Journal of Nonverbal Behavior, **25**, pp.167-186.

A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression

Michael A. Sayette, Jeffrey F. Cohn, Joan M. Wertz, Michael A. Perrott, and Dominic J. Parrott
University of Pittsburgh

Abstract

The Facial Action Coding System (FACS) (Ekman & Friesen, 1978) is a comprehensive and widely used method of objectively describing facial activity. Little is known, however, about inter-observer reliability in coding the occurrence, intensity, and timing of individual FACS action units. The present study evaluated the reliability of these measures. Observational data came from three independent laboratory studies designed to elicit a wide range of spontaneous expressions of emotion. Emotion challenges included olfactory stimulation, social stress, and cues related to nicotine craving. Facial behavior was video-recorded and independently scored by two FACS-certified coders. Overall, we found good to excellent reliability for the occurrence, intensity, and timing of individual action units and for corresponding measures of more global emotion-specified combinations.

Introduction

After a long hiatus, research on emotion has emerged as an important topic of psychological inquiry (see Ekman, 1998; National Advisory Mental Health Council, 1995; Russell & Fernandez-Dols, 1997). Much credit for this renewed interest is attributable to the development beginning in the 1970s of observational coding systems to identify facial expressions thought to be associated with emotion. These include the Facial Action Scoring Technique (FAST: Ekman, Friesen, & Tomkins, 1971), A System for Identifying Affect Expressions by Holistic Judgment (AFFEX: Izard, Dougherty, & Hembree, 1983), EMFACS (Ekman & Friesen, 1982), Monadic Phases (Tronick, Als, & Brazelton, 1980), the Maximally Discriminative Facial Movement Coding System (MAX: Izard, 1979), and the Facial Action Coding System (FACS: Ekman & Friesen, 1978).

Many of these systems enable researchers to make judgments about emotion state and to code facial expression using emotion labels. Such systems can be learned relatively quickly and have become highly influential (e.g., Campos, Barrett, Lamb, Goldsmith, & Stenberg, 1983). They have several limitations. Different systems give the same labels to different facial actions (e.g., Oster et al., 1992); they implicitly or explicitly assume that facial expression and emotion have an exact correspondence, which is problematic (see Camras, 1992; Fridlund, 1992; Russell, 1994), and paralinguistic and other nonverbal expressions, such as the brow flash, used as a greeting display in many parts of the world (e.g., Eibl-Eibesfeldt, 1989), are omitted. As a consequence, more descriptive methods of describing facial expression have become increasingly influential in emotion science (e.g., Ekman & Rosenberg, 1997). Chief among these are the

Maximally Discriminative Facial Movement Coding System (MAX: Izard, 1979) and the Facial Action Coding System (FACS: Ekman & Friesen: 1978).

FACS (Ekman & Friesen, 1978), informed by the pioneering work of Hjortsjo (1969), is more comprehensive than MAX (Oster et al., 1992). Using FACS and viewing videotaped facial behavior in slow motion, coders can manually code all possible facial displays, which are decomposed into 30 action units (AUs) and 14 miscellaneous actions. AUs have a specified anatomic basis, while miscellaneous actions (e.g., jaw thrust) are ones for which the anatomic bases have not been established (Ekman & Friesen, 1978). By comparison, MAX (Izard, 1979) provides less complete description of facial actions (e.g., Malatesta, Culver, Tesmna, & Shephard, 1989), fails to differentiate between some anatomically distinct expressions (Oster et al., 1992), and considers as separable expressions that are not anatomically distinct (Oster et al., 1992).

Combinations of FACS action units may be described with emotion labels if investigators choose to do so; but this inferential step is extrinsic to FACS. FACS itself is purely descriptive. Because of its descriptive power, FACS has emerged as the criterion measure of facial behavior in multiple fields including computer vision (Bartlett, Hager, Ekman, & Sejnowski, 1999; Lien, Kanade, Cohn, & Li, 2000; Tian, Kanade, & Cohn, in press), computer graphics (Parke & Waters, 1996), neuroscience (Bruce & Young, 1998; Katsikitis & Pilowsky, 1988; Rinn, 1991) forensic science (Frank, 1996), and developmental (Camras, 1992; Fox & Davidson, 1988; Oster, Hegley, & Nagel, 1992), social (Frank & Ekman, 1997) and clinical (Part II, Ekman & Rosenberg, 1997) studies of emotion.

Much of the research using FACS has involved posed expressions (Rosenberg, 1997). Participants are directed to voluntarily contract specific muscles, often with the aim of producing expressions believed to represent emotion prototypes (e.g., Ekman, Levenson, & Friesen, 1983). Although studies of posed facial expression have generated a range of important findings, the allure of FACS is that it also may provide an immediate, objective, unobtrusive, and reliable analysis of spontaneously generated expressions (see Ekman & Rosenberg, 1997; European Conference on Facial Expression, Measurement, and Meaning, 1999). The potential implications for such research rest, however, on the assumption that FACS can be used reliably to code spontaneous emotional expressions.

