4. Does the face provide accurate information?

PAUL EKMAN, WALLACE V. FRIESEN, AND PHOEBE ELLSWORTH

The question of whether the face can provide accurate information about emotion has been the central issue since the beginning of research on the face. Although there may even well be legitimate research questions if the face provides only inaccurate information, for example, in understanding a source of stereotyping and misinformation in person perception, the determination of accuracy has been pivotal in the ebb and the flow of research activity on the face. In this chapter we shall document our claim that there is now sufficient evidence of accurate information to merit renewed and vigorous research on the face and emotion. In so doing, we shall directly challenge the misinterpretations, based in part on misinformation, provided by past reviewers of this work.

The reader may wish to refer to Chapter 1, where the problems of establishing accuracy criteria were discussed. *Accuracy* was defined as correct information of some nature being obtained by some means from facial behavior. As such, accuracy does not necessarily entail accurate information about emotion; in addition to the finding of accuracy, relevance of the accuracy to some aspect of the phenomena described as emotional must be demonstrated. Major methodological problems encountered in the main types of accuracy criteria, which will be discussed in this section, were reviewed. In this chapter, we shall say that an investigator obtained evidence of accuracy whenever the observers were correct more often than would be expected by chance ($p < .05$).

Although either a judgment or a component approach (see Chapter 2) could be employed in experiments on accuracy, almost all of the research has used a judgment design. We shall consider these first and then discuss the few studies that examined accuracy by measuring facial components.

4.1. Can judgments of emotion be accurate?

The question of whether observers could make accurate judgments of emotion was the key issue in the first period of research from 1914 to 1940, and the answer to this question was an important determinant of interest in the face and emotion, except for those who turned to the study of the vocabulary of emotion judgments. In their highly influential reviews of this literature, Bruner and Tagiuri (1954, p. 635) and Tagiuri (1968, p. 399) wrote:

> Some writers have reported that, whatever the nature of the expressive stimulus, the number of correct recognitions of emotions on the part of their subjects did not exceed the number that would be expected on a chance basis (for example, Fernberger, 1927, 1928; Guilford, 1929; Jarden & Fernberger, 1926; Landis, 1924, 1929; Sherman, 1927a – all of whom employed photographs of real emotions elicited in the laboratory). Others have shown that emotional expressions can be labeled with considerable accuracy (for example, Darwin, 1872; Feleky, 1914; Goodenough, 1931; Langfeld, 1918; Levitt, 1964; [P.K.] Levy, 1964; Munn, 1940; Ruckmick, 1921; Schulze, 1912; Stratton, 1921; [D.F.] Thompson and Meltzer, 1964; Woodworth, 1938).

It is no wonder that investigators might lose interest in this uninviting topic or, at the least, in the question of whether the face provides valid information about emotion. But Bruner and Tagiuri were factually incorrect and misleading. They enhanced the credibility of the negative findings on accuracy by saying that all of these experimenters utilized photographs of real emotions elicited in the laboratory. This is true only of Landis and Sherman. Fernberger, Guilford, and Jarden and Fernberger, whom they also credited with such laudable research methods, instead studied artists' drawings, not photographs, of posed or remembered behavior, not of real emotions elicited in the laboratory. Guilford studied the Rudolph faces, which are sketches made from photographs of an actor posing; Fernberger and Jarden and Fernberger studied the Boring and Titchener (1923) version of Piderit's drawings of the face, which presumed to show emotions in terms of separate facial features. Earlier (Section 2.7), we discussed the reasons why the use of drawings of the face have only the most dubious relevance to studies of accuracy and why we have excluded all such studies from this review. Among the studies cited by Bruner and Tagiuri as providing positive evidence on accuracy, we have excluded Langfeld because he also used the Rudolf faces, Levy and Stratton because they did not study the face, Schulze because his book was not available, and Ruckmick because he used only nine observers. The negative studies remaining in Bruner and Tagiuri's list, those of Landis and Sherman, were widely criticized and at least partially contradicted in the literature prior to their 1954 review.

No research on accuracy has completely satisfied the requirements outlined earlier in our methodological framework. Nevertheless, we shall show that reliable evidence of accurate judgment can be obtained for studies of posed behavior by taking into account findings across a number of experiments; judgments do coincide with the posers' intent. Although there is not as much evidence in regard to spontaneous behavior, what there is suggests a positive, rather than a negative, answer. Before considering these studies, however, we shall first analyze the Landis and Sherman experiments, bringing together past criticisms, our own framework, and relevant other experiments, all of which raise doubts about their findings, in an attempt to lay these two experiments finally to rest.

The Landis and Coleman experiments

In his 1924 and 1929 studies Landis took still photographs of his 25 subjects in a series of 17 situations, which included listening to music, looking at pornographic pictures, smelling ammonia, being shocked, decapitating a live rat, etc. Brief introspective reports were obtained after each situation, but these were kept short because Landis wanted a "cumulative disturbance" (an aim that provides the basis for one of the methodological criticisms to be discussed shortly). Four of the subjects were later asked to remember each situation and pose a facial behavior for each. Photographs were taken by the investigator when he noticed a change in the face. Landis selected 77 from the 844 photographs for use in his judgment study; 56 were from the initial situations, 21 from the remembered, posed situations; the sampling included the behavior of 22 of the 25 stimulus persons. Landis said he selected pictures for use in his judgment study that he thought were expressive. Forty-two observers judged the photographs, describing in their own words the emotion felt by the stimulus person, the situation that might have elicited the reaction, and their feeling of certainty about their judgment.

Landis reported that the results clearly showed that the emotions judged and the situations described by the observers were completely irrelevant to both the actual and the posed situations as well as to the introspective reports made during the actual situations. He attributed the discrepancy between the inaccuracy of his observers and the accuracy found by other investigators to the latter having used posed facial behavior, considered by Landis to be a specialized, conventional, languagelike behavior, which does not occur when people actually feel emotion.

Although there are many grounds for criticizing Landis's experiment, we shall look into only three of these, which can, at least in part, be supported by reexamination of Landis's data.

The first criticism arose from Davis's (1934) reanalysis of Landis's data on the components of facial behavior (a separate study by Landis of the records from this experiment). Davis found a tendency for the behaviors shown in the later situations to correlate with each other more than with the earlier situations. Davis interpreted this as resulting from the cumulative effect of experiencing the various situations in Landis's experiment, and as previously noted, Landis purposely kept the subjects' self-reports brief in order to enhance a cumulative disturbance. If, however, the disturbance were cumulative, that is, if there were a tendency for the emotions experienced in one situation to carry over to the next and for a disturbed or stressed reaction to build, then it would have been extremely difficult for observers, when they saw the facial behavior from each situation, either to discriminate separate, different emotions or to guess the specific eliciting circumstance. Only if the situations were not cumulative, only if each situation elicited a different reaction, would there be a chance for the occurrence of differing facial responses that could provide a systematic basis for the observer to appraise the particular emotion or situation when viewing a particular face. (Coleman, 1949, took Davis's criticism of Landis's experiment seriously, built rest periods into his experiment to diminish any cumulative effect, and attributed his positive results to his succeeding in eliminating a cumulative effect. We shall consider Coleman's study shortly.)

A second criticism of Landis's experiment, first raised by Frois-Wittmann (1930), is that Landis's situations might not have elicited the same emotion in all of his subjects; Arnold (1960) and Honkavaara (1961) later raised the same question. Furthermore, it is possible that each situation might have evoked more than one emotion, either simultaneously as a blend or consecutively. Looking at pornographic pictures, for example, or reading Krafft-Ebing case histories might elicit disgust in one subject, happiness in another, disgust–anger in a third, etc. If such variations did occur, then it would be highly unlikely that observers would be able to achieve accuracy on one of Landis's two accuracy criteria, correctly guessing the eliciting circumstance. But his other accuracy criterion, the subjects' self-reports of their experiences, could provide a more useful basis for measuring accuracy because, if subjects responded differently and reported these differences, observers might correctly judge information related to these self-reports, even if they failed to identify the eliciting circumstance. Such was not the case, however; Landis explained failure on this second accuracy criterion as error inherent in such flimsy data as introspection. After first considering the criticism under review, that Landis's situations were not emotion-specific across subjects, we shall then consider a third criticism, that Landis's experimental procedures led his subjects to mask or control their facial behavior and not to reveal their actual feelings. If that were so, then self-reports would indeed be a poor source of information for establishing accuracy.

Let us examine now two aspects of Landis's data that support the second criticism, that the eliciting situations were not emotion-specific across subjects. Landis's listing of the subjects' introspective reports showed that in only 2 of his 17 situations did even half of his stimulus persons report feeling the same emotion. Furthermore, his observers failed to judge accurately the posed facial behavior that his subjects recreated for these situations, thus contradicting Landis's belief that stereotyped posed behaviors would be accurately judged. Landis did not attempt to explain why his observers were unable to judge his subjects' poses accurately. An explanation consistent with the second criticism is that most of his 17 situations were probably not associated with any single emotion for all of his posers, who thus emitted various facial behaviors, depending on the emotion they believed to be associated with a given situation. Landis's negative results on the judgment of spontaneous behavior in his situations would be more credible if he had found that poses of behavior thought to be relevant to these situations could be accurately judged.

The third basis for criticism, first made by Murphy, Murphy, and Newcomb (1937) and repeated by both Arnold and Honkavaara, is that Landis might have unintentionally encouraged his subjects to inhibit or mask their facial responses to his situations. Landis mentioned this possibility but, unaccountably, dismissed it. A number of aspects of the experimental setting indicate the operation of display rules, either to neutralize the facial responses or to mask them with a positive affect. All of Landis's subjects knew him; most were psychologists who had had other laboratory experiences. Not only did they know they were being photographed, but, because Landis had marked their faces with burnt cork in order to measure the components of facial behavior in his other use of these records, they knew that Landis was interested in their facial behavior.[1]

Some other aspects of Landis's results are also consistent with this criticism. Landis reported that smiles were frequent in all of his situations, though he was convinced that his subjects were not feeling happy. Landis interpreted this as evidence against the meaningfulness of the smile; we interpret it as possible evidence of masking. A second source of support for the contention that Landis's experiment encouraged the operation of masking and neutralizing display rules is the introspective data. More than a third of the 17 situations elicited a report of "no feeling" from the majority of the subjects. If that is true, Landis failed with some frequency to elicit emotion. However, these self-reports, which he did not believe, do appear improbable in view of the anecdotal evidence he provided and the experience of others with such eliciting circumstances. The alternative, then, is that the reports of no feeling suggest an unwillingness to acknowledge being emotionally aroused, which could well have been duplicated in the facial behavior. In a completely different situation, Ekman and Friesen (1969a) provided some evidence that when individuals want to conceal their feelings and mask their emotions with a socially acceptable feeling, their faces typically either display deceptive masks or convey contradictory information. This may well have happened in Landis's experiment.

In summary, Landis's findings, that observers could not make accurate judgments, as compared with either the expected emotional nature of the eliciting circumstance or the subject's self-reported experience, should be credited only if (1) the same or similar reactions were elicited during at least some of the situations in most of the subjects, (2) the elicited reactions were different for at least some of the different situations, and (3) the selection of subjects and experimental arrangements did not encourage the subjects to mask or otherwise to control their facial behavior and/or to falsify their self-report.[2] The three criticisms discussed suggest that these conditions were probably not met.

[1] It is interesting to note that Hunt (1941), a supporter of Landis, belittled the importance of Munn's (1940) accuracy findings on the grounds that his subjects might have known that they were being photographed but failed to consider that this criticism was even more applicable to Landis.

[2] Our analysis does not presume that all of these problems operated in a similar fashion for all subjects. Some might have shown cumulative disturbance; others might have been more concerned with concealing their responses; it is also possible that concealment was more salient at the beginning of the experiment for most subjects, whereas toward the end, the effects of the cumulative disturbance became manifest.

The last argument against Landis's findings is based on Colman's (1949) study using comparable eliciting circumstances in which the observers did achieve accuracy. Coleman took motion pictures of the facial responses of six men and six women to eight situations, comparable to Landis's and of their subsequent attempts to pose the appropriate facial response for each situation. Because Coleman's interest was in comparing judgments of the top and bottom half of the face, which required the tedious chore of blacking out part of each motion picture frame, he utilized the films of only 2 of the 12 persons in his judgment experiment. He selected stimulus persons whose behavior he believed to be natural, not exaggerated, but showing a variety of expressions, and whose self-reports revealed that they were strongly affected by the experiment. Later we shall consider his results on the judgments of the partial faces and consider now only his results on the judgments of the full faces.

The motion pictures of the natural, or original, responses and the posed responses of two subjects were shown to 379 observers, who judged which of nine situations was the one in which the pictures were taken; (in addition to the eight actual situations, a ninth was added to decrease judgment by elimination). Judgments were accurate, in that the correct situations were identified for each stimulus person, in both the natural and posed versions, more than would be expected by chance. Posing enhanced accuracy, but for different situations for the two stimulus persons.

Why did Coleman obtain positive and Landis negative results?

First, Coleman's situations might have elicited more similar emotions across his two stimulus persons, with different reactions to at least some of the situations by both subjects. Coleman did, after all, purposefully select subjects whom he thought had been affected by the experiment, which might mean that he picked those who had shown or reported different experiences across his eight situations. Coleman explained the difference in his results from Landis's as a result of his inclusion of rest periods to diminish any cumulative disturbance, which might obscure different reactions to his eight situations. Evidence in support of the possibility that most of Coleman's situations were associated with different emotions (and by both subjects) comes from his positive results on the judgment of posed behavior. Landis's failure to obtain accuracy in judgments of posed behavior suggests, it will be recalled, that his subjects might not have shared the same emotional experience and therefore did not attempt to pose the same emotions. The fact that Coleman did obtain accuracy in judgments of posed behavior suggests, contrarily, that at least some of his situations were associated with the same emotion by his two subjects.

Second, Coleman's judgment task may have allowed for more complex or inferential judgments in that he did not ask his observers to judge emotion, as did Landis, but instead to pick the situation during which a film was taken. If an observer sees a subject smile and knows nothing of the situation, he may simply call that smile "happy."[3] But if he knows the nature of the various eliciting situations, he may well consider the possibility that the smile is a mask or an embarrassed reaction to the feelings elicited by one of the situations (e.g., Coleman's situation of crushing a snail). Coleman's judgment task of matching a situation with a film clip leaves open, however, the question of whether the accuracy obtained was dependent on information about emotion. Perhaps accurate judgments could have been based on behavior not usually described as emotional, i.e., defensive behavior. It would be helpful if Coleman had shown the same films to another group of observers and asked them to judge the emotions shown.

Third, Coleman showed motion pictures, whereas Landis obtained judgments on still photographs. Earlier, we argued that film or videotape were more appropriate than stills for recording spontaneous behavior because they do not fragment the natural flow of behavior. Still photographs are more appropriate for recording static poses. Although Landis fired his camera whenever he thought something was happening, his basic unit, a single still, would provide less information about the onset and duration of the facial response, even if his own reaction time was so rapid that he adequately captured the response at its apex, or most extreme moment. The type of complex, inferential judgment referred to previously would probably be easier to make from films, which show the sequence of reactions, duration, etc., than from still photographs.

Fourth, Coleman's subjects might have been less motivated to inhibit or mask their facial behavior than Landis's subjects. Unlike Landis, Coleman did not select subjects who knew him nor did he mark their faces; thus he may have avoided Landis's error of inadvertently encouraging the operation of those display rules that inhibit candid facial responses.

Certainly Coleman's study is far from conclusive as an accuracy study. The generality of the findings across persons cannot be determined with only two stimulus persons nor, for reasons just mentioned, can it be

[3] In instances where it is unavoidable, the masculine pronoun has been used in the generic sense to mean he or she.

determined that the accuracy obtained necessarily involved judgments of emotion. But Coleman's study does underline the three major methodological problems in Landis's experiment and strengthens our contention that the results of Landis's experiment should be discredited. Correcting some of the flaws in Landis's experiment, Coleman was able to achieve accuracy. It is regrettable that no further work has been done utilizing such eliciting circumstances, for the most conclusive evidence for discounting Landis' findings would be further studies similar to Coleman's with a larger number of stimulus persons.

The Sherman experiment

As previously mentioned, the only other study that obtained negative results on accuracy among those described by Bruner and Tagiuri (1954) and Tagiuri (1968) as utilizing real behavior elicited in the laboratory, and that actually did so, was Sherman's (1927a) study of observers' judgments of very young infants. His research has remained influential, despite challenges to his interpretation of his results as long ago as 1931 by Goodenough, again by Murphy, Murphy, and Newcomb (1937), and by Honkavaara (1961) and despite the results of other inquiries contradicting his findings. As with Landis, we shall critically examine his results in detail, not because of the merits of his study, but because of its continuing acceptance despite enormous flaws. We shall suggest a number of grounds for challenging the validity of his results, as well as his interpretation of them, and shall present a summary of contradictory results from a number of studies from the same era.

Sherman actually performed four separate experiments, with different conditions, only one of which is sufficiently free of confounding sources of variability to be considered. In that experiment, he recorded, on motion picture film, the behavior of two infants, one 74 hours old and the other 145 hours old, as they were subjected to four eliciting circumstances: hunger, defined as prolonging the time for the infants' scheduled feeding by 15 minutes;[4] suddenly dropping the infant; restraining the infant by holding the head and face down on a table; and applying a needle to the cheek six times. Two groups of observers (graduate and freshmen psychology students) were shown the behavior immediately after the elicitation and asked to judge the emotion and the eliciting circumstance in their own words.

His second experiment was said to be the only instance in which accuracy and agreement were achieved. Here the observers saw, not only the postelicitation films, but also the filmed behavior during the elicitation itself. The results are difficult to interpret for the following reasons: These observers had two, not one, sources of additional information – knowledge of the eliciting circumstance, which Sherman intended, and access to the full range of the infants' facial behavior during the elicitation procedure, which may have been redundant with or different from the postelicitation facial behavior shown in the first experiment. Another problem in comparing the results from the first and second experiments is that Sherman coached the subjects in the second, but not in the first experiment, about what behaviors to observe. A third problem is that half of the observers in the second experiment were observers in the first study. Thus any difference in performance between the first and second experiments could be, not only because of coaching, knowledge of the elicitor, or exposure to additional facial behavior during elicitation, but also because of the benefits of practice or memory for those observers who were already in the first experiment.

