Report #17-UI-051

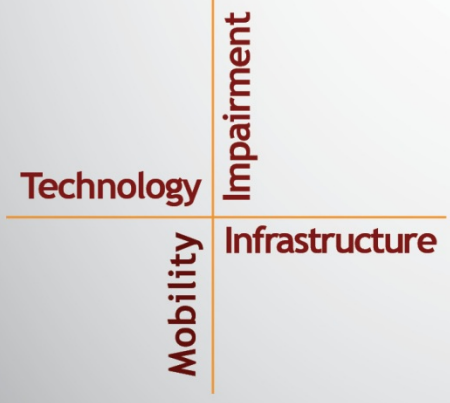
NSTSCE

National Surface Transportation Safety Center for Excellence

Development of a Protocol to Classify Drivers' Emotional Conversation

Greg Fitch • Sheldon M Russell • Julie McClafferty

Submitted: May 4, 2017



Housed at the Virginia Tech Transportation Institute
3500 Transportation Research Plaza • Blacksburg, Virginia 24061

ACKNOWLEDGMENTS

The authors of this report would like to acknowledge the support of the stakeholders of the National Surface Transportation Safety Center for Excellence (NSTSCE): Tom Dingus from the Virginia Tech Transportation Institute, John Capp from General Motors Corporation, Chris Hayes from Travelers Insurance, Martin Walker from the Federal Motor Carrier Safety Administration, and Cathy McGhee from the Virginia Department of Transportation and the Virginia Center for Transportation Research.

The NSTSCE stakeholders have jointly funded this research for the purpose of developing and disseminating advanced transportation safety techniques and innovations.

EXECUTIVE SUMMARY

To facilitate future analyses of emotion in naturalistic driving study (NDS) data, a protocol was developed to rate the emotional content of video samples collected during NDS. The protocol required data reductionists to observe video footage of the driver's face and rate their emotional demeanor in a reasonable amount of time.

The Facial Action Coding System (FACS; Ekman, 1978) was used to guide the development of the emotion reduction protocol. Similar to FACS, the protocol instructed reductionists how to classify the driver's emotion into one of six categories: Neutral/No Emotion Shown, Happy, Angry/Frustrated, Sad, Surprised, and Other. Once reductionists rated the type of emotion expressed by a driver, they then indicated the intensity of the emotion expression, using a four-point scale derived from the five-point scale used in FACS. Although FACS was used to guide development, the protocol was developed to capture the overall emotion of the driver, not necessarily specific facial muscle activations on a frame-by-frame basis.

Seventy-two cases for reduction were selected from previously collected NDS data drawn from studies of light vehicle drivers and heavy-truck drivers (Blanco et al., in press; Fitch et al., 2013; Hanowski et al., 2008). Each case was categorized by the experimenters for its specific emotion and intensity level content. The protocol was applied by two groups of reductionists, experienced and novice, in order to determine if training level would impact ratings. Results showed that both experienced and novice reductionists rated cases with similar levels of reliability. Furthermore, both groups of reductionists exhibited inter-rater reliability that was significantly different than chance for all rating types.

For both experienced and novice reductionists, accuracy was moderate to good; however, there was evidence of confusion for certain cases. Specifically, confusion existed when a driver exhibited low-intensity emotion. Rescoring the accuracy results to estimate if emotional content was presented by a driver (originally rated as marked or severe emotion present) and or not presented by the driver (originally rated as no emotion or slight emotion) further improved the reductionists' accuracy. Accuracy using rescored data was 85%, suggesting a high degree of accuracy for detecting emotion reaction. It is expected that future iterations of the protocol will show improved accuracy with slight modifications.

Future work applying the protocol to other NDS data sets can support the investigation of emotional cell phone conversation while driving. With further development, the protocol will ultimately be used to shed additional insight into the safety-critical event (SCE) risk of cell phone conversations while driving, and has the potential to be developed for use as a generic and standardized means of classifying the emotions experienced by drivers not only in naturalistic driving studies, but also in driving studies using other methods, including simulation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

EMFACS	Emotional Facial Action Coding System
FACS	Facial Action Coding System
MVOSS	Motor Vehicle Occupant Safety Survey
NDS	naturalistic driving study
NHTSA	National Highway Traffic Safety Administration
NOPUS	National Occupant Protection Use Survey
SCE	safety-critical event