Psychometric knowledge of FACS has not kept pace with the increasing and multiple uses of FACS coding in emotion science and related fields and in changes in FACS over time. While a proficiency test is required for certification as a FACS coder, relatively little information is available about the reliability of FACS coding for individual action units and related measures (e.g., action unit intensity), especially for the use of FACS with spontaneous facial expression. It may be unwise to assume that good reliability for these measures in posed expression indicates good reliability in spontaneous expressions. Spontaneous expressions are believed to differ from voluntary ones in both their morphology and dynamics, including velocity and smoothness of motion (Hager & Ekman, 1995). In addition, rigid head motion and face occlusion, which can impede coding (e.g., Matias, Cohn, & Ross, 1989; Kanade, Cohn, & Tian, 2000), are more likely to occur during spontaneous expressions. To accommodate the greater range of head motion

found in studies of spontaneous expression, investigators often use wider camera angles, which reduces face size relative to the video frame and makes coding of subtle motion more difficult. The FACS certification test requires coders to score videotapes of spontaneous expression with a high level of agreement with a group of reference coders. Consequently all FACS-certified coders presumably have achieved reliability. Nevertheless, many published studies of spontaneous expression either fail to report FACS reliability (e.g., Banninger-Huber, 1992; Brummett et al., 1998; Chesney, Ekman, Friesen, Black, & Hecker 1990; Ellgring, 1986; Heller & Haynal, 1997) or provide only incomplete information. It is important to test whether the reliability achieved for FACS certification is maintained in research studies.

At least four types of inter-observer reliability (i.e., agreement between observers) are relevant to the interpretation of substantive findings. One is the reliability of individual AUs. Most studies report only reliability averaged across all AUs, which may mask low reliability for specific ones (Ekman & Rosenberg, 1997). Failure to consider reliability of individual AUs may be of little concern when investigators analyze aggregate measures. When specific AUs are the focus of hypothesis testing, reliability at this more micro-level is needed. Otherwise, statistical power may be reduced by measurement error and negative findings misinterpreted. Information about reliability of measurement for measures of interest is important in planning new studies. Even when reliability of individual AUs is assessed, investigators typically fail to use information statistics such as kappa (Cohen, 1960; Fleiss, 1981) that correct for chance agreement between coders. If agreement statistics are not corrected for chance agreement, reliability estimates will fail to generalize to populations in which the marginal distribution of AUs varies.

A second type of reliability is the temporal resolution of FACS coding. AUs typically are coded using stop-frame video, which affords a temporal resolution of either $1/30^{th}$ or $1/25^{th}$ of a second depending on the video standard (NTSC and PAL, respectively; higher resolution is available with special purpose video decks). It is unknown whether investigators can detect change in AUs on this time base. Ekman, Friesen, and Simons (1985), for instance, reported mean differences in AU onset between coders, but not the distribution of errors as a function of variation in the time base. When hypotheses entail very small differences in latency or the duration of response (e.g., Cohn & Elmore, 1996; Ekman et al., 1985), precision of measurement is a critical concern.

Third, facial motion varies in degree as well as in type of AU shown. The intensity of facial movement is scored for five of the 30 action units (Ekman & Friesen, 1978). Differences in the intensity of AUs are believed to relate to differences in intensity of subjective experience (Camras, Oster, Campos, Miyake, & Bradshaw, 1992; Rosenberg & Ekman, 1994; but see also Fridlund, 1992; Russell, 1994). In contrast to deliberate expressions, spontaneous facial actions are believed to show greater symmetry in intensity of motion between left and right sides of the face (Hager & Ekman, 1985). Little is known, however, about whether coders can agree on intensity scoring. The FACS certification test, for instance, omits intensity differences altogether, and research reports often fail to evaluate measurement error in this regard, even when intensity differences are focus of study (e.g., Camras et al., 1992). Hypotheses about

intensity variation in AU intensity depend on whether coders can reliably agree on intensity. With few exceptions, reliability of intensity scoring of individual AUs is unknown.

Fourth, in many studies, investigators are interested in testing hypotheses about emotion-specified expressions, which are considered to represent combinations of AUs. Emotion-specified expressions include discrete emotions (e.g., joy, surprise, sadness, anger, fear, and disgust) and more molar distinctions between positive versus negative emotion. The reliability of emotion–specified expressions will of course depend on the constituent AUs. By assessing the reliability of these aggregates directly, one can more accurately estimate their reliability.

Recent changes in the minimum criteria for scoring AUs (Friesen & Ekman, 1992) make reliability assessment especially compelling. Under the traditional 3-point scoring, "trace" levels of AUs were ignored. Under the new 5-point rules, AUs at trace levels (AU intensity "a") are now coded. The original decision to ignore them was made in part because of the difficulty in attaining reliability about very small motion. The effect of scoring AUs at the trace level on FACS reliability is unknown. At issue is the comparability of studies that use the 3-point versus the 5-point approach to intensity scoring.

The present research evaluated inter-observer reliability of FACS in studies of spontaneous expression. Reliability was assessed for the occurrence of single AUs, the precision with which AUs could be coded at temporal resolutions up to $1/30^{th}$ of a second, and inter-observer reliability for AU intensity and emotion-specified expressions. Emotion-specified expressions, which are defined in terms of specific combinations of AUs, included both those thought to represent discrete emotions and more molar categories of positive and negative emotion. To ensure a representative test-bed of spontaneous expression, we included data from three experiments involving independent samples of subjects. Participants were videotaped under typical emotion-eliciting laboratory conditions. Facial behavior was FACS coded from videotape by experienced pairs of coders certified in the use of FACS. Agreement between coders was quantified using kappa coefficients to control for chance levels of agreement.