The third set of data was gathered from medical students and nurses who were shown the live, not the recorded, postelicitation behavior of an unspecified number of infants. (A screen blocking their view was present during elicitation and then removed.) We reject these data because the judgments are confounded by the observers' exposure not only to the facial behavior as in the first study but also, as Sherman readily acknowledged, to the vocal behavior; all infants cried during postelicitation. We should note that these two groups of observers were small and that, in most instances, the majority of their judgments were of events, not emotions, e.g., colic, just awakening, tight bandage.

The last study is one in which the eliciting circumstance film from one situation was followed by the postelicitation film from another situation as if they were sequential. This experiment suffers from most of the same flaws as that in which the observers saw the films of the actual eliciting circumstance in addition to the postelicitation films.

Based on methodological flaws in all of the other experiments, our decision to consider only the data from the first experiment, in which observers saw only the postelicitation behavior, does not ignore the results Sherman himself considered crucial. Sherman interpreted the

[4] Sherman did not explain the basis for the presumption that, in infants of that young age, the feeding schedule would be sufficiently established for a 15-minute deviation to make a sufficient difference to elicit any reaction. It is to be hoped that Sherman had other bases for knowing that his two infants were hungry at the time of the experiment.

data from that experiment as most damaging to any claim that accurate judgments could be made from infants' facial behavior.

There are three major flaws in Sherman's first experiment that serve to raise serious doubt about the validity of his conclusion. First, Sherman's data analysis is oversimplified. He did not distinguish between judgments of emotion, utilizing the usual emotion vocabulary, and judgments of events or internal states, i.e., taking medicine or being hungry. Instead, he counted both types of responses in his measures of whether observers could make accurate emotion judgments. Further, he ignored the possibility that some of the emotion terms might have been synonyms, although F. Allport had published his emotion categories a few years earlier (1924); thus, Sherman considered rage and anger, for example, as different emotion judgments and pain and hurt as different judgments.

The second and more serious criticism of Sherman's experiment is the probability that all four situations might have elicited the *same* reaction from the infants and, of course, as explained earlier in connection with Landis, if the situations do not elicit different reactions, then there is little reason to expect the observer to make different judgments. In other words, Sherman's accuracy criterion was the emotion *he* expected in each of his four situations; but he provided no empirical basis for his expectation and, if he were wrong, if the situations all elicited the same response, then the failure to find accurate judgments would be meaningless. Two aspects of Sherman's data suggest that the reactions across the four stimulus situations were similar. Anger was the most frequent judgment for all four of his situations, which Sherman saw as evidence of inaccuracy; but quite conceivably his infants *were* angry, either during all four situations or after each of the four stimulus situations. For it must be remembered that we are considering judgments made of postelicitation behavior; it is possible that each of the four elicitations immediately produced a distinctive response, was rapidly dissipated a few moments later and not seen in the postelicitation behavior, which may well have shown anger. Honkavaara (1961) might have had this in mind in criticizing Sherman for sampling infant behavior only during crying. From Sherman's other article (1927b), describing a separate study in which observers listened to the sounds made by the infants but did not see the films (again, failing to achieve accuracy in terms of Sherman's expectations), Honkavaara discovered that the infants in the study we are discussing here cried in all four postelicitation situations. Although crying does not necessarily signify that *only* anger (or only some other emotion) was present in the postelicitation behavior, one must question whether Sherman really did succeed in preserving different emotions for the four situations, without which he could have no basis for interpreting his results as showing inaccuracy.

The third criticism of Sherman's experiments, raised by Goodenough (1931), Murphy et al. (1937), and Honkavaara (1961), involves the age of the infants, both less than one week old. They argue that the failure of such young infants to show a differentiated facial response across the four eliciting situations would not be a conclusive demonstration that the face is unrelated to emotion nor that social learning provides the sole basis for any such relationship, which might exist later. Maturation may be such that the differentiated perception of the situations necessary for differential facial response, or the differentiation of the facial responses themselves, is not unfolded prior to an age of 150 hours.

The last criticism is that a sample of two infants is far too small to permit any conclusions; and, further, Sherman did not report his data separately for the judgments of the 74-hour-old and the 145-hour-old infant, so it cannot be determined whether the observers' judgments were similar for both or different because of maturational or other factors.

Although we shall not consider as a separate substantive issue the nature of the development of facial behaviors associated with emotion, we shall explore four articles reporting findings that indirectly contradict Sherman's to complete the case we have constructed for dismissing Sherman's experiment.

Goodenough (1931) showed eight photographs of a 10-month-old infant to 68 observers. The observers were given a choice among 12 possible judgments, each judgment describing both an emotion and an eliciting situation. There were four more choices provided than stimuli to decrease the chance that the observers would choose by elimination. Goodenough reported that 47% of the judgments were accurate. In our reanalysis of her data, we first discarded all of the stimuli and judgments involving the description of facial appearance rather than an inference about emotion (satisfied smile; roguish smiling; crying). We also omitted two stimuli described too generally to be relevant to accuracy regarding specific emotions (grimacing, dissatisfaction). This left three stimuli that were relevant to the question of whether accurate judgments could be made of the infant's emotions. In considering these data, we counted a judgment as correct if it accurately identified the emotion, regardless of whether or not it included a correct identification of the particular eliciting situation. For example, for the photograph taken when the infant had been astonished by the sight of a bright-colored toy, we decided that the observers were accurate if they chose that judgment or if they chose the judgments of "astonishment at the mother counting loudly," or "astonishment while listening to the ticking of a watch." The results were as follows: 94% correct judgments for the astonishment face; 79% correct judgments for the pleasure face; and 21% correct judgments for the fear face. With only one stimulus person and only three relevant stimuli, Goodenough's study was certainly enormously limited but showed accuracy on two out of three emotions tested.

In another article, Goodenough (1932/1933) reported data she felt contradicted Sherman's. She observed a 10-year old, blind–deaf child who, because of these handicaps, had little opportunity to learn facial behaviors. Goodenough dropped a doll inside the child's dress and described the facial reactions as being similar to the facial behaviors associated with different emotions in normal children. This is interesting anecdotal evidence of an innate tendency to show emotion in the face, but the nature of the data reported limits it to that.

J. Thompson (1941) and Fulcher (1942) both conducted studies of blind and sighted children, and we shall revaluate them in Section 4.2 with regard to the components of facial behavior. Thompson studied spontaneous behavior in 26 blind and 29 sighted children ranging from 7 weeks to 13 years of age; Fulcher studied the posed emotions of 50 blind and 118 sighted children ranging from 4 to 21 years old. Both noted maturational factors; both noted, in their analysis of the components of facial behavior, similarities between the blind and sighted children in at least some emotional facial behaviors – laughing, smiling, crying, and anger for Thompson and happiness, sadness, anger, and fear for Fulcher. Although differences in the extent of muscular movements were found between blind and sighted children, there was evidence of similarity in the particular muscles moved for each emotion.

Both investigators also utilized a judgment procedure with observers. Thompson had four trained psychologists observe the facial behavior of the blind and sighted children and judge emotion in 11 categories, which she then analyzed in terms of F. Allport's category scheme. There was both high agreement among observers and accuracy between observers and investigator in that the emotion judged corresponded with the investigator's impression, based on situational context as well as total behavior shown, for judgments of both blind and sighted children with no significant differences as a function of sightedness. Fulcher had five observers who knew the intended emotion judge the "adequacy" of the pose. For both blind and sighted children, happiness and sadness were judged as more adequately portrayed than anger and fear, but across all emotions the sighted portrayals were judged as more adequate than the blind ones. In discussing their results, Thompson and Fulcher cited Landis and Sherman as providing the main evidence for contradiction of their findings.

To summarize our discussion of Sherman's work and related studies, Sherman's evidence for inaccuracy rests on the presumptions that a different emotion was elicited in each situation, that it was the same emotion for each of the two infants, and that the emotion elicited during the situational manipulation was preserved long enough to appear in the postelicitation film shown to the observers. These presumptions are of dubious validity. Further, the validity of Sherman's findings can be questioned because of his failure to consider maturational processes and the small size (two) of his sample of stimulus persons. A brief review of studies by Goodenough, Thompson, and Fulcher, all studies of children's facial behavior that considered blind as well as sighted children and some of which studied different age periods, shows consistent contradiction of Sherman's conclusions.

The Landis and the Sherman experiments, with their questionable negative findings, have, in our opinion, had unmerited influence in the investigation of judgment of emotion from facial behavior. Our lengthy discussion of these studies has been an attempt to set them in perspective. We shall now turn to the positive evidence.

The first set of studies (Hanawalt, 1944; Munn, 1940; Vinacke, 1949) used as stimuli commercial magazine photographs of presumably spontaneous, naturally occurring, emotional behavior. Accuracy was measured in terms of the observers' ability to judge the emotion that presumably occurred in the situation. In the second set of studies (Ekman & Bressler, 1964; Ekman & Friesen, 1965b; Howell & Jorgenson, 1970; Lanzetta & Kleck, 1970), spontaneous reactions elicited in a standard stress interview, spontaneous behavior when anticipating electric shock, and clinical interviews with psychiatric patients were used as stimuli. Accuracy was measured in terms of the correspondence between the observers' judgments of emotion and anticipated differences between the stress and catharsis portions of standard interviews, between the pre- and posthospitalization interviews of psychiatric patients, and between observers' judgments of the eliciting circumstance, shock or nonshock. The last set of studies (Drag & Shaw, 1967; Dusenbury & Knower, 1938; Ekman & Friesen, 1965; Frijda, 1953; Kanner, 1931; Kozel & Gitter, 1968;

Levitt, 1964; Osgood, 1966; D.F. Thompson & Meltzer, 1964; Woodworth, 1938) used posed behavior as stimuli. Accuracy was measured in terms of the observers' success in judging the emotion intended by the poser.

Accuracy in judging candid photographs: Munn, Hanawalt, and Vinacke

Munn (1940) explained his decision to have observers judge the emotion shown in magazine photographs, taken during presumably emotional situations, as an attempt to resolve Woodworth's (1938) doubts about whether accuracy was possible for spontaneous as well as posed facial behavior. Munn's primary aim was to determine the influence of knowledge of the situation upon judgment of emotion from facial behavior by comparing the judgments of observers who saw the face alone with those who saw the entire photograph. Though he found that the number of observers making an accurate judgment increased when the entire picture was seen, accuracy was achieved with most of his stimuli even when only the face was seen. These comparative results will be discussed later in connection with the question of how contextual information influences the judgment of emotion from facial behavior (Chapter 6). We shall take into account here only Munn's data on the judgments of the face alone.

Hanawalt (1944) borrowed Munn's procedure of utilizing candid photographs from magazines and used a number of Munn's actual stimuli in addition to ones of his own. His purpose was not to study accuracy but to compare judgments made when either the top or bottom half of the face was seen; these results will be considered later in connection with the question of how judgments of emotion are influenced by the components of the face observed (Chapter 5). Only Hanawalt's results on the judgments of the full face will be considered here.

Vinacke (1949) also drew stimuli from magazines but chose his own different set of pictures. His purpose was not to study accuracy but to compare judgments made by different ethnic groups; those results will be scrutinized later in connection with the question of how judgments of emotion may vary across cultures (Chapter 7). Here we shall consider only Vinacke's results on the judgments by Caucasian observers.

Munn recognized that there were difficulties with his two accuracy criteria. He should have sought some means for estimating what emotional reaction was experienced independent of the facial behavior which occurred. He could have approximated that by showing the full photographs but obscuring the face so that the observer could see only the situation and determine whether there were any agreed upon expectations about the probable emotion. (This procedure might not be workable if the photograph of the situation did not adequately show what the nature of the setting was or what the elicitor was, etc.; see Chapter 1). However, he did not do that; instead, both of his accuracy criteria were contaminated by knowledge of the face as well as knowledge of the situation. One criterion was his own expectation which was an unreliable basis because he was not present when the behavior occurred and was contaminated by his inspection of the faces. The other criterion was the judgment of the observers who saw the situation and the face; it was similarly contaminated so that it is not possible to know whether their expectation about emotion was influenced primarily by the situation or by their judgment of the face. Neither Hanawalt nor Vinacke concerned himself with accuracy criteria; they analyzed their results solely in terms of observer agreement under the different conditions employed in their experiments.

The best basis for building an accuracy criterion for use with candid photographs from magazines is not available, precisely because the pictures are selected long after the event, with no access to the relevant sources of information, viz., the self-report of the individual, the reports of other people present in the situation, and the data on the antecedent and consequent events. Though less satisfactory, another basis for establishing an accuracy criterion for such stimulus materials is to determine what single emotion, if any, is usually associated with the situation in which the candid photograph appeared to have been taken. We conducted a simple experiment in which 35 college students were given the list (but no photographs) of situations as described by Munn, Hanawalt, and Vinacke and were asked to judge the most probable emotion, utilizing the list of proposed emotion categories from Table 3.2 and also the choice of "no emotion." Only situations yielding at least 50% agreement about a particular emotion were taken to be relevant for examining the accuracy of observers' judgments. The data on 5 of the 14 Munn stimuli were excluded because of lack of agreement about probable emotion as were the data on 7 of Hanawalt's 20 stimuli and 14 of Vinacke's 20 stimuli. Correspondence between the majority of original judgments of the faces in the photographs and the judgments we obtained of the verbal descriptions of the situations constituted the measure of accuracy. Table 4.1 gives the verbal description of the situation as it was noted in the articles by these authors and as it was given to our college students, the percentage agreement about the expected emotion made

Table 4.1. *Results on selected stimuli from candid photograph studies (in percentages)*

		Judgment of face		
Verbal description	Judgment of verbal description	Munn (1940)	Hanawalt (1944)	Vinacke (1949)
Girl laughing	100 Happy	—	97 Happy	—
Jitterbug clapping hands to music	97 Happy	86 Happy	—	—
Girl running into ocean	91 Happy	97 Happy	—	—
A man smiling standing between two other men	88 Happy	—	—	59 Happy
Baseball fan vociferously cheering	82 Happy	—	—	47 Happy
Girl in sack race	66 happy	49 Sad	—	—
Man escapes Nazis	56 Happy	—	65 Fear	—
Girl escapes explosion	53 Happy	—	96 Horror	—
Man in shower as water is unexpectedly turned on	89 Surprise	—	87 Surprise	—
Girl discovers photographer as she lifts hoop skirt to go through door	85 Surprise	—	80 Surprise	—
Girl in amusement park with dress going up	74 Surprise	—	90 Happy	—
Girl discovers photographer has her covered	58 Surprise	—	62 Surprise	—
Girl photographed over transom while dressing	56 Surprise	66 Surprise	89 Surprise	—
Girl photographed over transom while in bath	38 Surprise 12 Fear	26 Surprise 28 Fear	—	—
Girl running from ghost	96 Fear	94 Fear	92 Fear	—
Boy caught in revolving door attended to by policeman	61 Fear	—	—	1 Fear
Porter leading burned man from scene of airplane crash	56 Fear	8 Fear 33 Anxiety	—	—
Man with hand stretched toward hostile crowd	54 Fear	63 Distress/ Anxiety	—	—
Frenchman shows grief as colors of lost regiment are exiled to Africa	79 Sad	—	84 Sad	—
Man wrapped in blanket after failure to swim English Channel	71 Sad	—	—	4 Sad 16 Exhaustion

Table 4.1 (*cont.*)

		Judgment of face		
Verbal description	Judgment of verbal description	Munn (1940)	Hanawalt (1944)	Vinacke (1949)
Woman disheveled weeping telephoning	60 Sad	—	—	51 Sad
Girl sitting in a police station after one of her suitors was killed in a quarrel over her affections	53 Sad	—	—	32 Sad
Man who is holding strikebreaker by the coat collar	68 Anger	8 Anger	65 Anger	—
Lady awaiting news of mine disaster	94 Disgust/ Contempt	—	71 Sad	—

on the basis of reading the verbal descriptions and the judgments made by the original observers of the face alone. (These last data were reorganized in terms of the categories listed in Table 3.2 to facilitate comparisons across experiments.)

Accuracy was found with all three sets of data: for 6 of the 9 Munn stimuli listed, for 8 of the 12 Hanawalt stimuli, for 4 of the 6 Vinacke stimuli. The table also shows that this accuracy was achieved with happiness, surprise, fear, and sadness stimuli; the one anger and the one disgust/contempt stimulus yielded either inconsistent or inaccurate results.

The instances of inaccuracy are difficult to explain. Either set of observers could be wrong; or both could be correct if the stimulus person experienced more than one emotion and the camera captured the nonnormative one, or if the stimulus person had an idiosyncratic reaction, or if the verbal description of the situation failed to include relevant information. But this is the limitation of this indirect method of establishing an accuracy criterion.

To summarize, these three experiments show that observers can make accurate judgments of spontaneous behavior, in the sense that observers of a face can judge the emotion that other observers who read a description of the situation predict. There are four limitations on these results. First, the behavior studied (candid photographs taken from magazines) may not all actually have been spontaneous. The person shown in the photographs might have been aware of the photographer or, even worse, might have completely reenacted or staged the behavior for the press.

The second limitation is the accuracy criterion. Although the one we fashioned (determining what emotion would be expected in the eliciting circumstance) is preferable to the one employed by the original authors, it is still not totally satisfactory for reasons previously discussed.

A third limitation is the sampling of emotions; accuracy was shown for four of the seven emotion categories listed in Table 3.2; the sample for the other two emotion categories was very small (only one stimulus each).

The fourth and, perhaps, most serious limitation is in regard to the representativeness of the findings. Hunt (1941) appropriately noted the need to establish how often facial behavior in situations like those studied by Munn can provide the basis for accurate judgments of emotion. Are informative faces a rare event, usually lost within a sequence of noninformative facial behavior? Or is such informative facial behavior shown only by some special group of people, highly extroverted persons, for example, but not by a more representative sample? Did Munn or the photographer pick out the one rare moment, or the few rare people, who happen to provide accurate information in their spontaneous facial behavior? The answers to these questions would require information about sampling that is not available. Information would be needed both from the photographer (how the subject of the photograph was chosen and how the photographs that were published were chosen from those shot) and from Munn, Hanawalt, and Vinacke (how many published photographs they inspected in choosing the particular ones employed in their studies).