CHAPTER 1. INTRODUCTION

Naturalistic driving data provide a valuable data set that can be used to study driver use of cell phones, including how emotions present during a phone conversation may impact driving performance. There has been little research focusing on the emotional content of conversations in the driving context, and there are mixed findings in the literature regarding the relationship of emotion to driving. Emotion has been shown to have both positive and negative effects on driving performance (Cai & Lin, 2011; Grimm et al., 2007). Emotional conversation has been reported to lead to driver error and visual tunneling (Briggs, Hole, & Land, 2011). In another study, Briggs et al. (2011) found that emotional conversation, as exhibited by drivers involved in a conversation about their fear, increased their mental workload, led to more driving errors in a driving simulator, and induced cognitive tunneling. Investigating the effect of emotional cell phone conversation on driver performance and the associated risk of a safety-critical event (SCE) using naturalistic driving data is desired, but first requires a method for identifying drivers' emotional state.

FOUNDATION FOR THE PROTOCOL – THE FACIAL ACTION CODING SYSTEM

Significant research exists on how to assess peoples' emotional expressions by analyzing images of their face. As early as 1978, Paul Ekman developed the Facial Action Coding System (FACS; Ekman, & Friesen, 1978). FACS is a detailed, anatomically based coding system that describes how to categorize facial behaviors by perceiving the activation and relaxation of specific facial muscles, called action units. The coding scheme has been updated several times, with the latest version being FACS 2002. Typically, becoming FACS certified requires 100 hours of training (Hager, 2003e). Furthermore, because of the work involved in identifying specific action units, a 1-minute video typically takes 3 hours to code (Movellan, Frank, Bartlett, & Sejnowski, 2013). As such, the method is quite laborious.

The FACS manual teaches the specific action units but does not teach what they mean (Hager, 2003a), thus ensuring that FACS coding can be objective without a rater's biases. However, under the assumption that facial expressions have a communicative function and convey human emotion, there is a belief that certain facial expressions are associated with specific emotions (Hager, 2003c). EMFACS (Emotional FACS) was developed to use the objective scoring of FACS to identify facial expression (Ekman, Friesen, Irwin, & Rosenberg, 2003). EMFACS requires the ability to identify the specific action units, their intensity, and their symmetry. Although EMFACS scoring is not done on a frame-by-frame basis—and can be done in one-tenth of the time of FACS—it does require FACS certification to know which action units are engaged (Hager, 2003b). Nevertheless, Hager, a student and employee of Dr. Paul Ekman, synthesized decades of research by describing what facial expressions correspond with human emotions (Hager, 2003c). Although there are numerous types of emotions, scientific research has shown that people can reliably assign facial expressions to seven categories of emotion (Hager, 2003c): Happy, Sad, Anger, Surprised, Fear, Disgust, and Other. When applying FACS, raters also assess the intensity of each action unit on a five-point scale: (1) Trace, (2) Slight, (3) Marked or Pronounced, (4) Severe or Extreme, and (5) Maximum (Hager, 2003d). Each intensity level possesses criteria that are present in the lower-intensity levels.

MODIFICATION OF THE CODING SYSTEM

The protocol was originally developed to be applied to a naturalistic driving study (NDS) data set that had recorded audio of the driver's voice. However, it was subsequently modified to be applied to an NDS data set that did not have recorded audio. As such, the first constraint was that the protocol would need to be applied without knowing the context of the conversation. The second constraint was that the protocol would need to be applied relatively quickly. This is because NDS data reduction often involves inspecting thousands of samples under aggressive timelines and budgets. As such, the protocol was developed to capture an overall emotion by the driver, rather than facial muscle activations on a frame-by-frame basis. The third constraint was that there was only one view of the driver's face available for reduction, and it was taken from the right side of the driver's face. Although this constraint was less severe, it may have prevented the reductionists from perceiving the symmetry of facial expressions.

Of particular interest is the level of training required to successfully and reliably rate a driver's emotional state. As previously noted, the FACS certification is intensive. That level of detail may be unnecessary to make a determination of the overall emotion of a driver in a video segment. The approach presented below will outline tests between raters of different experience levels to determine if they exhibit similar ratings. The goal is to develop a protocol that can be quickly and easily implemented with general agreement across raters. If reductionists can be given the protocol and implement it reliably and accurately with little training, the protocol will serve this goal. The following sections present an approach to generating a protocol that can be quickly implemented with little training along with an initial test of the results of implementing the protocol on a series of conversations collected in multiple NDSs.