The first experiment induced emotion through the administration of odors that were pre-selected to elicit emotional responses. Odors often have been used to manipulate emotions (Aggleton & Mishkin, 1986; Engen, 1992), and of most relevance to the present study, have produced facial expressions associated with affect (see Soussignan & Schaal, 1996). In the second experiment, smokers underwent a cue exposure manipulation, in which they were asked to light, hold, look at, but not smoke a cigarette. This procedure has also been found to elicit potent emotional reactions across a range of response systems (Rohsenow, Niaura, Childress, Abrams, & Monti, 1990-1991), including facial expression (Sayette & Hufford, 1995). The third experiment induced an emotional response by instructing participants to present a self-disclosing speech pertaining to their physical appearance. This speech instruction has been found to increase negative affect across a range of measures (Levenson, Sher, Grossman, Newman, & Newlin, 1980; Sayette & Wilson, 1991; Steele & Josephs, 1988), including facial expressive behavior (e.g., Sayette, Smith, Breiner, & Wilson, 1992). These studies provided suitable material for a rigorous test of the reliability of FACS for the occurrence of single AUs, AU intensity, and emotion-specified aggregates at varying levels of temporal resolution.

Method

Overview

The three laboratory experiments described here all involved the elicitation of spontaneous expressions. In each case, participants' facial movements were videotaped using a single S-VHS camera and recorder that provided a frontal view of the subject's chest and face. Segments for coding were pre-selected for the likelihood of reflecting emotion in the face. These segments were independently coded by two individuals who were certified in the use of FACS by the Social Interaction Laboratory at the University of California San Francisco (DP and JW, Experiment 1; MP and JW, Experiments 2 and 3). Inter-observer agreement was quantified with coefficient kappa, which is the proportion of agreement above what would be expected to occur by chance (Cohen, 1960; Fleiss, 1981). Coefficients of 0.60 to about 0.75 indicate good, or adequate reliability; coefficients of 0.75 or higher indicate excellent reliability (e.g., Fleiss, 1981).

Brief Description of the Three Experiments

In Experiment 1, 58 participants (30 male and 28 female) were asked to rate the pleasantness of a series of eight odors. Odors included cooking extract oils (e.g., coconut, peppermint, banana, lemon) as well as a Vicks mixture, a vinegar solution, a floral odor, and a neutral water odor [for additional information about this study, see Sayette and Parrott (1999)]. Participants were videotaped using a Panasonic 450 Super-VHS recorder and Super-VHS videotape. Videotapes of the facial expressions were then coded using FACS. Two of the eight segments were randomly selected to be coded by both raters, producing 116 segments, each lasting about 12-14 seconds.

In Experiment 2, reliability data were obtained by having two FACS raters independently code the expressions of a randomly selected subset of participants in an investigation of smoking craving (Sayette, Martin, Wertz, Shiffman, & Perrott, under review). Data were collected on eleven women and eight men who were either light (i.e., smoke five or fewer cigarettes at least two days/week, $\underline{n} = 7$) or heavy (smoke 22-40 cigarettes daily, $\underline{n} = 12$) smokers participating in a smoking cue exposure procedure, in which they held and looked at a lit cigarette. [For details about the study's methods, see also a similar study by Sayette and Hufford (1995)]. To elicit a wide range of emotions, some of the smokers were deprived of nicotine for seven hours ($\underline{n} = 6$) while others were nondeprived ($\underline{n} = 13$). Facial behavior was recorded with the same equipment used in Experiment 1. Six time periods were independently coded by two FACS coders, producing 114 segments, each ranging from 5-10 secs. These time periods included moments when participants first were presented with the cigarette, when they were holding a lit cigarette, and when they were first permitted to smoke the cigarette.

In Experiment 3, reliability data were obtained from 25 (13 male and 12 female) participants who were randomly selected from 169 participants who completed the experiment [for additional information about this study, see Sayette, Martin, Perrott, Wertz, and Hufford (in

press)]. Participants were asked to present a three-minute speech about what they liked and disliked about their body and physical appearance. They also were told that this speech would be video recorded and evaluated by psychologists on a variety of psychological variables. Facial expressions were recorded using a JVC Super-VHS recorder and Super-VHS videotape. Two time periods previously found to be especially sensitive to this manipulation (the 10-seconds after being informed of the speech topic and the final 20-seconds before beginning the speech) were coded using FACS. Because some of the participants were informed of the speech topic on two occasions, there were 61 segments ranging from 10-20 seconds.

Data Analysis

In previous research, reliability of FACS AUs was typically assessed as the proportion or percentage of agreement between coders (i.e., [Agree / (Agree + Disagree]). As noted above, because this statistic fails to account for agreements due to chance, coefficient kappa is the preferred statistic (Cohen, 1960; Fleiss, 1981). Kappa was used in the present study and is reported below. To increase stability of estimate, data were pooled across the three studies prior to computing kappa coefficients.