Although it can be said that accuracy did occur, it is not possible to specify how frequently facial behaviors in situations such as those studied by these authors provide the basis for accurate judgments. Thus, there is doubt about the representativeness of their findings, in terms of generality across persons and of generality across time within the situation. The next group of experiments to be considered remedies this limitation, but they are weaker than the ones just discussed in that accuracy was sought on gross discriminations rather than on specific emotions.

Accuracy in the judgment of spontaneous behavior

Let us first note how the design of experiments on spontaneous behavior, which we shall next examine, answers Hunt's criticism of the candid photograph studies by providing evidence of generality across persons and generality across time. The first question about generality is resolved if there is representative sampling of stimulus persons, that is, if the experimenter does not preselect some atypical group of people, who, because of special training, instruction, or proclivity, are likely to be more facially facile than others. Although there was no information about the sampling of persons in the candid photograph studies, in the studies of spontaneous behavior, the sampling of persons was reasonably random within the constraints of utilizing volunteers and the usual sources for subjects. The stimulus persons were either college students or mental patients, but in neither case were good expressors preselected. The number of stimulus persons in each experiment was small; but by considering the finding of accuracy across all of the experiments, this limit is remedied.

The matter of how to design experiments to resolve doubts about the generality of the findings across time is more complicated. Let us examine how answers could be furnished to a skeptic who, like Hunt, holds the view that the face rarely emits information that allows for accurate judgment. This skeptic would have no problem in dismissing the candid photograph studies discussed earlier, for there is no evidence on the sampling of behaviors in those studies to counter the skeptic's claim that both photographer and investigator probably chose that one-in-a-million slice of life in which the face happened to show something decipherable and relevant to the eliciting circumstance. If the skeptic is shown evidence of accuracy in an experiment where the investigator did not take a single slice but showed observers a continuous sample of some length on film or videotape (Howell and Jorgenson used 60-second samples, Lanzetta and Kleck used 12-second samples), then the skeptic would have to yield, but only somewhat. The skeptic could no longer claim that the rare moments when facial behavior is informative are usually lost when embedded in the total sequence of random, meaningless, facial behavior. If that were so, then observers who saw a sequence would not achieve accuracy. However, the skeptic could still argue that only in *rare* moments is the face informative, but rather than being lost, those rare moments provide the basis for accurate judgment. Perhaps one signal in 12 or 60 seconds of facial noise was the basis for accurate judgment, the skeptic would argue, and this provides no evidence that the face often conveys accurate information.

The answer to this remaining claim, that the face is an infrequent output system, requires an experiment in which many separate samples of facial behavior are randomly drawn from within an eliciting situation.

If observers are able to make accurate judgments for most of the samples, the skeptic is answered and the representativeness of findings in terms of generality across time is established. There is one artifact in such a research design that can decrease the probability of obtaining accurate results. Selecting slices of behavior in a random fashion may well fragment the natural flow of behavior. For example, a 5-second slice might show the end of one facial behavior and the beginning of another rather than the beginning, middle, and end of a facial behavior and thereby increase the difficulty of judging the behavior. Nevertheless, in the experiments by Ekman and his associates, randomly drawn multiple samples of behavior did provide the basis for accurate judgments, establishing that the face often provides accurate information, and thus answering doubts about the generality of findings across time.

Let us now consider this set of experiments. Tables 4.2 and 4.3 show the methodological features and results of these studies. In some of the studies, the sample of facial behavior was obtained by recording a naturally occurring event; Ekman and Bressler (1964) and Ekman and Rose (1965) used facial behavior shown during interviews conducted at different points in inpatient psychiatric hospitalization. Some utilized a laboratory-contrived situation to elicit emotion; Ekman (1965) and Howell and Jorgenson (1970) used standardized interviews in which the interviewer's manner and style changed; Lanzetta and Kleck (1970) used the anticipation of receiving either a shock or nonshock. Only the studies by Ekman and his associates were designed primarily to study accuracy, but the other studies do provide information relevant to this issue. For all but the Lanzetta and Kleck experiment, the observers who judged emotion knew nothing about either the situation in which the stimuli had been recorded or the nature of the persons photographed.

In the first study listed in Table 4.2, the accuracy criterion was the expected difference in emotions elicited by the interviewer being hostile (stress-inducing) and then explaining the purpose of his hostility and praising the subject for resiliency under stress (catharsis-inducing). The observers rated stimuli from the stressful parts of the interviews as more unpleasant than those from the cathartic parts of the interviews.[5] Whereas the difference in pleasantness rating is small, it is significant, and it should be remembered that these stimuli were selected at random. As mentioned earlier, selecting still photographs at random may well frag-

[5] There were also significant differences in the ratings on the attention–rejection dimension, but these were not reported because ratings on this scale were highly intercorrelated with ratings on pleasantness.

Table 4.2 [illegible] experiments of spontaneous behavior

	Ekman (1965)	Howell & Jorgenson (1970)	Lanzetta & Kleck (1970)
Methodology			
Number of stimulus persons	5	4	12
Number of stimuli	60: with 12 of each person, half from each of the 2 conditions	8: with two 1-minute film clips of each stimulus person one clip from condition one film clip from relief condition	—
Type of record	Still photographs of face, randomly selected from larger record	Motion picture film of head and shoulders	12-Second videotape of stimulus person while they anticipated shock or nonshock; each observer judged 20 episodes for each of 6 persons
Number of observers	35	53	6 Observers for each stimulus person
Judgment task	Ratings on Schlosberg's three dimensions: pleasant–unpleasant, attention–rejection, sleep–tension	Pleasant–unpleasant dichotomy	Whether the videotape shows shock anticipation trial or nonshock trial
Sampling situation	Standardized interview with stressful and cathartic parts	Standardized interview with stressful and relief parts	Stimulus persons saw a red light if they were to receive a shock, a green light if no shock on that trial
Accuracy criterion	Compare judgments of stimuli from two parts of interview expected to differ in emotion experienced	Compare judgments with emotions expected in two parts of the interview	Is judge correct in identifying whether person is in shock anticipation (red) or nonshock anticipation (green)?
Results	Median stress stimuli: 4.6 unpleasant Median cartharsis stimuli: 5.7 pleasant	61% Correct identification	Accuracy significant across stimulus persons observed, although also significant differences among stimulus persons observed. Chance judgment would be 50% correct, and the range across stimulus persons was 55% to 83% correct, with a median of 62%

Table 4.3. *Methodology and results of two accuracy experiments of spontaneous behavior*

	Ekman & Bressler (1964)	Ekman & Rose (1965)
Methodology		
Number of stimulus persons	10	6[a]
Number of stimuli	40 sequences; each was five photos taken in 5 seconds; four sequences for each person; half from each of the two conditions	96 Sequences, each was five photos taken in 5 seconds; 16 sequences for each person, half from each condition
Type of record	Five rapid stills, showing face and body, randomly selected from larger record	Five rapid stills, showing face and body, randomly selected from larger record
Number of observers	34	244
Judgment task	Ratings on: pleasant–unpleasant scale, mobile–immobile scale	Ratings on: pleasant–unpleasant scale, Immobile–mobile scale
Sampling situation	Standardized psychiatric interview with in-patients, two interviews with each patient, one rated as most depressed and one rated as most improved by interviewer	Standardized psychiatric interview with in-patients, two interviews with each patient, one at time of admission to hospital and one at time of discharge from hospital.
Accuracy criterion	Compare judgments of stimuli from the two interviews when different emotional experience would be expected	Compare judgments of stimuli from the two interviews when different emotional experience would be expected
Results	Median depressed stimuli: 3.9 unpleasant, 2.3 immobile, Median Improved stimuli; 5.1 pleasant, 3.7 immobile	Median admission stimuli: 2.8 unpleasant, 2.9 immobile median discharge stimuli: 5.0 pleasant, 4.9 mobile

Notes: [a]These six patients were also stimulus persons in Ekman & Bressler's (1964) study.

ment the natural flow of behavior and, for that reason, provide a low estimate of accuracy, but this random sampling procedure was used because the study was intended to evaluate what the face might show over the course of an entire interview, not what the face might show at its most informative moments. In another study designed to assess the *maximum* accuracy possible, Ekman used as stimuli photographs that observers who saw both face and body had rated as maximally stressful or cathartic. When just the faces of these pictures were shown to another set of observers who judged the pictures on Schlosberg's three dimensions of emotion without knowledge of the interview situation, the difference in pleasantness ratings was large – a mean difference of 3.5 points on a 7-point scale.

Howell and Jorgenson (1970) performed an experiment that was very similar in both the eliciting situation and judgment task. Their major interest, however, was in comparing accuracy when the observers saw the face, read or heard the words, or received a combination of sources. We shall report here only on their results on the judgment of the facial behavior. Their interviewer's behavior changed from unfriendly and challenging to reassuring, in order to induce stress and relief from stress. The observers were shown 60 seconds of motion picture film from the stress phase and a 60-second sample from the relief phase and were asked to judge whether the person felt pleasant or unpleasant. Despite differences in accuracy achieved for particular stimulus persons and differences in level of accuracy achieved for the relief compared with the stress sample, overall accuracy was found.

Lanzetta and Kleck (1970) focused primarily on the interrelationships among three phenomena: how accurately the stimulus persons' facial behavior could be judged, the galvanic skin response (GSR) indication of psychophysiological arousal during the elicitation, and the stimulus person's performance as an observer of others. We shall discuss only the results on accuracy in judging the facial behavior.[6] Stimulus persons were recorded on videotape for 12 seconds while they watched a red light signaling that they would receive shock or a green light signaling nonshock. The observers were shown the short videotape episodes and

[6] Lanzetta and Kleck did not attempt to study accuracy but rather to determine the relationship among observers' abilities to judge facial behavior, the extent to which their own facial behavior could be judged by others, and psychophysiological measures. The results on these interrelationships are quite interesting but are not reported here because the authors considered the findings tentative owing to the small number of stimulus persons. This study is very suggestive of the fact that explorations of these variables will be fruitful (see also Buck, Savin, Miller, & Caul, 1969; Jones, 1950). These studies and later ones by these and authors are discussed in chapter 8.

were required to indicate whether they were watching persons anticipating shock or nonshock. There were 12 stimulus persons; each was judged by five other persons and a self-report was obtained. As in the preceding experiment, accuracy varied with the stimulus person observed, from a low of 55% to a high of 83% correct judgments, with the median accuracy across all stimulus persons observed better than chance, 62%. There are two problems with this experiment as an accuracy study of facial behavior. The videotapes showed more than the face; the body from the waist up was seen, and the observers may have used nonfacial sources for some, or most, of their judgments. Judgment of the eliciting circumstance may or may not have required any information about emotion; perhaps coping behavior provided the basis for accuracy.

The accuracy criterion in the next two experiments was based on the expected differences in emotions between the acute and remitted phases of a psychotic disorder, confirmed by the ratings of the treatment staff. Ekman and Bressler (1964) found that stimuli randomly selected from interviews of depressive patients during the acute phase of their illness were rated as more unpleasant and more immobile than those selected from the remitted phase. In a replication (Ekman & Rose, 1965) with a larger sampling of stimuli for each person but a smaller number of persons than in the prior experiment, stimuli from the interview closest to the patient's admission to the hospital were judged as more unpleasant and immobile than those from the interview closest to the patient's discharge from the hospital. The mobility ratings were, however, a description of actual movement, not an interference about emotions.[7]

In summary, these five studies consistently show that observers can accurately judge emotions shown in spontaneous facial behavior, in the sense that their judgments agreed with the emotion expected by virtue of the nature of the eliciting circumstance. This set of studies is particularly important because it establishes that accuracy has generality across persons judged and across time. Through the use of representative sampling of stimulus persons and behaviors, these experiments refute the argument, raised in connection with the candid photograph experiments, that perhaps accuracy is possible for only the rare stimulus person or the rare moment in time.

[7] Whereas in both of these studies on depressive patients the body and the face were shown in the photographs, there is little likelihood that the observers' judgments of pleasantness could have been based on body cues rather than facial behavior because in other experiments Ekman and Friesen (1965b, 1967a) found that observers could not agree in pleasantness judgments when they were restricted to viewing just the body in still photographs.

The major limitation in these experiments is that the emotion judged was general rather than specific. Accuracy has been shown only for the distinction between positive and negative emotional states, not for any of the distinctions within those groupings, such as happiness, interest, anger, fear, and disgust. The next set of experiments to be explored provides evidence on just this point, showing that accurate judgments are possible for specific emotions. However, we shall no longer be dealing with spontaneous behavior, but with posed behavior, and must evaluate the problem of the relevance of posed to spontaneous facial behavior.

Accuracy in the judgment of posed behavior

Why consider posed facial behavior in a discussion of accuracy? One answer is that there *is* an accuracy question: can an observer viewing a pose accurately judge the emotion intended by the poser? If, however, posing is a very special or unique eliciting circumstance, then establishing accuracy in the judgment of poses has little bearing on whether spontaneous facial behavior provides accurate information (see the discussion of posing in Sections 1.2 and 2.2). This issue will be addressed after we direct our attention to the set of experiments to be discussed next.

Whereas most investigators of the judgment of emotion from facial behavior have employed posed photographs as their stimuli, only a few presented their data in a manner that allows examination of whether the observers accurately judged the emotion intended by the poser.[8] Most of these studies have methodological problems, many of which can be resolved by considering the findings across all of the experiments.

Table 4.4 shows both the methodological features and the results of the eight experiments having as their focus, at least in part, accuracy in the judgment of posed emotions. In the early experiments, the sample both of stimulus persons and of stimuli for each emotion was small; but, as the table shows, these problems were resolved in some of the later studies. Most of the studies used still photographs; three studies used live behavior, a kind of stimulus introducing potential problems (see section 2.7). However, in two studies (Levitt and Kozell & Gitter), rec-

[8] Frijda (1953) performed an accuracy study, but it will not be reported because he utilized an admittedly subjective rating of whether the observers had successfully judged the emotions intended or experienced by the two stimulus persons. With both the still photograph and motion picture film presentations, Frijda concluded that he had demonstrated accurate judgments of emotion.

Table 4.4. *Methodology and results of nine accuracy studies of posed behavior*

	Kanner (1931)	Woodworth (1938) Feleky (1914)	Dusenbury & Knower (1938)	D. F. Thompson & Meltzer (1964)	Levitt (1964)	Osgood (1966)	Drag & Shaw (1967)	Kozel & Gitter (1968)	Ekman & Friesen (1965)
Methodology									
Number of stimulus persons	1	1	2	50	50	50 or 5[a]	48	10	6
Number of stimuli for each emotion	1–3	2	2	100	50	5	10	5	2
Method of presenting stimuli	Still	Still	Still	Live	Motion picture film	Live	Live	Motion picture film	Still
Number of observers	409	100	388	4	24	110	4	44	57
Number of judgement categories other than those listed below	[b]	1	5	2	0	6	3	1	3
Percentage of accurate judgments on:									
happy	—	93	100	76	86	55	71	86	65
surprise	76	77	86	—	43	38	68	69	—
fear	75	66	93	74	58	16	62	80	35
anger	32	31	92	60	62	39	42	79	—
sad	33	70	84	52	—	19	49	59	88
disgust/contempt	66	74	91	67	45	50	41	55	0
pleasantness factor	—	—	—	—	—	0.38[c]	—	—	—
intensity factor	—	—	—	—	—	0.32[c]	—	—	—
control factor	—	—	—	—	—	0.50[c]	—	—	—

[a] See text discussion on Table 3.2.
[b] Kanner allowed free labeling; in reanalyzing his results, many of his responses, which were not obviously in one of the categories, were not verified.
[c] These are correlations between intended emotion and observed emotion when the emotion word data were reordered in terms of factor

ords of sequential behavior (motion picture film) were used, and because their results are broadly comparable to the others, the finding of accuracy is not limited to still photographs. Though earlier studies used only professional actors and only a selected sample of the presumably best photographs of them, all but one of the studies since 1938 used untrained posers, and all but one of the studies presented all poses, not just the best attempts. (Both exceptions were in the Kozel & Gitter experiment.) Thus, the findings can be said to have generality across a large number of persons and to a broader range of behavior than might be represented by a preselected photograph of the possibly rare moment when a good pose was emitted. The number of observers is adequate, except in the Thompson and Meltzer and Drag and Shaw studies, the findings of which are substantially the same as those of the other experiments.

Before considering the results shown in Table 4.4, a few words of explanation are necessary about particular experiments. (1) Both Kanner's and Woodworth's data are based on judgments of the Feleky posed photographs; and Woodworth's findings are a reanalysis of Feleky's data. (2) Drag and Shaw found significant differences in observer accuracy depending on whether the male or female posers were judged.[9] But even the most poorly judged group (men) was judged with better than chance accuracy. In the table, we combine data for male and female stimulus persons. (3) The Ekman and Friesen (1965a) study falls somewhere between posed and spontaneous behavior. Psychiatric patients were asked to show a camera how they were feeling. The patients did not simulate a specified unfelt emotion as in all of the other studies described in the table. On the other hand, their facial behavior cannot be taken as spontaneous because it was occasioned by the investigator's request. Ekman and Friesen asked the patients to describe their feelings in their own words after they had shown their facial expression. Depressive patients were asked to engage in this task upon admission to the hospital and again at discharge as it was expected that they would have different feelings at these two points. The still photographs taken at these two times were shown to observers who did not know that the pictures were of psychiatric patients. The observers utilized the emotion

[9] There have been a few more recent studies that have attempted to isolate some of the variables associated with whether an individual is a well understood or poorly understood emotion poser. The race and sex of the poser have been found to interact with the emotion posed and the race and sex of the observer (Black, 1969; Gitter & Black, 1968; Kozel, 1969), and personality measures and skin conductance have been found related to posing (Buck et al., 1969). These issues will be considered along with other studies in Chapter 8.

category system proposed by Tomkins (see Table 3.2) to record their judgments of the emotions shown in the photographs. The accuracy criterion was conformity between observer judgment and patient self-description.

