Scene37	0.75	0.43	0.32
Scene40	0.25	0.43	0.27
Scene41	0.59	0.49	0.28
Scene43	0.58	0.49	0.31
Scene44	0.80	0.40	0.43
Scene45	0.66	0.47	0.33
Scene46	0.54	0.50	0.35
Scene49	0.84	0.37	0.32
Scene50	0.38	0.49	0.31
Scene51	0.95	0.21	0.18
Scene53	0.53	0.50	0.34
Scene54	0.82	0.38	0.14
Scene55	0.92	0.28	0.18
Scene58	0.92	0.28	0.08
Scene59	0.85	0.35	0.19
Scene60	0.72	0.45	0.19
Scene61	0.81	0.40	0.15

Table 17 (cont.)

Scene62	0.89	0.31	0.26
Scene63	0.77	0.42	0.36
Scene64	0.74	0.44	0.29
Scene66	0.75	0.43	0.19
Scene67	0.83	0.38	0.28
Scene69	0.98	0.15	0.19
Scene70	0.95	0.21	0.15
Scene71	0.74	0.44	0.27
Scene72	0.93	0.26	0.37
Scene73	0.98	0.13	0.33
Scene74	0.92	0.28	0.27
Scene75	0.73	0.44	0.28
Scene76	0.86	0.35	0.27
Scene77	0.47	0.50	0.23
Scene80	0.71	0.46	0.27
Scene81	0.98	0.14	0.18
Scene83	0.87	0.34	0.18
Scene84	0.95	0.23	0.27
Scene88	0.74	0.44	0.29
Scene89	0.79	0.41	0.09
Scene90	0.87	0.33	0.29
Scene91	0.91	0.28	0.20
Scene92	0.65	0.48	0.40

Table 18

Consensus MD Scoring

	Item Mean	Item Standard Deviation	Corrected Item-Total Correlation
Scene01	2.83	1.31	0.54
Scene02	2.76	1.58	0.47
Scene04	2.74	1.60	0.45
Scene05	2.60	2.00	0.56
Scene06	2.44	1.62	0.52
Scene08	2.49	1.37	0.46
Scene11	2.74	1.47	0.49
Scene12	2.53	1.88	0.53
Scene15	2.61	1.33	0.50
Scene16	2.45	1.46	0.51
Scene18	2.54	1.37	0.56
Scene20	2.89	1.88	0.59
Scene21	2.92	1.55	0.57
Scene24	2.31	2.60	0.43
Scene25	2.66	1.79	0.58
Scene27	2.67	1.90	0.58
Scene28	2.61	1.61	0.52
Scene29	2.47	1.88	0.65
Scene30	2.70	1.35	0.60
Scene31	2.44	1.64	0.67
Scene33	2.86	1.71	0.71
Scene34	2.89	1.50	0.54
Scene35	2.41	2.52	0.57
Scene37	2.89	1.22	0.53
Scene39	2.59	2.10	0.57
Scene40	2.62	1.43	0.63
Scene42	2.46	1.70	0.54
Scene43	2.70	1.81	0.59
Scene44	2.57	2.13	0.58
Scene45	2.53	1.57	0.57
Scene46	2.90	1.54	0.62
Scene47	2.47	1.72	0.72
Scene48	2.52	1.40	0.67
Scene49	2.49	1.71	0.67

Table 18 (cont.)

Scene52	2.65	1.60	0.64
Scene53	2.82	1.78	0.62
Scene55	2.45	1.85	0.58
Scene56	2.43	1.64	0.57
Scene57	2.69	1.95	0.63
Scene58	2.23	2.24	0.57
Scene59	2.75	1.54	0.53
Scene60	2.55	2.22	0.61
Scene62	2.72	1.32	0.45
Scene64	2.65	1.60	0.67
Scene68	2.62	1.32	0.68
Scene69	2.46	1.70	0.61
Scene71	2.55	1.69	0.59
Scene74	2.48	1.81	0.63
Scene75	2.56	1.28	0.52
Scene76	2.53	1.99	0.46
Scene77	2.48	2.26	0.42
Scene78	2.63	1.41	0.49
Scene80	2.68	1.34	0.63
Scene81	2.51	1.92	0.70
Scene83	2.43	1.83	0.65
Scene84	2.65	1.64	0.57
Scene85	2.59	1.67	0.53
Scene86	2.56	1.72	0.71
Scene87	2.76	1.62	0.71
Scene88	2.49	1.70	0.61
Scene89	2.58	1.99	0.58
Scene91	2.59	1.83	0.67
Scene92	2.68	1.47	0.53

Table 19

Expert MD

	Item Mean	Item Standard Deviation	Corrected Item-Total Correlation
Scene02	2.79	1.59	0.48
Scene04	2.80	1.42	0.41
Scene05	3.26	1.55	0.56
Scene06	2.66	1.56	0.52
Scene08	2.70	1.51	0.42
Scene09	3.10	1.53	0.45
Scene11	3.56	1.55	0.48
Scene12	2.63	1.53	0.50
Scene14	3.01	1.68	0.55
Scene15	2.54	1.23	0.43
Scene16	2.45	1.30	0.46
Scene19	2.99	1.38	0.60
Scene20	2.83	1.93	0.47
Scene21	3.10	1.55	0.59
Scene22	2.88	1.57	0.60
Scene24	3.78	1.76	0.40
Scene25	2.73	1.96	0.50
Scene26	2.96	1.42	0.60
Scene27	2.92	2.01	0.59
Scene28	2.79	1.69	0.56
Scene29	2.58	1.92	0.66
Scene30	2.85	1.50	0.61
Scene31	2.72	1.84	0.69
Scene33	3.29	1.77	0.71
Scene34	3.03	1.17	0.51
Scene35	2.81	2.52	0.59
Scene36	2.77	1.10	0.36
Scene37	2.79	1.19	0.38
Scene38	3.10	1.57	0.49
Scene40	3.39	0.96	0.42
Scene43	2.83	1.62	0.54
Scene44	2.61	2.08	0.47
Scene45	2.64	1.61	0.57
Scene46	3.07	1.61	0.57
Scene48	2.69	1.50	0.62
Scene49	2.68	1.69	0.69
Scene50	3.58	1.50	0.44
Scene52	2.97	1.47	0.65
Scene53	3.05	1.90	0.61
Scene54	2.92	1.60	0.59

Table 19 (cont.)

Scene55	3.10	1.47	0.57
Scene58	2.32	2.17	0.41
Scene59	2.94	1.57	0.58
Scene60	2.53	2.40	0.37
Scene62	2.97	1.55	0.45
Scene64	2.92	1.56	0.64
Scene66	2.88	1.67	0.66
Scene69	2.57	1.67	0.62
Scene71	2.61	1.55	0.51
Scene74	2.52	1.85	0.63
Scene75	2.64	1.28	0.51
Scene76	2.72	1.80	0.45
Scene77	2.62	2.30	0.39
Scene80	2.74	1.21	0.62
Scene81	2.79	1.87	0.70
Scene82	2.61	1.50	0.55
Scene83	2.73	1.47	0.63
Scene84	3.10	1.86	0.62
Scene85	3.25	1.33	0.51
Scene86	2.78	1.45	0.70
Scene88	2.59	1.87	0.61
Scene89	2.71	2.19	0.57
Scene91	2.66	1.69	0.66
Scene92	2.86	1.53	0.45