The reliability of FACS was assessed at four levels of analysis: 1) Occurrence/non-occurrence of individual AU scoring; 2) temporal precision of individual AU scoring; 3) AU intensity; and 4) emotion-specified expressions. In assessing precision of scoring, we used time windows of 0, 5, 10, and 15 frames, which correspond to $1/30^{th}$, $1/6^{th}$, $1/3^{rd}$, and 1/2 second, respectively. Coders were said to agree on the occurrence of an AU if they both identified it within the same time window.

In addition to individual AUs, we assessed molar positive and negative expressions. Coded as positive were AU 12 (lip corner pull) and AU 6 + 12 (cheek raise with lip corner pull), which could be accompanied by any of the following: 1+2 (brow raised both medially or laterally), 25 or 26 (jaw parted or jaw lowered) (Ekman, Friesen, & O'Sullivan, 1988; Ekman, Friesen, & Ancoli, 1980; Ekman, Davidson, & Friesen, 1990; Sayette & Hufford, 1995; Smith, 1989). For expressions to be considered positive, AU 12 had to receive a minimum intensity rating of "b" using Friesen's and Ekman's (1992) updated 5-point "a" to "e" intensity scale if it co-occurred with AU 6, and AU 12 had to receive a minimum intensity rating of "c" if it did not appear with AU 6. Negative emotional expressions were defined by the absence of AU 12 and the presence of at least one of the following AUs: 9 (nose wrinkle); 10 (upper lip raise); unilateral 14 (dimpler); 15 (lip corner depress); 20 (lip stretch), and 1+ 4 (pulling the medial portion of the eyebrows upwards and together). These AUs are thought to appear during the expression of negative emotion (Ekman & Friesen, 1982; 1986; Ekman et al, 1980; Gosselin, Kirouac, & Dore, 1995; Rozin, Lowery, & Ebert, 1994; Soussignan & Schaal, 1996; Vrana, 1993). For negative AUs, a minimum intensity rating of "b" was required in order to meet criteria (Friesen & Ekman, 1992). In addition, two negative emotion-specified expressions (sadness and disgust) as defined in the Investigator's Guide that accompanies the FACS manual (Ekman & Friesen, 1978) and in published studies referenced below occurred with sufficient frequency to permit analysis.

Results

Nineteen AUs met the 48-frame inclusion criterion and were included for analysis. With the exception of AU 5, we were able to include the actions that are most common in studies of emotion and paralinguistic expression (e.g., Cohn et al., 1999).

Inter-Observer Reliability for the Occurrence of Specific AUs and for Precision of Measurement

Using a 1/2-second tolerance window, all but two AUs (AUs 7 and 23) had good to excellent reliability (see Table 1). As the tolerance window decreased in size, the number of AUs with good to excellent reliability decreased. Even at the smallest tolerance window, however, 11 of 19 AUs continued to have good to excellent reliability.

Insert Table 1 about here

Inter-Observer Reliability for Action Unit Intensity

To examine the effect of using a 3-point vs. a 5-point intensity scale on reliability, kappas were calculated for the four AUs that were coded for intensity. Reliability was better for 3-point intensity scoring than for 5-point scoring (see Table 2). Intensity scoring of AU 10 and AU 12 was acceptable even with a zero frame ($1/30^{th}$ second) tolerance window. Intensity scoring of AU 15 was acceptable beginning with a $1/6^{th}$ second or larger tolerance window and a 3-point but not 5-point intensity scale. Intensity scoring of AU 20 was borderline acceptable on the 3-point scale, with a ½ second window.

Insert Table 2 about here

Inter-observer reliability for emotion-specified expressions

Table 3 shows the corresponding values for positive- and negative- AU combinations using the 3-point intensity scale. For positive AU combinations, reliability was excellent even at the most stringent (to the frame) level of analysis. Negative AU combinations also were coded reliably. In addition, negative emotion was examined more specifically. Reliabilities for disgust and sadness are also presented in Table 3. Kappas remained excellent for disgust and sadness. [Reliabilities were comparable using 5-point intensity scoring (kappas within .05), with the exception of the zero frame tolerance window (kappas reduced by .07 to .08 with the 5-point scale)].

Insert Table 3 about here

Discussion

Our major finding was that FACS had good to excellent reliability for spontaneously generated facial behavior. Across three experiments, reliability was good to excellent for nearly all (90%) AUs. These included AUs in all regions of the face. Many of these AUs involve

subtle differences in appearance. Moreover, these AUs are central to emotion expression and paralinguistic communication.

Only two AUs, AU 7 and AU 23, had fair reliability even at the slowest sampling rate. Cohn, Zlochower, Lien, and Kanade (1999) found relatively low, though acceptable, reliability for these AUs in a large sample of directed facial action tasks. The tightening of the lower lid in AU 7 is a relatively small appearance change that often is mistaken for AU 6, which is controlled by the same muscle. These AUs often co-occur, which makes the distinction difficult as well. Similarly, AU 23 (lip tightening) is a relatively subtle change in appearance often mistaken for AU 24 (lip pressing). AU 23 and AU 24 are both controlled by the same muscle and co-occur frequently. Because AUs 23 and 24 are both associated with emotion-specified anger, confusion between them may have little consequence at this level of description. Nevertheless, FACS training and documentation would benefit by increased attention to sources of confusion between those action units for which systematic confusions were found.