The results of each experiment reported in Table 4.4 were reanalyzed in terms of the emotion categories proposed earlier (Table 3.2) to facilitate comparison across experiments. Kanner had subjectively scored his observers' free emotion labels and judgments of the situation; these data were not used; instead his published raw data were analyzed to provide the results shown in the table. Woodworth himself recast Feleky's data, which we then further modified in terms of the seven emotion categories.

Do these studies show that observers can accurately judge the emotions intended by the posers? Generally, looking across emotions and across experiments, the answer is yes.[10] All of the percentages listed are the modal, or most frequent, response to the intended emotion. In all but a few instances the poser's intended emotion was the emotion most frequently perceived by the observers. Although the results are far from perfect for any single emotion category across experiments or for any single experiment across categories, there is certainly more correspondence between intended and judged emotion across these data than might be expected by chance.

It has been customary to dismiss accurate judgments of posed behavior as by and large irrelevant to the question of whether facial behavior is systematically related to emotion and, more specifically, to the study of spontaneous behavior. As we described earlier (Section 1.2), the argument (see Hunt, 1941; Landis, 1924) has been that posed behavior is a specialized, language like set of conventions or stereotypes, which might conceivably be understood, but that, by definition, such behavior is different from what the face actually does when emotion is spontaneously aroused. As we mentioned in our discussion of eliciting circumstances (Section 2.3), the only direct study of the differences in the components of posed and spontaneous facial behavior was the dubious experiment by Landis, who failed to find any relationship between the face and emotion for either kind of eliciting circumstance.

[10] For all the experiments except those of Levitt and Osgood, there were more emotion categories in the original data than those listed in Table 4.4. Therefore our estimate of accuracy is low in that the percentages were calculated by dividing the number of correct responses by the total responses to the stimulus including unlisted categories rather than by dividing by the total number of responses that fit into the six categories used by all these authors.

Indirectly, however, there is considerable evidence that posed behavior is not a specialized, languagelike set of conventions unrelated to real emotional behavior. If it were not in some way reflective of emotion, posed behavior in one culture would not be understood by people from different cultures. Later, in exploring cross-cultural studies (Chapter 7), we shall review a large body of data from a number of experiments in which the posed facial behavior of Westerners was judged as the same emotion by members of 13 literate cultures and one preliterate culture and one experiment in which the poses by members of a preliterate culture were accurately judged by members of a literate culture. For these findings to emerge, the behaviors occurring during posing must have developed in the same way across cultures. One reasonable explanation of such development would be that they are in some way based on the repertoire of spontaneous facial behaviors associated with emotion.

We believe that when an investigator asks a poser for an emotion there is an implicit request that the person show an extreme, uncontrolled version of the emotion. When the investigator asks for a pose of anger, the subject typically will imagine and try to show extreme anger and will not attempt to deintensify, mask, or neutralize facial appearances. If the investigator were to ask him to pose an emotion by specifying a low-intensity word, such as annoyance, then the subject would attempt to show facial behavior appropriate to moderate or low-intensity emotion. It would also be possible to ask the subject to show the facial behavior that would occur if a display rule were operating; e.g., anger at a superior in a situation in which the poser could not directly manifest anger. With such an instruction, posing might well yield facial behavior that is quite similar to much spontaneous conversational behavior where display rules for the management and control of facial appearance are operative.

If we are correct in our speculation about how the subject typically interprets the posing instruction (viz., as an occasion to display an uncontrolled version of the emotion), then the obtained poses are not dissimilar from all spontaneous behavior but approximate only that spontaneous facial behavior that occurs when a person is not applying display rules to deintensify, mask, or neutralize. However, poses of extreme, uncontrolled emotion may still differ from spontaneous, unmodulated, high-intensity emotion in duration and in complexity of muscle use.

Grouping the results from a number of experiments allows the conclusion that posed facial behavior can be accurately judged, in that the majority of the observers will correctly identify the intended emotion. This result is not limited to expert posers, to the best moments in the posing situation, or to still-photographic representations of posing. The results are limited, however, to the six emotion categories considered. Conceivably, further studies might achieve accurate judgments of poses of other emotions.

Summary on the accuracy of judgments

At the outset of this section, we quoted Bruner and Tagiuri's listing of studies that produced negative and positive evidence on accurate judgments of emotion. Most of the negative studies they cited were irrelevant to the question of accuracy because those studies utilized drawings rather than real behavior, thus providing only the artist's conception of emotion as the basis for determining correct judgments. The remaining two negative studies on accuracy, those of Landis and of Sherman, were thoroughly criticized, and contradictory findings from other studies were presented to support our contention that these two experiments henceforth should be disregarded.

Contrary to the impression conveyed by previous reviews of the literature that the evidence in the field is contradictory and confusing, our reanalysis showed consistent evidence of accurate judgments of emotion from facial behavior. Without question, the evidence based on posed behavior is far stronger than that based on spontaneous behavior, where a fully adequate study remains to be done. Such a study is needed in order to show accuracy in the judgment of specific emotions in addition to the judgment of positive and negative state. There is a need for further study of accurate judgments among the different negative emotions in spontaneous facial behavior, but it seems unnecessary to continue to question whether accurate judgments are possible. More useful research would determine under what conditions, for what kinds of people, in what kinds of roles and social settings, and with what types of accuracy criteria facial behavior provides correct information about emotion; and, conversely, research would also determine in what kinds of settings and roles and for what kinds of people facial behavior provides either no information or misinformation.

As we mentioned at the beginning of this chapter, there are two research approaches to the question of whether the face can provide accurate information about emotion. Thus far we have considered in this chapter only studies that use the judgment approach, determining whether observers can make accurate inferences about emotion from viewing facial behavior. The success of such studies makes the other approach, the measurement of facial components, very important because the judgment studies can only tell us that the information is there, somewhere in the face, and capable of being interpreted accurately by observers. The judgment approach cannot tell us what facial behaviors are providing this accurate information, what particular muscular movements or wrinkles in the face allow the observer to determine that an individual is in a stressful rather than a cathartic part of an interview, or how to distinguish when an individual is posing anger rather than disgust. To resolve these problems and to specify just which facial behaviors are distinctively related to which emotions, we must consider the second approach to the study of accuracy and measurement of facial components and this will be the subject of the next section.

4.2. Can measurement of facial behavior provide accurate information?

In component studies, facial behavior is the dependent variable, or response measure, rather than the independent variable, or stimulus, as it is in judgment studies. We do not attempt to determine what observers can say about faces but what the measurement of facial components can indicate about some aspect of a person's experience. In a component study, we might ask, for example, "What components of facial behavior differentiate among faces sampled when the subject was afraid and those sampled when the subject was disgusted?" In a judgment study, we might ask, "Can observers tell when looking at a face whether the subject was afraid or disgusted?" (the difference between component studies and judgment studies was reviewed earlier; see Section 2.1).

There have been remarkably few component studies. The scarcity of research is not due to the difficulty in establishing independent variables, that is, eliciting circumstances in which to sample facial behavior with some criterion of how the person feels, because these difficulties are also encountered with judgment studies, of which there have been many. It is probably due to the difficulty in deciding what to measure in the face. Today there is still no accepted notion of the units of facial behavior nor any general procedure for measuring or scoring facial components. (This may soon be changing; see Chapter 9.) Investigators have improvised their own techniques, rarely using techniques tried by others and almost invariably combining the facial component units into a few global scores. Progress is being made, and three measurement procedures have been developed, and there is some evidence to support the validity of one of them.

Tables 4.5 and 4.6 summarize the methodological features and findings from seven studies.[11] All obtained positive results, but each has shortcomings in either interpretation or generalization of the results. In the Landis and Hunt (1939) study, strong evidence of a specific facial response to a startling stimulus was detected. However, the facial responses documented for reactions to a sudden noise (a pistol shot) do not resemble the stimuli that observers customarily judged as showing surprise. The startle facial reaction was extremely brief, followed by a secondary reaction, presumably an emotion about the initial startle, which varied across subjects; Landis and Hunt did not determine whether this secondary reaction had systematic properties related to the subjects' reported feelings or to manipulations in the setting, which might have caused the sudden noise to be associated with fear, interest, or anger.

Trujillo and Warthin's (1968) finding that ulcer patients have more vertical creases in their brow when asked to frown than have other medical patients may or may not be relevant to emotion. They cite Darwin's (1872) and Bell's (1847) notion that the permanent creases in the face result from the most frequently experienced emotions and, on that basis, suggest their findings have relevance to emotion. However, they acknowledge that they did not control for chronic pain, which is not considered to be an emotion by most authors, although there is evidence (Boucher, 1969) that it does have some distinctive facial components. Their findings, even if relevant to emotion, are too general to be useful, because vertical creases in the brow can be found with anger, fear, or sadness, and they did not examine other facial components that might further distinguish among these emotions.

Leventhal and Sharp's (1965) findings are open to similar questions about whether facial components of pain or of some specific emotion during childbirth are responsible for their results. They use Tomkins's term *distress* to describe the emotion they studied. Earlier (in Section 3.1),

[11] Landis also conducted a component study on the same materials he showed to observers and failed to find any components related to his eliciting circumstances or the subjects' self-reports. His results will not be discussed, both for the reasons already outlined (Chapter 4), which raise serious doubts about his study, and because of two additional problems relevant to his components analysis. In computing the participation of the various muscles, he included pictures taken before and after the actual eliciting stimuli; to what extent the ubiquitous nervous smile he refers to is a function of including one "expectancy" situation (the before-elicitation pictures) with every situation cannot be gauged. Also, in analyzing his data, Landis used a technique that was both extremely conservative (Frois-Wittmann) and inappropriate for the problem (Davis).

	Landis & Hunt (1939)	Trujillo & Warthin (1968)	J. Thompson (1941)	Fulcher (1942)
Methodology				
Eliciting circumstances	Experimental presentation of sudden, intense stimuli (primarily .22 pistol shot)	Chronic duodenal ulcer	Naturally occurring activities; a few stimuli introduced; emotion inferred from context	Instruction to pose different emotions
Emotions sampled	Startle	Not specified as emotion	Laughter, smiling, crying (also isolated inferences of fear and anger)	Anger, fear, happiness, sadness
Number and type of subjects	Normals; also infants, animals, psychotics, epileptics, deaf persons, patients with neurological disorders, subject injected with adrenalin, hypnotized subjects	126 ulcer paitents, 274 patients with other medical disorders	29 Seeing children, 7 weeks–13 years; 26 blind children 7 weeks–13 years	118 Seeing children, 4–16 years; 50 blind children, 6–21 years
Type of facial components measured	Eyeblink, widening of mouth, forward movement of head	Number of vertical folds between eyebrows	Three-point scale: some/much/no involvement of	Six-point scale: amount of movement; six-point scale: amount of distortion (eye and mouth separately); Yes–no judgement of involvement of 18 muscles; six-point "adequacy" scale.
Results	Strength of response varies directly with intensity and suddenness of stimulus. Some facial response (minimally eyeblink) *always* elicited by stimulus of sufficient strength except in epileptics. Primary stimulus of sufficient pattern shows very little variation; secondary responses vary across subjects	Three or more vertical folds in 85% of ulcer patients as compared with only 6% of control patients when asked to frown	Pattern of muscular activity same in blind and seeing for each type of emotional behavior; seeing subjects show more uniformity of pattern	More facial activity in seeing subjects; blind subjects show same general patterns but less differentiation among emotions

Table 4.6. *Methodology and results of three component studies of facial behavior*

	Leventahl & Sharp (1965)	Rubenstein (1969)	Ekman & Friesen (1972)
Methodology			
Eliciting circumstances	Prechildbirth labor, total time in labor divided into four intervals	Depressed patients asked to smile before shock treatment and 1 hour after treatment; control group of nonpatients tested twice	Each subject watched a neutral travelogue and a stress-inducing film of sinus surgery
Emotions sampled	Distress	Happiness	The self-report of the subjects showed that the emotion experienced in the neutral film was slight happiness and in stress film was interest, surprise, fear, pain, disgust, and sadness
Number and type of subjects	52 Women with prior childbirth experience; 19 women with no prior childbirth experience (55 subjects returned Welch anxiety scale and were divided into high/low anxious by median split).	17 depressive patients; 16 control subjects	25 college students
Type of facial components measured	Forehead: 4 behaviors, 2 index[a] Brow: 8 behaviors, 3 indexes Eyelids: 6 behaviors, 2 indexes Nose: 5 behaviors Eyes: 12 behaviors, 2 indexes Mouth: 36 behaviors, 3 indexes Score: frequency of behavior during 5-minute observation interval	Amount of development of facial muscles derived from obtaining a series of profile shots taken rapidly on motion picture film within a facial expression	Facial Affect Scoring Technique measured the presence of fear, anger, surprise, disgust, sadness, happiness: Brow: 8 behaviors Eyes: 17 behaviors Lower face: 45 behaviors
Results	Forehead, brow, eyelid indexes show increased discomfort (wrinkles, movement) as labor progresses; other facial indicators insignificant	More displacement of facial muscles during smile following shock treatment than before; no change from pre- to post treatment in the control subjects	More surprise, sadness, disgust, anger in stress than neutral; more happiness in neutral than stress stituation

[a] Indexes of comfort, discomfort (major, minor, or unspecified), change, created by grouping individual behavior measures.

we pointed out that the term is problematic in that it can refer either to sadness–grief or to pain–hurt–suffering. Their discomfort indexes, built from scores on the eyebrows, forehead, and eyelids, may well have measured either pain or sadness or both. Their study is noteworthy, however, in that their facial behavior measures were related not only to the severity of labor but to the number of previous births and anxiety.

Both Fulcher (1942) and J. Thompson (1941) analyzed their results primarily by comparing blind and sighted children. They reported their data in a way that makes it difficult to determine what the precise differences in the facial components were for each emotion they studied within either sample of children, yet they both reported more extensive lists than most other investigators of facial components, which they hypothesized as distinctive for each emotion. Thompson's results on smiling, laughing, and crying showed similarities in the distinctive movements of the facial components for each of these reactions for her blind and sighted subjects. Less information is provided about anger and sadness, although she said they also had distinctive facial components in both blind and sighted children. Fulcher's study of the posed emotions of blind and sighted children provides information on a wider sampling of emotions and with more information about the distinctive components for each emotion, but not in sufficient detail to check his hypotheses about whether the components of facial behavior are distinctive for each emotion posed. His findings do suggest that facial components unique to each posed emotion could be isolated and measured. For new studies on this question utilizing posing procedures with sighted adults or children see Section 8.2.

Rubenstein's (1969) procedure for measuring facial components is novel but quite cumbersome. A 16-mm motion picture camera was rotated around the subject's face rapidly, acquiring a series of profile frames during a facial expression. His method of recording requires, however, that the subject freeze an expression for at least 5 seconds while the camera travels around the face and that the subject be in a rather immobilized position. This procedure is not only questionable in terms of its applicability to spontaneously occurring facial behavior, but the subjects are constantly made aware that their facial behavior is of interest. His finding, that depressive patients smile more broadly when asked to do so after shock treatment than before, does demonstrate that his measurement procedure works, but it adds little information about the facial components.

The most elaborate and complex study is the experiment by Ekman and Friesen (Ekman, 1972, 1973; Friesen, 1972) listed in Table 4.6. We shall report this experiment in some detail because of its complexity, the import of the findings, and the relevance of the methods and results to our discussion in two later sections, 5.2 and 7.2.

In conjunction with Averill, Opton, and Lazarus, Ekman and his associates collected records of the facial behavior, skin resistance, heart rate, and self-reported emotion of subjects in the United States and Japan, as they watched a neutral and a stress film. We shall discuss here only their findings on the facial behavior of the American subjects; in Chapter 7 we shall discuss the cross-cultural comparison with the Japanese subjects.

In this study a new tool for measuring facial behavior was utilized, viz., Ekman, Friesen, and Tomkins's Facial Affect Scoring Technique (FAST). The derivation of this technique and the details of its use are reported elsewhere (Ekman, Friesen, & Tomkins, 1971), but it will be necessary to provide some information about how the scoring procedure was used in order to explain the findings and convey something about the comprehensiveness of this measurement system. We shall first describe the use of FAST and then the experiment in which it was used, explaining the results listed in Table 4.6.

The Facial Affect Scoring Technique requires scoring of each observable movement in each of three areas of the face: (1) brows/forehead area; (2) eyes/lids; (3) lower face, including cheeks, nose, mouth, and chin. Rather than defining each scoring category in words, FAST employs photographic examples to define each of the movements within each area of the face that, theoretically, distinguish among six emotions: happiness, sadness, surprise, fear, anger, and disgust. For example, instead of describing a movement as "the action of the frontalis muscle which leads to raising of both brows in a somewhat curved shape, with horizontal wrinkles across the forehead," FAST utilizes a picture of just that area of the face in that particular position to define that scoring item. Figure 4.1 shows as an example the items across the facial areas considered to be relevant to surprise.

The FAST system is applied by having independent coders view each of the three areas of the face separately, with the rest of the face blocked from view. It should be emphasized that the FAST measurement procedure does not entail having the coders judge the emotion shown in the face they are coding. Rather, each movement within a facial area is distinguished, its exact duration determined with the aid of slowed motion, and the type of movement classified by comparing the movement observed with the atlas of FAST criterion photographs. If, for example, the coder is looking at the brows/forehead and sees a particular movement in that area of the face, he compares the movement with the eight photographs of brow/forehead movements in the FAST atlas and assigns to it the FAST atlas number of the criterion picture it most closely resembles. In addition to those for the brows, there are 17 criterion photographs of eyes/lids, and 45 criterion photographs of the lower face in the FAST atlas.