CHAPTER 2. APPROACH

APPLYING FACS

FACS was used to guide the development of the emotion reduction protocol. Coding the specific muscle activations on a frame-by-frame basis was beyond the scope of the research projects to which the protocol was to be applied. Similar to FACS, the protocol instructed reductionists how to classify the driver's emotion into one of six categories. Certain categories, such as Fear and Disgust, were combined into the Other category to simplify the reduction. An "Unable to Determine" category was also added to allow reductionists to indicate when they were unable to classify an emotion. The description of the facial expressions was also simplified. It is recognized that, in doing so, the "Happy" category does not allow reductionists to directly indicate whether a driver exhibits a social smile (i.e., only the zygomaticus major is activated, raising the corners of the mouth), or a Duchenne smile (i.e., in addition to the zygomaticus major being activated, the orbicularis oculi are also activated, squinting the eyes). Note that a Duchenne smile is said to only be exhibited when true emotion is expressed (Ekman, Davidson, & Friesen, 1990). Nevertheless, reductionists could capture true expressions of happiness in their intensity rating. Table 1 shows the emotion reactions and definitions used in the present work. The full protocol given to reductionists is presented in Appendix A.

Table 1. Driver emotion reaction definitions.

Emotion Category	Operational Definition
Unable to Determine	Cannot tell what emotion the driver is showing e.g., poor video quality
Neutral/No Emotion Shown	The driver has a straight face, does not smile or laugh, does not gesture.
Happy	The driver smiles or laughs. The driver gestures in excitement.
Angry/Frustrated	The driver lowers/squeezes eyebrows, wrinkling forehead. The driver clenches his/her teeth. The driver yells (opens mouth wide with eyebrows lowered). The driver gestures in anger/frustration. The driver raises his/her upper lip or tightens lips.
Sad	The driver has droopy eyebrows (raises inner eyebrows, lowers outer eyebrows). The driver frowns by lowering the outer corners of his/her lips.
Surprised	The driver's eyebrows raise. The driver's mouth opens. The driver gestures.
Other	Emotion reaction that does not fit into any other category

Once reductionists rated the type of emotion expressed by a driver, they then indicated the intensity of the emotion expression. The five-point rating scale used in FACS to assess the intensity of each muscle activation was converted into a four-point scale and applied to the overall emotion. Table 2 presents the intensity levels and the operational definitions used. It is recognized that the intensity levels are described in general and not described differently for different emotion categories. Nevertheless, they serve to identify extreme expressions of emotion

behind the wheel. It is also possible that other markers of severe emotion (e.g., driver is in tears, or visibly laughing) that are not listed could have led reductionist to rate emotion as severe.

Table 2. Driver emotion intensity definitions.

Intensity of Emotion	Operational Definition
Unable to Determine	Cannot tell the intensity of the emotion
Neutral/No Emotion Shown	The driver has a straight face, does not smile or laugh, does not gesture.
Slight (Emotion Somewhat Shown)	The driver no longer has a straight face. However, no gesturing or head movement is observed.
Marked or Pronounced (Emotion Very Much Shown)	The driver no longer has a straight face. The driver gestures one time in a reserved manner. The driver moves his head one time.
Severe (Emotion Extremely Shown)	The driver has wide eyes and a wide open mouth. The driver is screaming. The driver gestures wildly, or the driver moves his head frequently.

CASE SELECTION

Cases were drawn from archival naturalistic driving data collected during previous studies (Blanco et al., in press; Fitch et al., 2013; Hanowski et al., 2008). The cases that were selected from these data sets had previously been coded using the above emotion reduction protocol. The previous coding had been conducted by data reductionists and reviewed by senior reductionists for quality assurance purposes. In the event of disagreement between the two reductionists, another senior reductionist served as a third rater. This coding provided the classification for each case to aid in sampling for the current study. A total of 72 test cases were selected from this previously coded data in order to represent a wide range of emotions and intensities. Table 3 summarizes the initial category and intensity of cases sampled. Although cases were selected to provide the widest possible range of categories and intensities, the existing data sets did not contain equal numbers of each category and intensity combination. The sampled cases were as balanced as possible given the sources.

Table 3. Number of cases per category and intensity.