A second finding was that reliability estimates varied depending on time frame precision. By evaluating reliability using different tolerance windows ranging from exact frame reliability to reliability within 1/2 second, we found that reliability, while adequate at the precise frame-by-frame unit of measurement, was considerably improved when the tolerance window expanded to 1/2 second. Reliability improved significantly between the $1/30^{th}$ second and $1/6^{th}$ second frame tolerance windows. Kappas did not increase much, however, between the $1/6^{th}$ second and the $1/3^{rd}$ and 1/2 second windows. This pattern may reflect difficulty discerning the exact frame that an AU reaches minimum requirements for coding. Even a tolerance window of $1/6^{th}$ second, however, provides adequate latitude for temporal agreement. For most purposes, a 1/2 second tolerance window is probably acceptable. When brief latencies are crucial to hypotheses, however (e.g., Ekman et al, 1985; Hess & Kleck, 1990) smaller tolerance windows may be necessary. At a minimum, the tolerance window used should be noted when reporting reliabilities in studies that include AU duration as a variable.

Reliability also was good for AU intensity. Not surprisingly, agreement on the 5-point intensity scale was somewhat lower than that for the 3-point scale. There are at least two reasons for this difference. One, of course, is that five-point intensity scoring requires finer differentiation. A second, however, is that what counts as an occurrence differs depending on which intensity scale is used. Trace levels of occurrence are counted when scoring 5-point intensity, while they are ignored when scoring 3-point intensity. Consequently, reliability for AU occurrence varies depending on which intensity criteria are followed. This can be seen for AU 12. Reliability for occurrence of AU 12 was higher when 3- versus 5-point scoring was followed.

When molar distinctions were drawn between positive and negative emotion-related AUs, reliability was good to excellent. Even when scoring was to the exact frame, reliability remained good. These data are reassuring for investigators interested in assessing emotion-specified positive and negative affect (e.g., Frank, Ekman, & Friesen, 1993; Ruch, 1993; Sayette & Hufford, 1995). Results also indicated that reliability was excellent for AU combinations associated with disgust and sadness.

The current study as has some limitations. Although we included data from three separate studies, some AUs occurred with low frequency, which precluded reliability estimation. The reliability of FACS coding was not evaluated across different laboratories. It is important to establish that different laboratories are using FACS in the same way. Reliability studies are needed in which the same data sets are coded by multiple groups of FACS coders. Another limitation is that the validity of FACS as a measure of emotion was not assessed. There is a large literature relevant to this topic (cf. Ekman & Rosenberg, 1997; Russell & Fernandez-Dols, 1997), and a meta-analysis of studies using FACS would be timely.

The coding of facial expression to assess emotional responding is becoming increasingly popular. Nevertheless, to our knowledge, evidence to support reliable use of the most comprehensive facial coding system (FACS) has yet to be published. The present data, using three different emotion induction procedures, provide an important test of the reliability of FACS. Despite difficulties that can arise when coding spontaneous expressions in the laboratory, these data empirically confirm that FACS can be reliably used to code spontaneous facial expression.

Acknowledgements

This research was supported by National Institute on Drug Abuse Grant R01-DA10605 and National Institute on Alcohol Abuse and Alcoholism Grant R29-AA09918 to Michael Sayette, and National Institute of Mental Health Grant R01-51435 to Jeffrey Cohn. We thank Adena Zlochower and Peter Britton for their helpful comments and David Liu for his technical assistance.

Correspondence concerning this manuscript should be addressed to Michael Sayette, Department of Psychology, 604 Old Engineering Hall, University of Pittsburgh, Pittsburgh, PA 15260. Electronic mail can be sent to sayette@pitt.edu.

References

Proceedings of the 8th European Conference on Facial Expression, Measurement, and Meaning (September, 1999) Saarbruecken, Germany.

Aggleton, J.P. & Mishkin, M. (1986). The amygdala: Sensory gateway to emotions. In R. Plutchik & H. Kellerman (Eds.). Emotion: Theory, research and experience. Vol 3: Biological foundations of emotion (pp. 281-296). Orlando, FL: Academic Press.

Banninger-Huber, E. (1992). Prototypical affective microsequences in psychotherapeutic interaction. Psychotherapy Research, 2, 291-306.

Bartlett, M.S., Hager, J.C., Ekman, P., & Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. Psychophysiology 36, p. 253-263.

Bruce, V.& Young, A. (1998). In the eye of the beholder: The science of face perception. NY: Oxford.

Brummett, B.H., Maynard, K.E., Babyak, M.A., Haney, T.L., Siegler, I.C., Helms, M.J., & Barefoot, J.C. (1998). Measures of hostility as predictors of facial affect during social interaction: Evidence for construct validity. Annals of Behavioral Medicine, 20, 168-173.

Campos, J.J., Barrett, K.C., Lamb, M.E., Goldsmith, H.H., & Stenberg, C. (1983). Socioemotional development. In M.M. Haith & J.J. Campos (Eds.) (P.H. Mussen Series Ed.), Infancy and developmental psychobiology (pp. 783-916).

Camras, L. (1992). Expressive development and basic emotions. Cognition and Emotion, 6, 269-283.

Camras, L. Oster, H., Campos, J., Miyake, K., & Bradshaw, D. (1992). Japanese and American infants' responses to arm restraint. Developmental Psychology, 28, 578-583.