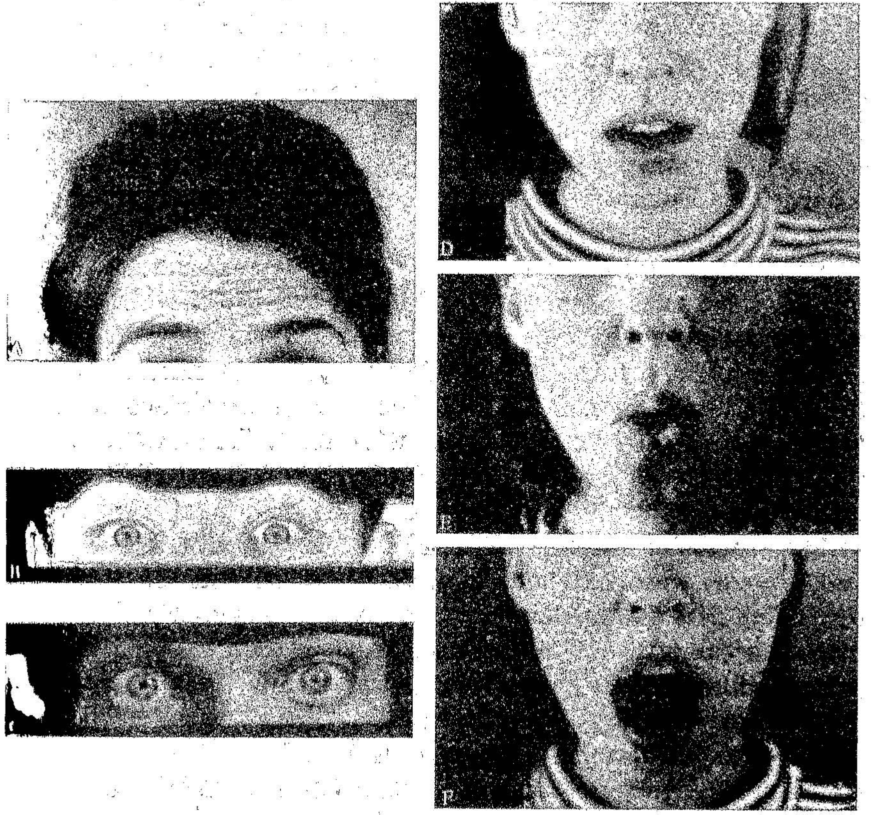

Figure 4.1 Examples of criterion items from the Facial Affect Scoring Technique (FAST) showing the brow/forehead (A), the eyes/lids (B and C), and the lower face items (D-F) for surprise.

Once the coders' scoring is complete, formulas are used to derive the emotion prediction for each facial movement, taking into account the scoring of more than one independent coder. For example, if the facial movement is coded by more than one coder as most closely resembling the FAST brow/forehead picture B9 (shown in Fig. 4.1), then that movement is labeled surprise. The output of the scoring system is a series of duration scores for anger, fear, surprise, sadness, disgust, and happiness for the brows/forehead, the eyes/lids, and the lower face.

Data analysis can be performed by measuring either the *frequency* of occurrence of each emotion within each facial area or the *duration* of each emotion within each facial area. The frequence or duration scores can be analyzed separately for each of the three facial areas or emotion scores for the total face can be obtained by utilizing another formula, which combines the scores for emotions shown across the face into a total face score for a single emotion or a blend of emotions. In the results reported in Table 4.6, total face scores were calculated only for a movement occurring at least two of the three facial areas and only for single emotions.

With a scoring system such as FAST, a system intended to measure facial behavior and which distinguishes among six emotions, the question must be raised as to whether or not it is valid. There are two types of validity, which we may call *personal* validity and *social* validity. In the next chapter we shall present the results of a study of FAST's social validity – whether measures of facial behavior can predict how people will judge the emotion shown in a face. In this chapter we have been discussing personal validity – whether measures of facial behavior can provide accurate information about the person, that is, about some aspect of his emotional experience or circumstance.

Let us turn now to the question asked in the Ekman and Friesen, experiment: Can the measurement of facial behavior accurately distinguish whether subjects watched a stressful or a neutral motion picture film?

In the data pool collected jointly by Ekman and his associates and by Averill, Opton, and Lazarus, each subject had been seated alone, watched first a film of autumn leaves and then a 3-minute stress-inducing film of sinus surgery. Unknown to the subjects, a videotape record was made of their facial behavior. Subsequently, the subjects answered a questionnaire about their emotional experience during the stress film. The FAST scoring system was applied to every observable movement in each of the three areas of the face for the 25 American subjects; approximately 3 minutes of their facial behavior during the neutral film and 3 minutes during the stress film were scored.

The results reveal an enormous difference in the facial behavior shown in these two eliciting circumstances. The total face scores, the scores for each of the separate facial areas, the scores based on frequency, and the scores based on duration all indicate that there was more behavior that FAST measured as surprise, sadness, disgust, and anger shown during the stress film and more behavior that FAST measured as happiness shown during the neutral film.

This study shows that measurement of facial behavior accurately discriminates between two eliciting circumstances, watching a neutral and a stress film. It is important to note that this difference between facial behavior shown in two different eliciting circumstances was obtained with a measurement system designed to measure six different emotions rather than being limited to the occurrence of one or two emotions or to the distinction between positive and negative feelings.

The experiment was not designed to provide evidence that FAST can accurately indicate each of the six emotions it was designed to measure. The only accuracy criterion available is the two film conditions, stress and neutral. Although self-reports were gathered, they are a poor accuracy criterion in this experiment because, although the stress film appears to have elicited different emotions in each subject, the self-report did not provide any information about the sequence of emotions experienced and the self-report data were gathered some time after the experience. What is required is a self-report on the felt emotion obtained immediately after a particular facial behavior occurs.

There are two other sources of information that imply that FAST does succeed in accurately differentiating particular emotions. The first, which we shall discuss in Chapter 7, is the high correlations between the specific emotions shown by American and Japanese subjects as measured by FAST. Although it cannot be said from those findings that FAST accurately measured each emotion, it can be said that FAST differentiated types of facial behavior and that these different types of facial behavior occurred with the same frequency in subjects from two different cultures who were placed in the same eliciting circumstances. For example, even though we cannot conclude that there is evidence that the FAST scores for disgust do actually measure disgust and those for surprise do actually measure surprise, it is encouraging that the FAST measurements show the same ratio of disgust to surprise behavior across members of two cultures subjected to the same eliciting circumstance.

The second piece of evidence that suggests that FAST can accurately measure specific emotions comes from a study to be outlined in the next chapter on the social validity of FAST. In that experiment, FAST scores accurately forecasted the specific emotions judged by those who simply observed the face.

One reason why we have described Ekman and Friesen's experiment at some length is because of the importance, in our view, of research that directly measures facial components. There has been too little of such research. Although judgment studies in which observers tell us their impressions about a face can be quite informative, they cannot provide knowledge about the specific facial behaviors relative to specific emotions, and many of the questions that need to be answered about the face and emotion cannot be approached solely through the use of judges. Most investigators have avoided direct measurement of facial components, and the few who did measure facial behavior, discussed earlier in this chapter, did not offer a general tool for measuring the occurrence of a number of different emotions. Ekman, Friesen, and Tomkins's FAST is intended as a general-purpose tool to measure the occurrence of six different emotions and blends of these six. Chapter 9 reports on a new technique for directly measuring facial behavior, which has been developed by Ekman and Friesen.

Two other scoring systems for measuring facial behavior have been developed by investigators following an ethological approach, Blurton-Jones (1969) and Grant (1969). Neither has yet performed any validity studies. All three systems, FAST and those developed by Blurton-Jones and Grant, have considerable overlap, although they differ in a number of regards. The FAST system is based on theory, attempting to specify only those facial behaviors that can distinguish one emotion from another, and the other two systems have attempted inductively to derive a descriptive system to cover all facial behavior observed in their samples of adults or children. The scoring items are depicted in terms of a photographic atlas in FAST, whereas the other two systems utilize a verbal description of particular muscular movements and wrinkles. The appearance of these three scoring systems is an exciting development, offering investigators a choice where there previously was none for measuring the face.

Summary on accuracy of measurements of facial behavior

The few studies on components of facial behavior are encouraging, suggesting that accurate information about some aspect of a person's experience (whether it be response to a gunshot, to childbirth, to a stress-inducing film) can be derived from measures of facial components; but much more work is needed to supply a definitive answer as to whether measurements of the face can provide accurate information about specific emotions. The evidence to date is limited to showing that accurate information about the distinction between positive and negative emotional reactions can be obtained from measurements of the face. (Accuracy is considered again in the review of studies from 1970 to 1977 in Chapter 8, Section 8.4. Unfortunately, not much progress has been made to answer the questions raised here.)

We believe this is one of the most crucial areas for further research and that the ability to measure the face directly, rather than solely relying on observers' global judgments, will be the key to a breakthrough in the next generation of questions about the face and emotion. (Chapter 9 critically reviews a number of new facial measurement procedures. See also, Chapter 8, Section 8.3.)

9. Measuring facial movement with the Facial Action Coding System

PAUL EKMAN AND WALLACE V. FRIESEN

Most researchers on facial behavior have not measured the face itself but instead have measured the information that observers were able to infer from the face. Examples of the questions asked are: Can observers make accurate inferences about emotion? Can observers detect clinical change or diagnosis? Do observers from different cultures interpret facial expression differently? Are observers influenced by contextual knowledge in their judgments of the face? Do observers attend more to the face than to the voice?

Few studies have involved measurement of the face itself. Examples of the type of questions that could be asked are: Which movements signal emotion? Do facial actions change with clinical improvement or differentiate among types of psychopathology? Do the same facial movements occur in the same social contexts in different cultures? Are certain facial actions inhibited in certain social settings? Which facial movements punctuate conversation? The differences between these two approaches to the study of facial behavior (i.e., observers' inferences versus facial measurement) were discussed in Chapter 1.

Research focused on the face has been impeded by the problems of devising an adequate technique for measuring the face. Over the years, various procedures for facial measurement have been invented. Early work (e.g., Frois-Wittmann, 1930; Fulcher, 1942; Landis, 1924; J. Thompson, 1941) is rarely cited by current investigators. Rather, current approaches to facial measurement have varied in methodology, ranging from analogic notations of specific changes within a part of the face (Birdwhistell, 1970) to photographic depictions of movements within each of three facial areas (Ekman, Friesen, & Tomkins, 1971) to verbal descriptions of facial gestalten (Young & Decarie, 1977).

No consensus has emerged about how to measure facial behavior. No tool has been developed as a standard to be used by all investigators.

Investigators have almost been in the position of inventing their own tools from scratch. The only exception has been that the category lists of facial behavior described by some human ethologists (Blurton Jones, 1971; Grant, 1969; McGrew, 1972) have influenced other human ethologists studying children.

Although differing in almost all other respects, most facial measurement techniques have shared a focus on what is visible, that is, on what raters can differentiate when they see a facial movement. An exception (Schwartz, Fair, Salt, Mandel, & Klerman, 1976) used electromyographic (EMG) measurement to study changes in muscle tone not involving a noticeable movement. EMG could also be used to measure visible changes in muscle tone not involving a noticeable movement. EMG could also be used to measure visible changes in muscle tone that do not involve a noticeable movement, but such work has not been done. Although EMG could also be employed to study visible movement, we think it is unlikely that surface electrodes could distinguish the variety of visible movements delineated by most other methods. Later in this chapter we shall describe a study comparing EMG and visible movement measurement.

Vascular changes in the face are another aspect of facial behavior that can occur without visible movement and, like muscle tonus, could be measured directly with sensors. No such work has been published in coloration or skin temperature, although Schwartz, in unpublished studies, has found thermal measures useful in measuring affective responses. Some of the measurement procedures that utilize observers to rate visible movement, have included a reference to a "reddened" face.

Elsewhere (Ekman, 1982) a comparison was made between 13 other methods for measuring facial movement and the FACS method, contrasting the assumptions that underlie each method, explaining how units of measurement were derived, and providing point by point comparisons of the measurement units. Here we shall only selectively contrast other methods with FACS to explain the technique.

9.1. Background to the development of the Facial Action Coding System

Our primary goal in developing the Facial Action Coding System was to develop a *comprehensive* system that could distinguish all possible visually distinguishable facial movements. Most other investigators developed their method just to describe the particular sample of behavior they were studying. Our earlier approach, the Facial Affect Scoring Technique (FAST) (Ekman, Friesen, & Tomkins, 1971) discussed in Chapters

4 and 5, also had this narrower objective. It was designed primarily to measure facial movement relevant to emotion. Although we remained interested in describing emotion signals, to do so we needed a measurement scheme that could distinguish among *all* visible facial behavior. We were also interested in a tool that would allow study of facial movement in research unrelated to emotion; e.g., facial punctuators in conversation or facial deficits indicative of brain lesions. With comprehensiveness as our goal, we wanted to build the system free of any theoretical bias about the possible meaning of facial behaviors.

The interest in comprehensiveness also led us to reject an inductive approach to developing FACS. Most other investigators devised their descriptive system on the basis of careful inspection of some sample of the behavior they intended to measure. Thus, although their system might contain gaps, as long as its purpose was simply to measure a prescribed sample of events, it was perfectly practicable. With comprehensiveness as a goal, an inductive method would require inspecting a very large and diversified sample of behavior.

We chose to derive FACS from an analysis of the anatomical basis of facial movement. Because every facial movement is the result of muscular action, we concluded that a comprehensive system could be obtained by discovering how each muscle of the face acts to change visible appearance. With that knowledge, it would be possible to analyze any facial movement into anatomically based minimal action units.

No other investigator has so exclusively focused on the anatomy of facial movement as the basis for the descriptive measurement system. Blurton Jones (1971) considered anatomy in developing his descriptive categories, but it was not the main basis of his measurement system. He did not attempt to provide a description of the full range of minimal actions.

Our interest in comprehensiveness was motivated not only by the diverse applications we had in mind but by an awareness of the growing need for a common nomenclature for this field of research. Comparisons of the measurement units employed by other investigators would be facilitated if the particular units used in each study could be keyed to a single comprehensible list of facial actions. Also, a complete list of facial actions would reveal to potential investigators the array of possibilities, so they could better select among them. And, of course, there might be some investigators who, like us, would want to measure, not just some facial behavior, but all possible movement that they could observe.

A constraint on the development of FACS was that it deals with what is clearly *visible* in the face, ignoring invisible changes (e.g., certain changes in muscle tonus) and discarding visible changes too subtle for reliable distinction. In part, this constraint of measuring the visible was willingly adopted, based on our interest in behavior that could have social consequences. In part, the constraint of dealing only with the visible was based on our interest in a method that could be applied to any record of behavior – photographic, film, or video – taken by anyone. If our descriptive system included the nonvisible, we would be limited only to situations where we ourselves could attach the apparatus (e.g., the leads for EMG). The visibility constraint was also dictated by our belief that if subjects know their face is being scrutinized, their behavior may differ radically. The odd results obtained by Landis (1924) may have been in part because of this (see Chapter 4 for discussion of the Landis studies). A method based on visible behavior would use video or motion picture film records, which could be gathered without the subject's knowledge.

Another limitation placed on the system was that FACS would deal with *movement*, not with other visible facial phenomena. These other facial signs are important to a full understanding of the psychology of facial behavior, but their study requires a different methodology. Elsewhere (Ekman, 1977) a variety of static and slow facial signs have been distinguished, contrasting the types of information they may contain with rapid facial movement. With FACS, visible changes in muscle tonus that do not entail movement are excluded. These changes can be measured through EMG or by having observers make global inferences about brightness, alertness, soberness, etc. Changes in skin coloration are not usually visible on black and white records. Facial sweating, tears, rashes, pimples, and permanent facial characteristics were also excluded from FACS. As the name states, the Facial Action Coding System was developed to measure only movement of the face.

Ideally, the Facial Action Coding System would differentiate every change in muscular action. Instead, it is limited to what humans can reliably distinguish because it is used by human operators viewing facial behavior, not a machine-based classification. The system includes most, but not all, of the subtle differences in appearance that result from muscle action. The fineness of the scoring categories in FACS depends on what can be reliably distinguished when a facial movement is inspected repeatedly in stopped and slowed motion.

A system for measuring visible facial movements can follow one of two approaches. Either the minimal units of behavior can be specified, which can, in combination, account for any total behavior, or a list of possible facial gestalten can be given. There are several reasons for selecting the minimal units approach. First, the sheer variety of possible actions the facial musculature allows argues for the minimal units solution rather than gestalten if comprehensiveness is the goal. Also, there are too many different possible total facial actions to list all of the gestalten. Third, if the method specifies facial gestalten (e.g., Young and Decarie's, 1977, the list of 42 facial gestalten), it cannot score facial actions that show only part of the gestalt or actions that combine some of the elements of several gestalten.

Although most investigators listed minimal units, they were not explicit as to how they derived their list. How did they determine whether an action was minimal or, instead, a composite of two actions that might appear separately? Usually the decision was based on a hunch, speculation about signal value, or simply what was observed in a limited sample of facial behavior. Because we decided that an answer would come from knowledge of the mechanics of facial action, we set about determining the number of muscles that can fire independently, and whether each independent muscular action results in a distinguishable facial appearance. Such an anatomically-based list of facial appearances should allow description and differentiation of the total repertoire of visibly different facial actions.

Some might argue that there is no need to make such fine distinctions among facial actions. Indeed, there might not be a need; many differently appearing facial actions may serve the same function, or convey the same message. There may be facial synonyms, but that should be established empirically, not on *a priori* grounds. Only a measurement scheme that separately scores visibly different facial actions will permit the research that can determine which facial actions should be considered equivalent in a particular situation.

Another consideration that guided our development of the Facial Action Coding System was the need to separate inference from description. We are interested in determining which facial behavior is playful, or puzzled, or sad, but such inferences about underlying state, antecedent, or consequent actions should rest on evidence. The measurement must be made in noninferential terms that describe the facial behavior, so that the inferences can be tested by evidence. Almost all of the previous descriptive systems have combined inference-free descriptions with descriptions confounded with inference; e.g., "aggressive frown" (Grant, 1969); "lower-lip pout" (Blurton Jones, 1971); "smile tight-loose" (Birdwhistell, 1970). Each of these actions could be described without inferential terms. Because humans do the measurement the possibility of inferences cannot be eliminated, but they need not be encouraged or required. If a face is scored, for example, in terms of the lip corners moving up in an oblique direction that raises the infraorbital triangle, the person scoring the face still may make the inference that what he is describing is a smile. Our experience has been that when people use a solely descriptive measurement system, as time passes they increasingly focus on the behavioral discriminations and are rarely aware of the "meaning" of the behavior.

Another problem that plagued previous attempts to measure facial movement was how to describe most precisely each measurement unit. Blurton Jones (1971) noted that facial activity could be described in three ways: the location of shadows and lines; the muscles responsible; or the main positions of landmarks such as mouth corners or brow location. He opted for the last, although he said the other two were used also. He decided not to base his descriptions on muscular activity because it would be "... more convenient if description could be given which did not require that anyone who uses them should learn the facial musculature first, although knowledge of the musculature obviously improves the acuity of one's observations" (p. 369).