Emotion Category	Intensity	Test Cases	Warm-up Cases
Neutral/No Emotion Shown	Neutral/No Emotion Shown	10	2
Happy	Slight (Emotion Somewhat Shown)	10	2
Angry/Frustrated/Impatient	Slight (Emotion Somewhat Shown)	10	3
Surprised	Slight (Emotion Somewhat Shown)	8	0
Happy	Marked or Pronounced (Emotion Very Much Shown)	10	2
Angry/Frustrated/Impatient	Marked or Pronounced (Emotion Very Much Shown)	10	1
Surprised	Marked or Pronounced (Emotion Very Much Shown)	4	0
Happy	Severe (Emotion Extremely Shown)	9	0
Angry/Frustrated/Impatient	Severe (Emotion Extremely Shown)	1	0
Total		**72**	**10**

RATERS

For the current study, each case was coded for both emotion reaction (Table 1) and intensity (Table 2) by six experienced raters. Six independent experienced ratings were therefore obtained for each case. Experienced raters were blind to the initial classification of each case. Experienced raters were defined as data reductionists who had accumulated at least 6 to 12 months of experience analyzing driver behavior on video and who had worked with earlier versions of the emotional state and intensity scales. They were retrained with the current version of the scales at the start of this study.

Cases were then rated by six novice raters. Six independent novice ratings were therefore obtained for each case. Novice raters were also blind to the initial classification of each event. A novice rater was defined as a data reductionist with less than 6 months of experience and who had never before applied emotional assessment scales to driving video. They received their initial training at the start of this study.

In both groups (experienced and novice), training was limited to asking the rater to read the protocol, become familiar with the reduction interface (video and survey-type interface), and code 10 “warm-up” cases. No feedback was provided after these 10 cases were rated unless the rater had specific questions about the protocol. The number of warm-up cases was limited to provide the largest number of test cases. The categories for each case are shown in Table 3. Within the 10 cases, one driver had 3 events, with the remaining 7 drivers having one event each.

Once the practice cases were completed, the 72 sampled cases were presented in an order that was randomized for each rater. As raters worked through the 72 randomized cases, they could make notes about observations and go back to previous events if they believed it was necessary

to provide consistent ratings. The 72 cases include samples from 53 different drivers, with a range of 1 to 4 events per driver and an average of 1.4 events per driver.

It is worth noting that the intent was to develop a protocol that utilized raters' innate ability to determine whether a driver was showing emotion. At the level of detail prescribed by the protocol, raters might classify subtle emotions as neutral. However, that limitation was deemed acceptable given that the protocol is intended to be used to identify emotions more generally, rather than the subtle distinctions between emotional states.

CHAPTER 3. RESULTS

Results are presented for the initial reliability and validity of the ratings applied by experienced and novice reductionists. Also presented are subsequent analyses to determine where inaccuracies occurred, in order to improve future implementations of the protocol. Rater confidence was collected during the reduction, but it was not included in any analyses or used to score any ratings.

RELIABILITY

Initially, reliability analyses were conducted to test reliability within each group of raters for both the emotion category and the intensity of the emotion. Both the emotional category and the intensity of the emotion were treated as nominal data for the purpose of this analysis. Fleiss' *k* (Fleiss, 1971) was calculated for both experienced and novice reductionists to determine if the ratings within each group were different from what could be expected by chance. High corrected reliability scores indicate that raters within a group are rating in a similar fashion, with a reliability score of 1 indicating perfect agreement between raters and zero indicating no agreement. Furthermore, reliability within each group serves as an initial test to determine if a validity assessment is warranted. That is to say, if reliability is less than chance then there is no reason to assume that raters have provided meaningful or accurate responses.

As shown in Table 4 and Table 5, reliability was significantly different than chance ($p < .05$) for experts and novices for both emotion type and emotion intensity. Reliability was moderate for emotion type for both experienced and novice reductionists. Reliability was lower for intensity for both experienced and novice reductionists. Experienced and novice reductionists showed similar reliability estimates, indicating little difference between the two groups. Although overall reliability was moderate, the reliability analyses provide sufficient evidence to proceed to validity analyses.

Table 4. Corrected reliability scores for experienced and novice reductionists—Emotion reaction.

Reductionist Experience Level	Kappa	S.E.	*p*
Experienced	.57	.019	<.000
Novice	.59	.019	<.000

Table 5. Corrected reliability scores for experienced and novice reductionists—Emotion intensity.