Carroll, J. M., & Russell, J. A. (1997). Facial expressions in Hollywood's portrayal of emotion. Journal of Personality and Social Psychology, 72, 164-176.

Chesney, M.A., Ekman, P., Friesen, W.V. , Black, G.W., & Hecker, M.H.L. (1990). Type A behavior pattern: Facial behavior and speech components. Psychosomatic Medicine, 53, 307-319.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20, 37-46.

Cohn, J.F. and Elmore, M. (1988). Effect of contingent changes in mothers' affective expression on the organization of behavior in 3-month-old infants. Infant Behavior and Development, 11, 493-505.

Cohn, J.F., Zlochower, A., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. Psychophysiology, 36, 35-43. Available from http://www.cs.cmu.edu/~face.

Eibl-Eibesfeldt, I. (1989). Human ethology. NY: Aldine de Gruyter.

Ekman, P. (1982). Methods for measuring facial action. In K.R. Scherer & P. Ekman (Eds.), Handbook of methods in nonverbal behavior research (pp. 45-90). Cambridge: Cambridge University.

Ekman, P. (Ed.) (1998) Third Edition of Charles Darwin's *The Expression of the Emotions in Man and Animals*, with introduction, afterwords, and commentaries. London: HarperCollins; New York: Oxford University Press.

Ekman, P., Davidson, R.J., & Friesen, W.V. (1990). The Duchenne smile: Emotional expression and brain physiology II. Journal of Personality and Social Psychology, 58, 342-353.

Ekman, P. & Friesen, W.V. (1978). Facial action coding system. Palo Alto: Consulting Psychologist Press.

Ekman, P. & Friesen, W. (1982). Rationale and reliability for EMFACS Coders. Unpublished.

Ekman, P. & Friesen, W.V. (1986). A new pan-cultural facial expression of emotion. Motivation and Emotion, 10, 159-168.

Ekman, P. Friesen, W. V. & Ancoli, S. (1980) Facial signs of emotional experience. Journal of Personality and Social Psychology. 39, 1125-1134.

Ekman, P. & Friesen, W.V, & O'Sullivan, M. (1988). Smiles when lying. Journal of Personality and Social Psychology, 54, 414-420..

Ekman, P., Friesen, W., & Simons, R. (1985). Is the startle reaction an emotion? Journal of Personality and Social Psychology, 49, 1416-1426.

Ekman, P., Friesen, W., & Tomkins, S.S. (1972). Facial affect scoring technique (FAST): A first validity study. Semiotica, 3, 37-58.

Ekman, P., Levenson, R.W., & Friesen, W.V. (1983). Autonomic nervous system activity distinguishes among emotions. Science, 21, 1208-210.

Ekman, P. & Rosenberg, E.L. (Eds.), (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). NY: Oxford.

Ellgring, H. (1986). Nonverbal expression of psychological states in psychiatric patients. European Archives of Psychiatry and Neurological Sciences, 236, 31-34.

Engen, T. (1982). The Perception of Odors. New York, NY: Academic Press.

Fleiss, J.L. (1981). Statistical methods for rates and proportions. NY: Wiley.

Fox, N. & Davidson, R.J. (1988). Patterns of brain electrical activity during facial signs of emotion in ten-month-old infants. Developmental Psychology, 24, 230-236.

Frank, M.G. (1996). Assessing deception: Implications for the courtroom. The Judicial Review, 2, 315-326.

Frank, M.G., Ekman, P. (1997). The ability to detect deceit generalizes across different types of high stake lies. Journal of Personality and Social Psychology, 72, 1429-1439.

Frank, M.G., Ekman, P., Friesen, W.V. (1993). Behavioral markers and recognizability of the smile of enjoyment. Journal of Personality and Social Psychology 64, 83-93.

Fridlund, A.J. (1992). The behavioral ecology and sociality of human faces. In M.S. Clark (Ed.), Review of Personality and Social Psychology, 13, pp. 90-121.

Friesen,W.V. & Ekman, P. (1992). Changes in FACS scoring. Unpublished manuscript, University of California, San Francisco.

Gosselin, P., Kirouac, G., & Dore, F.Y. (1995). Components and recognition of facial expression in the communication of emotion by actors. Journal of Personality and Social Psychology, 68, 83-96.

Hager, J. & Ekman, P. (1985). The asymmetry of facial actions is inconsistent with models of hemispheric specialization. Psychophysiology, 22, 307-318.

Heller, M. & Haynal, V. (1997). Depression and suicide faces. In Ekman, P. & Rosenberg, E. (Eds.), (1997). What the face reveals: Basic and applied studies of spontaneous expressions using the Facial Action Coding System (FACS). (pp. 398-407). Oxford University Press: New York.

Hess, U. & Kleck, R. (1990). Differentiating emotion elicited and deliberate emotional facial expressions. European Journal of Social Psychology, 20, 369-385.

Hjortsjo, C.H. (1969). Man's face and mimic language. Cited in V. Bruce & A. Young (1998), In the eye of the beholder: The science of face perception. NY: Oxford.

Izard, C.E., (1979).The Maximally Discriminative Facial Movement Coding System (MAX), Newark, Del.: University of Delaware, Instructional Resource Center.