We have taken almost the opposite position. The user of FACS must learn the mechanics – the muscular basis – of facial movement, not just the consequences of movement or a description of a static landmark. It is by emphasizing patterns of movement, the changing nature of facial appearance, that distinctive actions are described – the movements of the skin, the temporary changes in shape and location of the features, and the gathering, pouching, bulging, and wrinkling of the skin.

It is FACS's emphasis on movement and the muscular basis of appearance change that helps overcome the problems caused by physiognomic differences. Individuals differ in the size, shape, and location of their features and in the wrinkles, bulges, or pouches that become permanent in midlife. The particular shape of a landmark may vary from one person to another; e.g., when the lip corner goes up, the angle, shape, or wrinkle pattern may not be the same for all people. If only the end result of movement is described, scoring may be confused by physiognomic variations. Knowledge of the muscular basis of action and emphasis on recognizing movements helps to deal with variations caused by physiognomic differences.

9.2. Development of the Facial Action Coding System

Our first step in developing FACS was to study various anatomical texts to discover the minimal units. We expected to find a listing of the muscles that can fire separately and how each muscle changes facial appearance. We were disappointed to find that most anatomists were seldom concerned with facial appearance. The anatomy texts for the most part described the location of the muscles. Capacity for separate action or visible change in appearance was not the basis for the anatomists' designation of facial muscles. Instead, they distinguished muscles because of different locations, or if there was a similar location they separately named what appeared as separate bundles of muscle fibers.[1]

Duchenne (1862) was one of the first anatomists concerned with the question of how muscles change the appearance of the face. He electrically stimulated the facial muscles of a man without pain sensation and photographed the appearance changes. By this means he was able to learn the function of some of the muscles. His method was problematic for exploring the action of all of the facial muscles. Many of the muscles of the face lie one over the other, and surface stimulation will fire a number of muscles. Inserting a needle or fine wire through the skin to reach a particular muscle might fire others as well.

The work of Hjortsjö (1970) proved to be the most help. An anatomist interested in describing the visible appearance changes for each muscle, Hjortsjö learned to fire his own facial muscles voluntarily. He photographed his own face and described in drawings and words the appearance changes for each muscle. His aim was not to provide a measurement system, and so he did not consider many of the combinations of facial muscles, nor did he provide a set of rules necessary for distinguishing among appearance changes that are in any way similar.

Following Hjortsjö's lead, we spent the better part of a year with a mirror, anatomy texts, and cameras. We learned to fire separately the muscles in our own faces. When we were confident that we were firing the intended muscles, we photographed our faces. Usually there was little doubt that we were firing the intended muscle. The problem instead was to learn how to do it at all. By feeling the surface of our faces, we could usually determine whether the intended muscle was contracting. By checking Hjortsjö's account, we could see whether the appearance on our faces was what he described and showed in his drawings. There were a few areas of ambiguity for which we returned to a variation on Duchenne's method for resolution. A neuroanatomist placed a needle in one of our faces, inserting the needle into the muscle about which we were uncertain. With the needle in place, the muscle was voluntarily fired, and electrical activity from that needle placement verified that indeed it was the intended muscle. As this method was uncomfortable, we used it rarely and only when we were in doubt.

One limitation of this method of deriving facial units must be noted. If there are muscles that cannot be fired voluntarily, we cannot study them. This seems to be the case only with the *tarsalis* muscle, and as best we can determine, its action and effect on appearance are not different from those of one of the voluntarily controlled muscles, *levator palpebrae*.

Our next step was to examine the photographs taken of each of our faces, scrambling the pictures so that we would not know what muscle had been fired. Our purpose was to determine if all the separate muscle actions could be distinguished accurately from appearance alone. Often it was easy to determine, although it usually required comparing the appearance change with the resting, or baseline, facial countenance.

There were instances, however, in which we found it difficult to distinguish among the many muscles in a set to account for a photograph of a facial appearance. Sometimes we could tell one muscular action from another, but the differentiation seemed so difficult that we prejudged it as not likely to be reliable. Sometimes the appearance changes resulting from two muscles seemed to differ mostly in intensity of the action, not in type of appearance. In either instance, we designated and described the result as one *action unit*, which could be produced by two or three different muscles.

Note that we call the measurements *action* not muscle units. As just explained, this is because a few times we have combined more than one muscle in our unitization of appearance changes. The other reason for using the term *Action Unit* is because we have separated more than one action from what most anatomists described as one muscle. For example, following Hjortsjö's lead, the *frontalis* muscle, which raises the brow, was separated into two Action Units, depending on whether the inner or outer portion of this muscle lifts the inner or outer portions of the eyebrow.

Table 9.1 lists the numbers, names, and anatomical bases of 33 Action Units, most of which involve a single muscle. The numbers are arbitrary and do not have any significance except that 1 through 7 refer to brows,

[1] We are grateful to Washburn (1975, Pers. comm.) for explaining why the standard anatomy texts were of so little help and for encouraging our attempt to explicate the muscular basis of facial action.

Table 9.1. *Single Action Units (AU)*

AU number	FACS name	Muscular basis
1	Inner brow raiser	*Frontalis, pars medialis*
2	Outer brow raiser	*Frontalis, pars lateralis*
4	Brow lowerer	*Depressor glabellae; depressor supercilii*; corrugator
5	Upper lid raiser	*Levator palpebrae superioris*
6	Cheek raiser	*Orbicularis oculi, pars orbitalis*
7	Lid tightener	*Orbicularis oculi, pars palpebralis*
8	Lips toward each other	*Orbicularis oris*
9	Nose wrinkler	*Levator labii superioris, alaeque nasi*
10	Upper lip raiser	*Levator labii superioris, caput infraorbitalis*
11	Nasolabial furrow deepener	*Zygomatic minor*
12	Lip corner puller	*Zygomatic major*
13	Cheek puffer	*Caninus*
14	Dimpler	*Buccinator*
15	Lip corner depressor	*Triangularis*
16	Lower lip depressor	*Depressor labii inferioris*
17	Chin raiser	*Mentalis*
18	Lip puckerer	*Incisivii labii superioris; incisivus labii inferioris*
20	Lip stretcher	*Risorious*
22	Lip funneler	*Orbicularis oris*
23	Lip tightener	*Orbicularis oris*
24	Lip pressor	*Orbicularis oris*
25	Lips part	*Depressor labii*, or relaxation of *mentalis* or *orbicularis oris*
26	Jaw drops	*Masseter*; temporal and internal *pterygoid* relaxed
27	Mouth stretches	*pterygoids*; digastric
28	Lips suck	*Orbicularis oris*
38	Nostril dilator	*Nasalis, pars alaris*
39	Nostril compressor	*Nasalis, pars transversa and depressor septi alae nasi*
41	Lids droop	Relaxation of *levator palpebrae superioris*
42	Eyes slit	*Orbicularis oculi*
43	Eyes close	Relaxation of *Levator palpebrae superioris*
44	Squint	*Orbicularis oculi, pars palpebralis*
45	Blink	Relaxation of *levator palpebrae* and contraction of *orbicularis oculi*, pars palpebralis
46	Wink	*orbicularis oculi*

Table 9.2. *An example of information given in FACS for each Action Unit*

Action Unit 15 – Lip corner depressor

The muscle underlying AU 15 emerges from the side of the chin and runs upward attaching to a point near the corner of the lip. In AU 15 the corners of the lips are pulled down. Study the anatomical drawings that show the location of the muscle underlying this AU.

(1) Pulls the corners of the lips down.
(2) Changes the shape of the lips so they are angled down at the corner, and usually somewhat stretched horizontally.
(3) Produces some pouching, bagging, or wrinkling of skin below the lips corners, which may not be apparent unless the action is strong.
(4) May flatten or cause bulges to appear on the chin boss, may produce depression medially under the lower lip.
(5) If the *nasolabial furrow*[a] is permanently etched, it will deepen and may appear pulled down or lengthened.

The photographs in FACS show both slight and strong versions of this Action Unit. Note that appearance change (3) is most apparent in the stronger versions. The photograph of 6 + 15 shows how the appearance changes due to 6 can add to those of 15. Study the film of AU 15.

How to do 15

Pull your lip corners downward. Be careful not to raise your lower lip at the same time – do not use AU 17. If you are unable to do this, place your fingers above the lip corners and push downward, noting the changes in appearance. Now, try to hold this appearance when you take your fingers away.

Minimum requirements to score 15

Elongating the mouth is irrelevant, as it may be due to AU 20, AU 15, or AU 15 + 20.

(1) If the lip line is straight or slightly up in neutral face, then the lip corners must be pulled down at least slightly to score 15, or
(2) If lip line is slightly or barely down in neutral face, then the lip corners must be pulled down slightly more than neutral and not the result of AU 17 or AU 20.

[a] A wrinkle extending from beyond the nostril wings down to beyond the lip corners.

forehead, or eyelids. The table indicates where we have collapsed more than one muscle into a single Action Unit and where we have distinguished more than one Action Unit from a single muscle. The FACS names given in the table are shortened for convenience of recall and are not meant to describe the appearance change.

Table 9.2 lists an example of how an Action Unit (AU) is described in the FACS manual (Ekman & Friesen, 1978). The description includes four types of information:

1. The muscular basis of each AU is given in words and diagrams.
2. Detailed descriptions of the appearance changes are keyed to illustrative still photograph and film examples.
3. Instructions are given as to how to make the movement on one's own face. This aids in learning the appearance changes, particularly if FACS is learned by a group of people who can observe the variations in appearance on each others' faces. Learning how to do each AU also provides the user with a technique for later analyzing movements to be scored into their component parts. The user imitates the movement to be scored, noting which muscles had to be moved to produce the movement to be scored. By this means the scoring of any novel, complex facial action can be determined.
4. A rule is given specifying the minimal changes that must be observed in order to score a slight version of each AU.

The determination of the single AUs (Table 9.1) and their description (an example of which is shown in Table 9.2) were the first steps in developing FACS. The procedure of moving muscles, photographing the movement, and inspecting the pictures was reiterated for all possible combinations of two AUs. There was no need to describe AU combinations that could not interact. For example, pulling the lip corners down is done by a muscle that cannot affect the muscles controlling the position of the eyebrows. Two-way combinations were performed separately for the AUs controlling the brows, forehead, and upper and lower eyelids and for those AUs controlling the lower eyelids, cheeks, and lower regions of the face. There were several hundred combinations to perform and examine, for only in a very few instances did we discover that two AUs could not occur simultaneously.

Study of the photographs of the AU combinations showed that most of the appearance changes were additive. The characteristic appearance of each of the two-AU combinations was clearly recognizable and virtually unchanged. There were, however, a few AU combinations that were not additive. Their appearance changes may have incorporated some of the evidence of the single AUs, but new appearance changes from their joint action were also evident. All of these distinctive combinations were added to FACS, each described in the same detail as were the single AUs.

Inspection of the photographs of the AU combinations revealed that the appearance changes might be neither additive nor distinctive and that there might be a relationship of dominance, substitution, or alternation between AUs. In *dominance*, the strong AU overshadows the weak one. It may completely conceal the appearance of the subordinant AU or it may make the evidence of the subordinant AU very difficult to detect. In order to enhance agreement in scoring, rules were established to prohibit the scoring of subordinant AUs when there is clear evidence of a dominant AU. In *substitution*, the appearance of two different AU combinations is so similar that, in order to avoid disagreements, we designated only one of the combinations as the score to be used for either of the combinations. In *alternation*, two AUs cannot both be scored because both cannot be performed simultaneously; it is hard to distinguish one from the other; or the logic of other FACS rules does not allow both to be scored. The coder determines which of the two alternatives best describes a particular action.

After analyzing the pictures of all the two-AU combinations, the processes of performing, photographing, and then inspecting were reiterated, but this time with 3-AU combinations. Instead of hundreds there were thousands to so examine. Those that produced a distinctive, rather than an additive, combination of AUs were allotted their own entry in FACS with full descriptions as shown in Table 9.2. When we were ready to explore the 4-AU combinations, the number to be considered was so great that we decided to study them only selectively. On the basis of what we learned from the 2-AU and 3-AU combinations, we extrapolated those further combinations likely to result in distinctive facial movements. In total, between 4,000 and 5,000 facial combinations were performed and examined. This included *all* the possible combinations of AUs in the upper regions of the face and *all* the two- and three-way combinations in the lower face, plus *some* of the four-, five-, six-, seven-, and eight-AU combinations in the lower region of the face.

The manual for the *Facial Action Coding System* (Ekman & Friesen, 1978) was written in a self-instructional format, to serve as an initial tutor and subsequently as a reference in scoring facial behavior. It contains the following information:

1. textual material describing each AU listed in Table 9.1 in terms of its muscular basis, appearance changes, instructions for making the movement, and requirements to be met for scoring slight verisons (see the example in Table 9.2);
2. comparable information for each of more than 44 combinations of AUs;
3. a simple, less precise account of the 11 additional single AUs listed in Table 9.3, many of which do not involve the facial muscles (We have not described them in as much detail as was done in Table 9.2.);
4. descriptors that can be used to measure head and eye position;
5. tables comparing and contrasting over 400 AUs (or AU combinations) with only subtle differences;
6. scoring rules based on the dominance, alternation, and substitution relationships among AUs;

Table 9.3. *Simply defined AUs in FACS*

AU number	FACS name	AU number	FACS name
19	Tongue out	33	Cheek blow
21	Neck tightener	34	Cheek puff
29	Jaw thrust	35	Cheek suck
30	Jaw sideways	36	Tongue bulge
31	Jaw clench	37	Lip wipe
32	Lips bite		

7. a scoring sheet and a step-by-step procedure containing a number of internal checks designed to increase inter-rater reliability.

There are also still photographic and motion picture film examples of all the single AUs in Tables 9.1 and 9.3, of the 44 AU combinations, and the head and eye position descriptors. Additional still photographs and motion picture film examples of facial behavior are provided for practice in scoring facial movement. Correct scores are given, with commentary about the source of possible errors in scoring.

9.3. An example of scoring faces

It is not feasible in this chapter, without film or video, to illustrate the actual use of FACS in scoring a facial movement. The logic underlying FACS can be illustrated, however, with still photographs. For example, consider the seven facial behaviors shown in Figure 9.1. They all involve some common elements in appearance, in particular the down curve to the line of the mouth; they also differ. Analysis of these faces in terms of the single AUs involved will allow precise differentiation among them.

These seven faces include three single AUs and four combinations among these AUs. Figure 9.1 *A* is the appearance change resulting from AU 15, described earlier in Table 9.2. Figure 9.1*B* shows AU 17, described in Table 9.4; Figure 9.1*C* shows AU 10, also described in Table 9.4. If you read the verbal descriptions from Table 9.4 and match them to the photographs, you should then be able to "dissect" the other four faces in Figure 9.1 into their component AUs. Figure 9.1*D* combines AUs 10 and 15; Figure 9.1*E* combines AUs 10 and 17; Figure 9.1*F* combines AUs 15 and 17; Figure 9.1*G* combines AUs 10, 15, and 17.

Any complex facial behavior can be similarly analyzed into its component elements, if the single AUs have been learned and if rules regarding combinations have been studied. The scoring procedure leads the

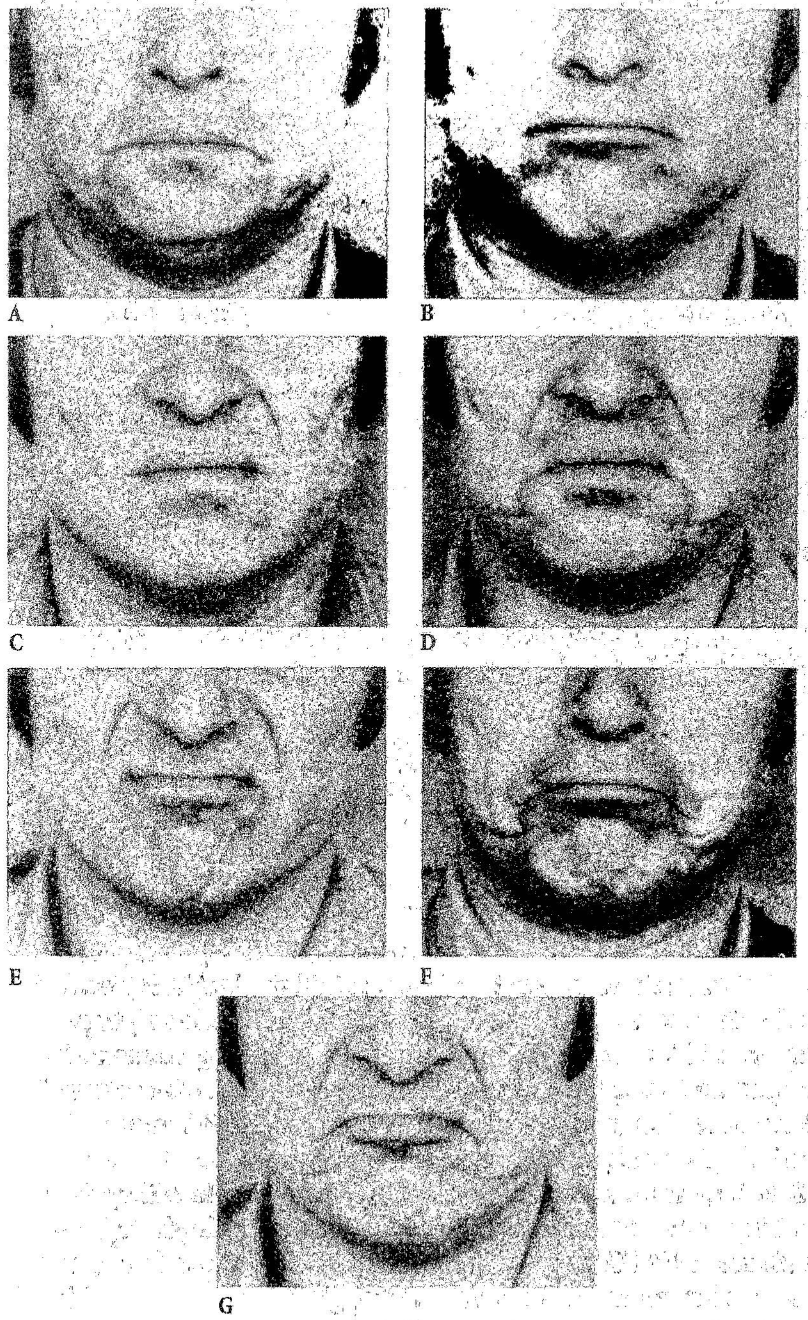

Figure 9.1 Action units that produce a down curve to the line of the mouth. (From "Measuring facial movement" by P. Ekman and W. V. Friesen, *Journal of Environmental Psychology and Nonverbal Behavior*, 1976, *1*, 56-75. (Copyright 1976 by P. Ekman. Reproduced by permission)

Table 9.4. *Appearance changes due to AU 10 and to AU 17.* (Copyright © Ekman & Friesen, 1976.)