Izard, C.E., Dougherty, L.M., & Hembree, E.A. (1983). A system for identifying affect expressions by holistic judgments. Unpublished Manuscript, University of Delaware.

Kanade, T., Cohn, J.F., & Tian, Y. (2000). Comprehensive data base for facial expression. Proceedings of the 4th IEEE international conference on automatic face and gesture recognition, pp. 46-53, Grenoble, France. Available from http://www.cs.cmu.edu/~face

Katsikitis, M. & Pilowsky, I. (1988). A study of facial expression in Parkinson's disease using a novel microcomputer-based method. Journal of Neurology, Neurosurgery, and Psychiatry, 51, 362-366.

Levenson, R.W., Sher, K.J., Grossman, L.M., Newman, J., & Newlin, D.B. (1980). Alcohol and stress response dampening: Pharmacological effects, expectancy, and tension reduction. Journal of Abnormal Psychology, 89, 528-538.

Lien, J.J.J., Kanade, T., Cohn, J.F., & Li, C.C. (2000). Detection, tracking, and classification of subtle changes in facial expression. Journal of Robotics and Autonomous Systems, 31, 131-146.

Malatesta, C.Z., Culver, C., Tesman, J.R., & Shephard, B. (1989). The development of emotion expression during the first two years of life. Monographs of the Society for Research in Child Development, 54 (Serial No. 219).

Matias, R., Cohn, J.F., & Ross, S. (1989). A comparison of two systems to code infants' affective expression. Developmental Psychology, 25, 483-489.

National Advisory Mental Health Council, (1995) American Psychologist, 50, 838-845).

Oster, H., Hegley, D., & Nagel, L. (1992). Adult judgments and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas. Developmental Psychology, 28, 1115-1131.

Parke, F.I. & Waters, K.(1996). Computer facial animation. Wellesley, MA: A.K. Peters.

Rinn, W.E. (1991). Neuropsychology of facial expression. In R.S. Feldman & B. Rime (Eds.), Fundamentals of nonverbal behavior. NY: Cambridge University.

Rohsenow, D.J., Niaura, R.S., Childress, A.R., Abrams, D.B., & Monti, P.M. (1990-1991). Cue reactivity in addictive behaviors: Theoretical and treatment implications. International Journal of the Addictions, 25, 957-993.

Rosenberg, E. (1997). Introduction: The study of spontaneous facial expressions in psychology. In Ekman, P. & Rosenberg, E. (Eds.), (1997). What the face reveals: Basic and applied studies of spontaneous expressions using the Facial Action Coding System (FACS). (pp. 3-18). Oxford University Press: New York.

Rosenberg, E., & Ekman, P. (1994). Coherence between expressive and experiential systems in emotion. Cognition and Emotion, 9, 33-58.

Rozin, P., Lowery, L., & Ebert, R. (1994). Varieties of disgust faces and the structure of disgust. Journal of Personality and Social Psychology, 66, 870-881.

Ruch, W. (1993). Extraversion, alcohol, and enjoyment. Personality and Individual Differences, 16, 89-102.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychological Bulletin, 114, 102-141.

Russell, J.A., & Fernandez-Dols, J.M. (1997). The psychology of facial expression. Cambridge, UK.

Sayette, M.A., Martin, C.S., Wertz, J.M., Shiffman, S., & Perrott, M.A. (manuscript under review). A multi-dimensional analysis of cue-elicited craving in heavy smokers and tobacco chippers.

Sayette, M.A., & Hufford, M.R. (1995). Urge and affect: A facial coding analysis of smokers. Experimental and Clinical Psychopharmacology, 3, 417-423.

Sayette, M.A., Martin, C.S., Perrott. M.A., Wertz, J.M., & Hufford, M.R. (in press). A test of the appraisal-disruption model of alcohol and stress. Journal of Studies on Alcohol.

Sayette, M.A. & Parrott, D.J. (1999). Effects of Olfactory Stimuli on Urge Reduction in Smokers. Experimental and Clinical Psychopharmacology, 7, 151-159.

Sayette, M. A., Smith, D. W., Breiner, M.J., & Wilson, G. T. (1992). The effect of alcohol on emotional response to a social stressor. Journal of Studies on Alcohol, 53, 541-545.

Sayette, M.A., & Wilson, G.T. (1991). Intoxication and exposure to stress: The effects of temporal patterning. Journal of Abnormal Psychology, 100, 56-62.

Smith, C.A. (1989). Dimensions of appraisal in psychological response in emotion. Journal of Personality and Social Psychology, 56, 339-353.

Soussignan, R. & Schaal, B. (1996). Children's responsiveness to odors: Influences of hedonic valence of odor, gender, age, and social presence. Developmental Psychology, 32, 367-379.

Steele, C.M. & Josephs, R.A. (1988). Drinking your troubles away II: An attention-allocation model of alcohol's effect on psychological stress. Journal of Abnormal Psychology, 97, 196-205.

Tian, Y.L, Kanade, T., & Cohn, J.F. (in press). Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23. Available from http://www.cs.cmu.edu/~face

Tronick, E., Als, H., & Brazelton, T.B. (1980). Monadic phases: A structural descriptive analysis of infant-mother face-to-face interaction. Merrill-Palmer Quarterly of Behavior and Development, 26, 3-24.