Action Unit 10

The muscle underlying AU 10 emerges from the center of the infraorbital triangle[a] and attaches in the area of the nasolabial fold.[b] In AU 10 the skin above the upper lip is pulled upwards and towards the cheek, pulling the upper lip up.

(1) Raises the upper lip. Center of upper lip is drawn straight up, the outer portions of upper lip are drawn up but not as high as the center.
(2) Causes an angular bend in the shape of the upper lip.
(3) Raises the infraorbital triangle; and may cause the infraorbital furrow to appear, or if it is evident in neutral, to deepen.
(4) Deepens the nasolabial furrows and raises the upper part of this furrow producing a shape as ⁄‾\.
(5) Widens and raises the nostril wings.
(6) When the action is strong the lips will part.

Action Unit 17

The muscle underlying AU 17 emerges from an area below the lower lip and attaches far down the chin. In AU 17 the skin of the chin is pushed upwards, pushing up the lower lip.

(1) Pushes chin boss upward.
(2) Pushes lower lip upward.
(3) May cause wrinkles to appear on chin boss as skin is stretched, and may produce a depression medially under the lower lip.
(4) Causes shape of mouth to appear ⌒.
(5) If the action is strong the lower lip may protrude.

[a]Roughly the cheek area.
[b]A wrinkle extending from beyond the nostril wings down to beyond the lip corners.

user to break down any action into a set of single AU scores. When in doubt, the user is encouraged to consult the verbal descriptions, photographic and film examples, and tables of contrasting subtle differences. The person is also encouraged to imitate the action seen, observing the mirror reflection, and noting what AUs must be used in order to reproduce the action observed.

It is important not to be mislead by the example of Figure 9.1 into thinking that FACS was designed for scoring still photographs. The emphasis of FACS is on movement and its chief use is for scoring facial actions seen on motion records, although it can be used with stills if there is also a picture of a "neutral" face.

Figure 9.1 was used to demonstrate how FACS scoring differentiates among the seven facial behaviors shown. Although these seven behaviors are not visibly the same, are they the same functionally, psychologically, or communicatively? Is one a sadness expression, another a pout, another a disbelief gesture, etc.? It is only if the facial measurement distinguishes among these behaviors that we can determine empirically how many of the distinctions are useful. Once we can measure their separate occurrences, we can examine the contexts in which the behaviors occur or we can study preceding or consequent actions of other persons, isolate concomitant behavior in the person showing the behavior, study observers' inferences from viewing each behavior, etc.

The Facial Action Coding System far exceeded our initial anticipation of what would be required to provide a comprehensive, descriptive system for measuring facial action. Certainly, FACS is a very elaborate system, considerably more comprehensive than any previous system. There is no facial action described by other systems that cannot be described by FACS, and there are many behaviors described by FACS not previously distinguished by others. In addition, FACS allows for measuring facial asymmetries, where different AUs appear on each side of the face. It does not, however, include a measure of the intensity of actions for every AU, although it does so for four of the AUs listed in Table 9.1. It would be possible for others to follow the procedure used for these AUs in order to provide intensity of action scoring for the other AUs.

We are reasonably confident that FACS is complete for scoring the visible, reliably distinguishable actions of the brows, forehead, and eyelids. Unfortunately, FACS probably does not include all of the visible, reliably distinguishable actions in the lower part of the face.The hinged jaw and rubbery lips allow a nearly infinite number of actions. We included everything we could see, everything anyone else included and what are probably the most common elements and combinations of actions in the lower part of the face among children and adults. As we and others use FACS, we expect that some other AUs may need to be added, although we hope not many. Also, others may be interested in more finely discriminating separate AUs from the list of broadly defined AUs in Table 9.3.

Some will ask the question whether FACS is too elaborate, too comprehensive, and too detailed? We believe it has been useful to attempt an approximation of the total repertoire of facial action, to isolate minimal Action Units that can combine to account for any facial movement. At the least, FACS provides a means to cross-reference the different scoring categories used by others with a common nomenclature. It may also serve to advise the investigators of their options, so that they may make explicit decisions about what to include in and what to omit from their measurements. No one knows at the outset how many of the variations in facial behavior can be ignored in any research study without losing important information. In preliminary observations, or pilot studies, investigators may wish to use FACS to make comprehensive measurements and, then, based on these results, to score selectively only certain AUs or AU combinations in their main study.

Apart from these more selective uses of FACS, there will be some who simply need a comprehensive measurement system. If we wish to learn all the facial actions that signal emotion (and those that do not) or whether facial emphasis markers are the same regardless of the content of speech so emphasized (to mention just two current interests), then a method such as FACS is needed.

9.4. Reliability of the Facial Action Coding System

Reliability issues

The fundamental reliability issue is whether independent persons would agree in their scoring of facial behavior; more specifically, whether persons who learn FACS without instruction from the developers would agree among themselves and/or with the developers.

To score facial behavior, two different operations, description and location, and thus two different reliability issues are required. By *description*, we mean what happened: What are the Action Units responsible for an observed change in facial behavior? By *location*, we mean when did it happen: At precisely what moment did whatever happened start and stop. Suppose the brows moved. To describe the movement, we would ask which type of movement it was: did the brows raise, lower, raise and draw together? did just the inner part raise or the entire brow, etc.? To locate the movement, we would ask at what videoframe (1/60 second) the movement, whatever it is, started and at what videoframe it ended? The two questions are independent to some extent. Reliability could be high for description but low for location, or vice versa.

For either description or location, reliability can be evaluated on either of two bases: (1) agreement among independent persons, or (2) agreement between a learner and an expert. We were interested in not only whether there was intercoder agreement but whether those who learned FACS without instruction from us would score facial behavior the way

we did. Data were reported for both types of agreement. The results were about the same.

The description of facial movement with FACS involves four operations and the reliability of each was studied:

1. *Determining which AUs are responsible for the observed movement.* The coder learns how to recognize the appearance changes due to each of 44 AUs, singly and in combination. The logic of the system is that any movement can be scored in terms of which AUs produced it. Theoretically, it is possible for about 20 AUs to combine to produce a single facial movement or as few as one. (All 44 cannot combine because some involve antagonistic actions and the occurrence of some actions conceals the possible presence of others.)
2. *Scoring the intensity of action for five of the 44 AUs.* While intensity scoring could have been provided for each and every one of the 44 AUs, we used intensity scoring only where we thought the magnitude of action could influence the recognition of a particular action unit or a related action. Intensity was scored in terms of three levels: low, medium, and high.
3. *Scoring whether an action is asymmetrical or unilateral.* FACS distinguishes unilateral actions, where there is no evidence of the action on one side of the face, from asymmetrical actions, where the movement is evident on both sides of the face but stronger on one side. There were very few instances of unilateral actions in the records we scored, too few to estimate the reliability of distinguishing unilateral from bilateral or asymmetrical actions. Later the reliability of asymmetry scoring will be reported.
4. *Scoring the position of the head and the position of the eyes during a facial movement.* This descriptive system is less rigorous than that provided for the AUs. Fourteen descriptors are provided, of which up to six can be scored for any event. Because head/eye scoring is a simpler system, agreement on it could have inflated agreement measures on the total scoring of a face. Results will, therefore, be reported separately, including and excluding head/eye position scores. In fact, however, it made little difference.

A final issue to be considered is whether agreement is substantially improved by having the independent coders arbitrate their disagreements. The agreement achieved by six independent coders (intercoder agreement and agreement with experts) will be contrasted with agreement achieved by three pairs of arbitrated scores. Arbitration improved agreement, but not by much.

The behavioral sample for initial reliability studies

For our initial reliability study, we selected behavior samples from 10 of the honest–deceptive interviews we have been studying over the past eight years (Ekman & Friesen, 1974a; Ekman, Friesen & Scherer, 1976). We selected the first two actions shown by a subject who was conversing about her reactions to a film while she was watching it and the first two actions shown when the interview continued after the film ended.

In order to increase the variety of behaviors subject to scoring, if the first two actions repeated an AU or AU combination already selected more than once, then the next nonredundant action was taken. By these means, a total of 40 items was obtained. Six were dropped because the videopicture was not acceptable, leaving 34 items.

Coders were given the videotape with the instruction to score whatever occurred within each of the 34 events. Note that by defining each event ahead of time, giving the coders the start and stop frame within which they should score, we eliminated decisions about location and studied just description reliability.

The coders

Seven persons previously unfamiliar with FACS learned the system as a group during a 5-week period, working on a half-time basis. We had minimum contact with them during this time so that their performance can be considered a fair test of whether FACS produces reliable scoring when learned without instruction from the developers. The results reported are based on six persons because one coder did not continue.

The six coders consisted of five women and one man. Two were research assistants who have bachelor's level education. Two were doctoral candidates, one in psychology, another in linguistics. Another was a postdoctoral fellow trained in developmental psychology. The last was a visiting associate professor of clinical psychology whose native language is German.

Procedure

The six coders independently scored the 34 events without any communication among them. After their scoring was completed, the six were grouped into three pairs and given their scorings on any event about which they disagreed. They were required to arrive jointly at an arbitrated final scoring.

We (the authors) jointly scored each of the 34 events. We then examined the scoring of the six learners and considered whether we would want to change our scoring in light of their performance. We did so only a few times, and those decisions did not increase the agreement between them and us.

Table 9.5. *Example of raw scores on one behavioral event*

Coder	AU numbers			
Experts	1 + 4 +		7	
Blossom	1 + 4 +		7	
Kathy	1 + 4 +	6		
Charlotte	4 +	5X +	7+	10X[a]
Linda	1 + 4 +		7	
Sonia	4 +		7	
Rainer	4 +		7	
Arbitrated				
Bl–Ka	1 + 4 +		7	
Ch–Li	1 + 4 +		7 +	10X[a]
So–Ra[b]	4 +		7	

[a] In FACS, X denotes a rating of low intensity.
[b] Note that arbitration was not necessary here as subjects were in agreement on original scoring.

Raw data matrix

Thirty-four events were scored by six independent persons producing 6 x 34 = 204 sets of action unit scores. Additionally, there are the three arbitrated pair scorings for each event. Table 9.5 shows the scores for one of the 34 items. The first row is our scoring. The next six rows show the scoring of this event by each of the six persons. The final three rows show the arbitrated scoring of the three pairings. (Note that Sonia and Rainer agreed on this event so they did not arbitrate.) The entries are the numbers for the AUs, which is the system used to record scores. The experts scored three AUs, 1, 4, and 7, which describe raising the inner corners of the brow (1), pulling the brows together (4), and tightening of the eyelids (7). There was agreement among all coders that AU 4 was present. Some did not score AU 1. One coder scored an outer eyelid action (6) rather than the inner eyelid action of AU 7. One coder also scored an upper eyelid raise (5X; the X meaning that she scored it as being low in intensity); and a low upper lip raise (10X).

An index of agreement

It was not obvious what type of measure of agreement should be employed. Reliability measures often are applied to situations where scoring involves a binary decision (e.g., present or absent) or assignment into one of a series of exclusive categories. In FACS there is a range of possible scores, from 1 to about 26 (about 20 AUs and 6 head/eye descriptors) that could be scored for any one event. There are many more opportunities for disagreement than is usually the case in psychological measurement.

We could have assessed reliability for each AU separately, determining how many times the six persons agreed about its presence or absence over the 34 items. This method, often used in reliability studies, would give as much credit to an agreement that an AU was not scored for an event as an agreement that it was scored. Such a method would have produced reliability scores much higher than the procedure we selected.

The index of agreement that we employed (Wexler, 1972) was a ratio calculated separately for each of the 34 events for each pair of coders and for each coder compared to the expert scoring. The arbitrated scoring was also evaluated with the same index. The formula was:

$$\frac{(\text{number of AUs on which coder 1 and coder 2 agreed}) \times 2}{\text{total number of AUs scored by the two coders}}$$

For example, if the scoring by one coder was 1 + 5 + 7 + 22 and the scoring by a second coder was 1 + 7 + 16, the ratio would be:

$$\frac{4\ (2\ \text{AUs agreed upon} \times 2)}{7\ (\text{total number of AUs scored by two coders})} = .57$$

Table 9.6 shows the matrix of ratios generated with this formula for the raw data shown in Table 9.5. The first six rows of Table 9.6 give the ratios calculated for the scoring of each individual person. The last three rows of the table give the ratios when the scoring reached through arbitration by a pair of persons was evaluated. We will use the first six rows to illustrate how the ratio represents agreement. The first column of numbers shows the ratio when each coder's scoring was entered into the formula with the scoring of the experts. Perfect agreement (in the case of Blossom and Linda) generated 1.00 ratios. Disagreements generated lower ratios. The other columns in the table show the ratios between each pair of coders. One can see that Sonia and Rainer agreed exactly, as did Linda and Blossom. The maximum disagreement was between Kathy and Charlotte.

The mathematics of the formula used are such that if only one or two AUs are scored for an event, a disagreement will lower the ratio more

Table 9.6. *Matrix of agreement ratios for the scoring of one behavioral event*

Coder	Experts	Blossom	Kathy	Charlotte	Linda	Sonia
Single-person scoring						
Blossom	1.000					
Kathy	.667	.667				
Charlotte	.571	.571	.286			
Linda	1.000	1.000	.667	.571		
Sonia	.800	.800	.400	.667	.800	
Rainer	.800	.800	.400	.667	.800	1.000
	Experts	Bl–Ka	Ch–Li			
Arbitrated pairs scoring						
Bl–Ka	1.000					
Ch–Li	.857	.857				
So–Ra	.800	.800	.667			

than if six of seven are scored. If two coders disagreed about only one AU and agreed about one AU, they would earn a ratio of .50. If they disagreed about one AU and agreed about four AUs, the ratio would be .80. Even though the disagreement is in both instances about only one score, it seems reasonable that the formula rewards agreement on a high proportion of actions that are present.

We checked on how many AUs were scored for each of the 34 events by the experts. The mode was three scores for an event, with about one-third of the 34 events having one or two scores and one-third having four to seven scores. Thus, if the absolute number of scores distorted the ratio of agreement, the 34 events produced a balanced distribution in this regard.

Two matrixes were generated. One matrix was composed of the ratios derived by comparing each person's scoring of each event with the experts' scoring, generating 204 data points (6 persons x 34 events). The second matrix disregarded the experts' scoring and calculated the ratio by comparing each person's scoring with each other person. With six persons, five such ratios were generated for each person (comparing that person with every other person) for each of the 34 events scored. The mean of those five ratios was taken as the measure of a particular person's average agreement with others for a particular event. This yielded a second matrix that again had 204 points, with each point representing the mean ratio of agreement with the other person's for each event scored (34 events x 6 persons).

Overall agreement

The mean ratio across all coders (6) and all events scored (34) was .822 when scoring was compared with experts' and .756 when intercoder agreement was evaluated. Figure 9.2 shows that the distributions of ratios were skewed toward high agreement. For example, 141 out of 204 ratios of agreement with the experts were .80 or above, and only 28 out of the 204 ratios were below .60. The figure also shows that the distribution of ratios representing intercoder agreement was similarly skewed toward agreement, with just as few low-value ratios but not as many ratios above .80 as when agreement with experts was calculated.

Since FACS was first published, more than thirty investigators have learned FACS using the training materials provided without any direct contact with us. Part of the training package makes available the videotape of the same behavior sample coded by the initial group of six learners. All of these new investigators achieved comparable reliability to the data reported in Figure 9.2.

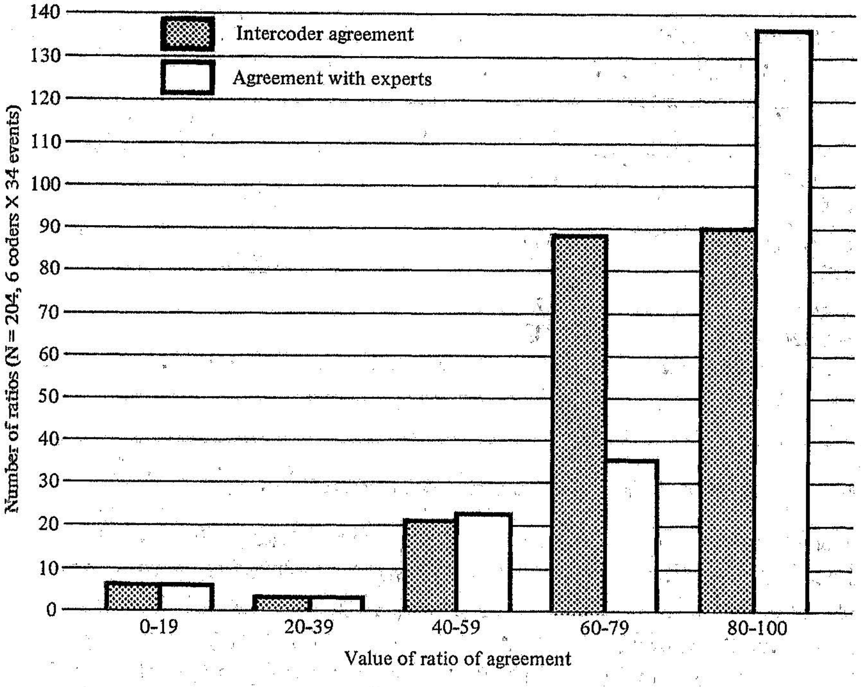

Figure 9.2 Distribution of agreement ratios among coders and between coders and experts. (From *The facial action coding system* by P. Ekman and W. V. Friesen. Palo Alto, Ca.: Consulting Psychologists Press, 1978. Copyright 1978 by P. Ekman. Reproduced by permission)

Did scoring head/eye position inflate reliability?

The answer is no. Recall that the measurement of head/eye position was a grosser descriptive scheme than that of the Action Units. Agreement on what might be an easier set of decisions might have inflated the agreement ratios, concealing disagreements about the scoring of AUs. When head/eye position scores were disregarded, however, and the ratios recalculated, the mean ratio across all coders and all events was .816 (as compared with .822 including head/eye) when scoring was compared against experts and .745 (as compared with .756 including head/eye) for intercoder agreement. The distributions were examined and were found to be not noticeably different from those shown in Figure 9.2. Results reported hereafter will include the head/eye position scores.

Table 9.7. *Mean ratios of agreement with experts, showing the benefits of having coders arbitrate their disagreements in pairs*

	Individual scoring	Arbitrated pairs
Blossom	.782	
Kathy	.827	.869
Charlotte	.859	
Linda	.858	.886
Sonia	.873	
Rainer	.732	.883

Does arbitrating differences enhance agreement?

The answer is slightly, but it depends on how much they disagreed and how low their individual agreement was prior to arbitration. Presenting the coders with their disagreements and asking them to arbitrate their differences could have produced lower, rather than higher, agreement. Each pair after arbitrating might diverge more from the other pairs or from the experts. Instead there was a slight increase in both agreement measures.

The mean ratio across all coders and all events went up from .822 to .863 in terms of agreement with experts and from .756 to .809 in terms of intercoder agreement. Table 9.7 shows that in terms of agreement with experts the benefit was negligible for the pair who had high agreement individually (Charlotte and Linda), moderate for a pair somewhat lower individually (Blossom and Kathy), and considerable for the pair where one member (Rainer) had the lowest coefficient of agreement. His gain through arbitration, however, was at the cost of a loss of the person with whom he arbitrated (Sonia). When the same comparisons were made utilizing the measures of intercoder agreement (rather than agreement with experts as shown in Table 9.7), the values were two to three hundreths lower but the pattern was the same. For example, the mean ratio of intercoder agreement for Sonia and Rainer's arbitrated scoring was .802 as compared with .833 for agreement with experts.

Two other methods for reconciling disagreements were explored. In one, a simple flip of the coin was used to determine who was "correct" on each disagreement. Using the coin flip as the basis for saying what the final score should be for items where a pair disagreed yielded ratios of agreement with the experts that were just as high as arbitration for the coder pairs who did not initially disagree strongly (Blossom and Kathy; Charlotte and Linda). For the pair with the one coder who showed the lowest agreement with the experts (Rainer), a coin flip did not yield as much increased agreement as did arbitration. The second method for resolving disagreements consisted of applying a set of logical rules to determine who was "correct" for any events where a pair disagreed. These rules benefited the pair who most disagreed (Sonia and Rainer) as much as did arbitration.

Agreement about intensity

The data analysis thus far has ignored any disagreements about intensity. Such disagreement could have occurred on the scoring of only 5 of the 44 AUs because FACS provides for intensity scoring on just those few. There were 19 instances in which the experts scored the AUs with an intensity rating, providing 114 opportunities for agreement with the experts (6 coders x 19 instances).

Exact agreement about intensity was reached on 55% of these scorings. Recall that intensity involves a three-point scale. There were no two-point disparities; instead about half the disagreements were one-point disparities, the other half were when one person entirely missed scoring an intensity-AU that was scored by the experts at the low-intensity level.

The scorings of the pairs of persons who disagreed on intensity were subject to arbitration, which enhanced the agreement with experts. Exact agreement about intensity rose to 74%.

Recall that the data reported previously disregarded disparities in intensity scores. The agreement ratios for each of the six coders compared with the experts' scoring were recalculated with a disagreement about intensity considered as a total disagreement. The mean ratio across all six persons and all 34 events was .778 when a difference of intensity was considered an error, as compared to .822 when intensity disagreement was ignored. Of course the reason why the ratio of agreement did not decrease further was that there were not that many instances where intensity could be scored. In another behavioral sample, in which there was a preponderance of behavior involving AUs for which intensity could be scored, the ratios of agreement might be lower.

Representativeness of the behavioral sample

The scores were tabulated for each AU across all coders and all events to provide a picture of the extent to which the behavioral sample offered opportunity for testing the reliability of all the AUs. For this tabulation we considered not only whether an AU was scored but also whether an AU was considered, even if not scored, during the coders' step-by-step scoring procedure. (Such information is readily retrieved from the scoring sheets on which the coders recorded every AU considered.)

Twenty-five of the 44 AUs were scored or considered many times; only 19 of the AUs were scored or considered less than 10 times. These 19 AUs are probably rare occurrences in most conversations between adults; for example, sticking out the tongue, tightening the *platysma* muscle, sucking the lips in to cover the teeth, puffing out the cheeks, etc. Although we cannot generalize from this study to the reliability that might be obtained if the behavior scored included such actions, there is no reason to suspect that reliability would be lower. Quite the contrary, the classification of many of these infrequent AUs probably involves an easier set of discriminations than is required for the AUs that were considered and scored in this study.

Reliability in the location of facial action

In the beginning we distinguished between two aspects of measuring facial action, description (what happened) and location (when something happened). Thus far we have considered only description. Let us now consider the reliability of scoring location.

Information is available from a dissertation by Ancoli (1979). Subjects sat alone in a room and watched two films. One film showed scenes that other subjects had rated as causing pleasant feelings. The other film had been rated as producing feelings of disgust and fear. The subjects were monitored on EEG, heart rate, EMG on skeletal muscles, and respiration. In addition, a videotape was made of their faces. Ancoli scored all of the facial behavior shown by 35 subjects, a total of 3 minutes during a pleasant film and 2 minutes during the unpleasant film for each subject.

Reliability was evaluated at two points in the study. After she scored the first 10 subjects, Ancoli randomly selected the facial behavior during one of the two films for each subject. This sample was then scored by a second person (Linda Camras, another of the people who had recently learned FACS). Later, a second sample was drawn, consisting of a 30-second period from the video records of each of the 25 remaining subjects. Again, Camras scored the randomly selected sample.

Location, unlike FACS description, can be regarded as a binary decision – something is happening or not at each frame in time. The decision should be easy with a large facial movement or when the face is completely still; it should be difficult when there is a very small movement. A set of minimum requirements is provided by FACS for the amount of change that must occur before a movement can be scored. The most difficult decision, and the main opportunity for disagreement, is when there is a small movement and the person must evaluate whether it is sufficient to meet FACS requirements for scoring. If it does not, the coders treat it as a no-movement.

When occur versus no-occur decisions are made point by point in time, a common way to assess reliability is to determine for each point in time whether two independent persons agree. Agreement is then represented as a percentage of total time considered. Each .10 second was so examined. In sample 1, the two coders agreed (as to whether or not something was occurring) 89% of the time. In sample 2, the two coders agreed 95% of the time. This calculation gave equal credit to agreement that nothing happened as to agreement that something happened. If the sample contained long periods of time in which the face was inactive, this measure of locational agreement would be inflated. In sample 1, the face was totally inactive or not scorable (action occurred but did not meet the minimum requirements demanded by FACS) 69% of the time; in sample 2, the face was inactive or not scorable 66% of the time.

There is, of course, quite a difference between agreement that nothing has occurred and agreement that something unscorable has occurred (i.e., that it does not meet the minimum requirements specified by FACS). Agreement about an unscorable action should represent the most difficult locational decision. Since Ancoli's study we have added a new action descriptor to FACS for unscorable actions. If this had been available in Ancoli's study, it would have been possible to calculate the percentage of time two coders agreed that a scorable action occurred, that an unscorable action occurred, and that no action occurred. Now that unscorable actions have been included in the scoring procedure, in future studies using FACS, we recommend that locational agreement be so examined. It will then be possible to isolate disagreements where one person said the action was scorable and another called it unscorable, and instances where one person said the action was unscorable and the other recorded no action. In either case, additional instruction can be given to increase locational reliability if a consistent pattern is found, consistent, that is, for a particular coder or a particular AU.

Another way to examine agreement about location, which avoids the problem of inflating the estimate by agreements on the absence of ac-

Table 9.8. *Percentage of agreement on total events located*

	Agree on beginning		Agree on end		Agree on beginning and end	
	within .10 sec	within .5 sec	within .10 sec	within .5 sec	within 1 sec	within 2 sec
Sample 1	25.0	59.1	13.6	38.6	47.7	68.2
Sample 2	64.5	74.5	61.3	67.7	74.2	74.2

Table 9.9. *Mean description reliability ratios on agreement*

	Including events scored by only one	Including only events scored by both	Including only events scored by both and excluding events agreed to be not scorable
Sample 1	.722	.878	.815
Sample 2	.791	.909	.824

tion, is to consider the occurrence of complete disagreements. The worst error in location is when one person scores an event that the other fails to score (either because they missed the event entirely or judged it as not reaching the minimum requirements dictated by FACS). In sample 1, such complete disagreement occurred with 18.4% of the behavior scored; in sample 2, such complete disagreement occurred with 12.9% of the behavior scored.

Location reliability can be studied in more detail by examining exactly how closely coders designated when an event began and when it ended. Table 9.8 shows that information for the two samples of Ancoli. The calculation of percentage of agreement used the total events scored by both (including events seen by only one) as the denominator. Agreement was higher for judgments of when an action began than for judgments of when it ended. Agreement was higher in sample 2 than in sample 1, perhaps because of experience. The last two columns in Table 9.8 show the percentage of events where both persons agreed within 1 second and within 2 seconds on both the start and stop of an action. A high percentage of agreement was found in sample 2.

Another look at description reliability

Ancoli's dissertation presents another opportunity to study the reliability of the FACS description. Table 9.9 reports the ratios of agreement (calculated as explained previously) for the two behavioral samples. The first numerical column shows the mean ratio when the events scored by only one coder were included in calculating the mean across all events. For those events scored by only one coder, the ratio was zero, thus allowing disagreement about location to lower the measure of agreement on description. The next column gives the percentage that include in the calculation of the agreement ratios only events scored by both persons. These ratios include agreements that in certain instances there was no scorable facial action. That is, of course, an important type of agreement, but it is not the same as agreement about how to describe what is present. In the last column are shown the ratios calculated excluding items in which both coders agreed that the event was a no-score, or neutral, action. The figures in this column are directly comparable to the ratios reported earlier for intercoder agreement among six persons because in that reliability test no neutral events were included in the behavior sample and the ratios are not deflated by disagreements about location. Agreement remains about as it was for the learners described previously.

Now that FACS provides an unscorable action descriptor, it is possible to analyze description reliability with one further refinement not shown in Table 9.9. Agreement ratios would be calculated for all events considered scorable or unscorable by one or the other of the coders, excluding from the ratio only agreements that no action had occurred. These ratios would give credit for any agreements that unscorable activity occurred.

9.5. Validity

The Facial Action Coding System was designed to measure any facial movement, not just movements that might be relevant to emotion. Although FACS contains hypotheses about the particular actions that signal each emotion, these hypotheses are quite separate from FACS and even if they were shown to be incorrect, FACS would still be valid as a measurement technique if it were found to measure the behavior it claims to measure. Conversely, the hypotheses about emotion signals could be correct and yet FACS might not be valid. There is a difference then between the validity of a measurement technique and the truth or falsity of any set of hypotheses about what the measurements might mean.

Figure 9.3

Descriptive validity

The validity of FACS requires evidence that it actually measures the behavior it claims to measure. Because FACS identifies the elemental muscular actions that singly or in combination produce any visible movement observed, the question is whether the muscular actions identified by FACS are the particular ones that actually produce a particular movement. Of course there may be more than one action that can produce a momentary change in facial appearance. Consider an expression entailing a down-curved lower lip, as shown in the drawing in Figure 9.3. If FACS scored this appearance as Action Unit 15 it is saying that the down curve to the lower lip was produced by a muscle which pulls the corners of the lips down (see Figure 9.1). Perhaps the down curve to the lower lip was produced by Action Unit 17 which pushes the chin and lower lip up in the center, or by Action 15 and Action 17 (see Figure 9.1).

In addition, FACS measures the intensity of some facial actions, such as whether the pulling down of the lower lip was slight, moderate, or extreme. The validity question is whether such measurements correspond to known differences in the intensity of such an action. The problem is how to know what facial action actually occurred; that is, what criterion to utilize independent of FACS? Two approaches have been taken: performed action and electromyography.

In the first approach, Ekman and Friesen trained a number of people to perform voluntarily various actions on request. Videotape records of such performances were scored without knowledge of the performances that had been requested. FACS accurately distinguished the actions the performers had been instructed to make.

In the other approach, in joint study with Schwartz, Ekman and Friesen placed EMG leads on the faces of subjects who were asked to produce actions on request. Utilizing the EMG measure of electrical activity in one or another muscle region as the criterion, FACS was found to differentiate accurately the *type* of action. This study also showed that FACS measurement of the *intensity* of facial action was valid; FACS scoring of intensity was highly correlated with EMG readings (Pearson $R = .85$).

Validity of emotion-signal hypotheses

The Facial Action Coding System contains more than a thousand hypotheses about the particular combinations of facial actions that signal the *type* of emotion (anger, fear, surprise, etc.), the variations in the *intensity* of each emotion, the *blends* of emotions, and the signs of attempts to control emotion. The problem in validating these hypotheses is to have some criterion independent of FACS to determine just which emotion, at what intensity, is being experienced at any given moment by a particular person. The traditional approach has been to use a poser's intent as the criterion. Such studies using FACS to measure the poser's facial actions have verified many of the hypotheses about emotion signals contained in FACS, but for the many reasons discussed in Chapter 4, one cannot generalize from posed to spontaneous emotions. Additionally, posing is probably one of the easiest types of facial behavior to measure. The onset is usually coordinated and abrupt, the apex frozen, and the scope very intense and exaggerated. Success with poses is no guarantee of success when emotional expressions occur spontaneously. Let us consider, then, a number of studies using spontaneous emotional behavior.

In one study autonomic nervous system changes were used as the criterion to test hypotheses about facial signs of emotion. The question asked was whether there was a difference in ANS activity when the face showed what FACS considered emotion or nonemotional activity? Ancoli (Ancoli, 1979; Ancoli, Kamiya, & Ekman, 1980) had female subjects watch pleasant and unpleasant films seated alone in a room, while heart rate, GSR, respiration and EEG were measured. The subjects did not know that their facial activity was being videotaped. When the subjects' ANS responses to the two films were compared, differences in the pattern of changes were found only when the face showed what FACS identified as positive or negative emotions. One limit of this study, however, was that it could only validate the predictions about the gross distinction between positive versus negative emotions, not finer distinctions within positive or negative emotional experience.

In another study of the subjects in this experiment, the subjective reports of emotional experience were utilized as the criterion to validate FACS predictions about finer distinctions among emotions. Ekman, Friesen, and Ancoli (1980) found that FACS hypotheses about a number of aspects of emotional experience predicted the subjects' report on multidimensional scales immediately after viewing the pleasant and unpleasant films. That is, FACS was able to discriminate (1) the intensity of happy feelings, (2) which of two happy experiences was the happiest, (3) the intensity of negative feelings, and (4) the occurrence of disgust as compared with fear, anger, or sadness.

In another study an environmental event was used as the criterion of what emotion was experienced. Ekman, Friesen, and Simons (1982) studied facial reactions in response to a blank pistol shot to test hypotheses about the facial startle response. They were able to discriminate the uniform pattern of facial actions that always occurs in response to such an unanticipated very loud noise from idiosyncratic responses. The pattern included the type of action that occurred and also the timing of the action, i.e., how quickly it began after the gun shot.

In yet another study, the emotion being experienced was identified in terms of the characteristics of the person showing the expression. Krause (1981) reasoned from the clinical literature that people who stutter should show many signs of anger during a conversation. Support for FACS predictions about the particular actions that signify anger was obtained because stutterers showed more of these particular actions than nonstutterers.

A variation in the experimental conditions during which an expression occurs was employed as the criterion in another study addressing a fundamental question about facial expressions of emotion. Is it possible to discriminate a purposeful facial action from a naturally occurring emotional expression? Ekman, Hager, and Friesen (1981) compared facial movements performed on request with naturally occurring movements in response to a joke or while watching pleasant or unpleasant films. They found that the requested actions were asymmetrical more often than the spontaneous emotional expressions and that the actions usually were more intense on the left side of the face for the requested, but not for the spontaneous, facial actions.

Many more studies are needed, of course, to validate all of the hypotheses contained in FACS and to replicate the studies just reported.

Utility

Because FACS was designed to measure any type of facial behavior, not just actions relevant to emotion, it is reasonable to ask whether there is evidence that FACS has general-purpose utility? Evidence of utility would require demonstration that FACS can measure people of different ages and can measure facial behavior that is providing signals other than emotional ones.

Oster (Oster, 1978; Oster and Ekman, 1978) provided information about some of the expression differences between the neonate and the young infant. For example, when the brows are raised, the horizontal furrows apparent in the child or adult will not appear in the neonate because of the fatty pad in the forehead. When FACS is used with Oster's modifications, all of the facial actions can be measured. Oster's studies provided evidence that certain complex, spontaneous facial actions observed in young infants are not random but represent organized patterns and sequences of facial muscle activity that are reliably related to other aspects of the infant's behavior.

The Facial Action Coding System has been found useful in studying conversational facial signals in which the facial action serves to illustrate or punctuate speech. Camras (in prep.) found differences in the syntactic form of questions that do and do not contain facial actions functioning as "question-markers." Ekman, Camras, and Friesen (reported by Ekman, 1980) found that the semantic context predicts which of two facial actions is used to provide speech emphasis. Baker (1982) used FACS to identify the facial actions shown by deaf persons when they sign. She was able to isolate particular combinations of facial actions that appear to serve syntactic functions.

9.6. Conclusion

Since it was first published, more than forty people have learned FACS and are beginning to use it to study quite diverse phenomena. Others have been learning Izard's (1979) technique for measuring facial movement. The next decade should see a great growth in knowledge about the face, now that the tools are available to measure facial action itself.