Vrana, S.R. (1993). The psychophysiology of disgust: Differentiating negative emotional contexts with facial EMG. Psychophysiology, 30, 279-286.

Table 1.

Kappa coefficients for single action units.

AU	Facial muscle	Description of muscle movement	Tolerance Window (seconds)				
			Frames	1/30 th	1/6th	1/3rd	1/2
0	Not Applicable	Neutral, baseline expression					
1	Frontalis, pars medialis	Inner corner of eyebrow raised	5124	0.73	0.79	0.81	0.83
2	Frontalis, pars lateralis	Outer corner of eyebrow raised	386	0.66	0.71	0.74	0.76
4	Corrugator supercilii, Depressor supercilii	Eyebrows drawn medially and down	323	0.58	0.64	0.67	0.70
5	Levator palpebrae superioris	Eyes widened	480	0.68	0.76	0.79	0.82
6	Orbicularis oculi, pars orbitalis	Cheeks raised; eyes narrowed	201	0.72	0.78	0.82	0.85
7	Orbicularis oculi, pars palpebralis	Lower eyelid raised and drawn medially	136	0.44	0.49	0.53	0.56
9	Levator labii superioris alaeque nasi	Upper lip raised and inverted; superior part of the nasolabial furrow deepened; nostril dilated by the medial slip of the muscle	48	0.67	0.76	0.81	0.83
10	Levator labii superioris	Upper lip raised; nasolabial furrow deepened producing square-like furrows around nostrils	96	0.69	0.76	0.79	0.81
11	Levator anguli oris (a.k.a. Caninus)	Lower to medial part of the nasolabial furrow deepened	<48	--	--	--	--
12	Zygomaticus major	Lip corners pulled up and laterally	800	0.67	0.71	0.74	0.76
13	Zygomaticus minor	Angle of the mouth elevated; only muscle in the deep layer of muscles that opens the lips	<48	--	--	--	--
14	Buccinator	Lip corners tightened. Cheeks compressed against teeth	168	0.59	0.67	0.72	0.75

15	Depressor anguli oris (a.k.a. Triangularis)	Corner of the mouth pulled downward and inward	71	0.54	0.65	0.69	0.72
16	Depressor labii inferioris	Lower lip pulled down and laterally	<48	--	--	--	--
17	Mentalis	Skin of chin elevated	136	0.55	0.63	0.66	0.68
18	Incisivii labii superioris andIncisivii labii inferioris	Lips pursed	57	0.65	0.71	0.74	0.75
19	Not Applicable	Tongue show	50	0.81	0.94	0.98	0.99
20	Risorius w/ platysma	Lip corners pulled laterally	162	0.38	0.47	0.54	0.60
22	Orbicularis oris	Lips everted (funneled)	<48	--	--	--	--
23	Orbicularis oris	Lips tightened	63	0.32	0.41	0.47	0.53
24	Orbicularis oris	Lips pressed together	259	0.50	0.58	0.62	0.64
25	Depressor labii inferioris, relaxation of mentalis, or orbicularis oris	Lips parted	789	0.57	0.62	0.65	0.67
26	Masseter; relaxed temporal and internal pterygoid	Jaw dropped	696	0.65	0.72	0.76	0.79
27	Pterygoids and digastric	Mouth stretched open	<48	--	--	--	--
28	Orbicularis oris	Lips sucked	90	0.61	0.70	0.76	0.79
41	Relaxation of levator palpebrae superioris	Upper eyelid droop	<48	--	--	--	--
42	Orbicularis oculi	Eyelid slit	<48	--	--	--	--
43	Relaxation of levator palpebrae superioris; orbicularis oculi, pars palpebralis	Eyes closed	<48	--	--	--	--
44	Orbicularis oculi, pars palpebralis	Eyes squinted	<48	--	--	--	--
45	Relaxation of levator palpebrae superioris; orbicularis oculi, pars palpebralis	Blink	<48	--	--	--	--
46	Relaxation of levator palpebrae	Wink	<48	--	--	--	--

	superioris; orbicularis oculi, pars palpebralis

Table 2.

Kappa coefficients for 3- and 5-point intensity scoring.

		Tolerance Window (seconds)							
		3-Point Intensity Scale				5-Point Intensity Scale			
AU	Frames	1/30th	1/6th	1/3rd	1/2	1/30 th	1/6th	1/3rd	1/2
10	96	0.63	0.70	0.74	0.77	0.61	0.67	0.70	0.72
12	800	0.69	0.74	0.77	0.79	0.57	0.61	0.63	0.66
15	71	0.48	0.60	0.65	0.68	0.44	0.53	0.57	0.59
20	162	0.36	0.46	0.53	0.59	0.31	0.39	0.45	0.49

Table 3.

Kappa coefficients for emotion-specified combinations

		Tolerance Window (seconds)			
	Frames	1/30th	1/6th	1/3rd	1/2
Action Unit Aggregates					
Positive emotion	335	.71	.78	.81	.83
Negative emotion	313	.64	.74	.79	.82
Disgust	103	.75	.82	.85	.86
Sadness	37	.47	.61	.67	.73

Note. Results shown for 3-point intensity scoring.