Reductionist Experience Level	Kappa	S.E.	*p*
Experienced	.23	.018	<.000
Novice	.32	.019	<.000

RATER ACCURACY

The following analyses estimate the overall validity of ratings. Typical validity analysis would test the rated values against a "gold standard" of ratings, such as patient outcomes in the case of

medical ratings. However, there is not a "gold standard" that can be used to determine the "ground truth" of the present sample, given that the initial classifications were ratings themselves, albeit a multi-rater consensus. For the current analysis, validity was estimated via accuracy, or agreement. Accuracy was estimated two ways: first by comparisons between the experienced and novice raters, and second by comparison of each group to the initial classification. The higher the percentage of agreement within the ratings for a case, the higher the likelihood that the case was rated accurately.

When making comparisons to the initial classification, agreement scores were computed both across raters and for each case. Results across raters consist of calculations for each rater for each case (i.e., six ratings for each case). This was defined as *raw agreement*; it encapsulates any time raters disagreed with the initial classification. Results computed for each case used the mode of the six raters to estimate what the overall rating would be for a given case. This was defined as *practical agreement*, as a typical video segment would be reviewed by more than one reductionists for quality assurance purposes. As was the case for the initial classification, the majority rating would likely be used to classify the case during a typical rating process. Results for both emotion reaction and emotion intensity are presented.

For comparisons between experienced and novice reductionists, there is no specific basis to compare any individual reductionist to any other individual reductionist. This means that raw agreement is not informative for comparing the two groups. For comparison between the two groups, only practical agreement scores were calculated. If both groups are rating similarly, it will be reflected in higher practical agreement. The term used for practical agreement between groups will be *inter-group agreement.*

Emotion Type Agreement

Raw agreement between raters and the initial classification was computed for both experienced and novice raters. Raw agreement was calculated across all cases for each rater group. Experienced and novice raters both showed 66% raw validity. Note that this reflects aggregate agreement and not identical agreement. Both groups showed disagreement on the same number of cases but exhibited disagreement on different cases.

Practical agreement for each case was computed for each group. Since the data collected were nominal, the mode of the raters was used for each case. Again, if the mode of the raters matched the initial rating, the case was considered to be rated accurately. Both groups showed nearly identical agreement, with 71% of cases matching the initial rating. Again, although the aggregate agreement was similar, experts and novices did not show the same agreement for all cases; they just showed disagreement on the same number of cases (21 cases did not match).

Inter-group agreement was 79%, indicating that the two groups of reductionists showed high agreement with each other.

Emotion Intensity Agreement

Overall, raters were not highly accurate in regard to intensity of emotion observed, regardless of experience level. Raw validity with experts coding for intensity was 45% across all cases. Novice validity was 47% across all cases.

For practical agreement, ratings for each case were calculated using the mode for each group of raters, in the same fashion as for the emotion coding. Experts showed 54% of cases as valid; novices showed 50% of cases as valid.

Inter-group agreement was 64%, suggesting moderate agreement between groups.

SUMMARY AND FURTHER ANALYSIS

Overall, there was little difference between experienced and novice reductionists in regard to ratings of the emotion types of the selected cases. The agreement shown can be considered good to moderate for classification ratings. Raters of both experience levels were more accurate when coding for specific emotion types rather than the intensity of the emotion. Of the 21 cases that did not match the initial classification, no specific category appears dominant (Table 6). As shown in Table 7, the majority of the discrepancies occurred for cases that had an intensity rating of "slight" identified at case selection. This suggests at least one source of confusion among raters regarding classification: they were only able to reliably determine the type of emotion when the intensity was categorized as Pronounced or Severe.

Table 6. Cases and classifications showing disagreement between experienced and novice reductionists—Type.

Emotion	Experienced Cases	Experienced Percentage	Novice Cases	Novice Percentage
Neutral	1	1.4%	3	4.2%
Happy	7	9.7%	6	8.3%
Angry	5	6.9%	5	6.9%
Surprised	8	11.1%	7	9.7%
Total	21		21	

Table 7. Cases and classifications showing disagreement between experienced and novice reductionists—Intensity.

Emotion	Experienced Cases	Experienced Percentage	Novice Cases	Novice Percentage
Neutral	1	1.4%	3	4.2%
Slight	15	20.8%	13	18.1%
Marked	3	4.2%	3	4.2%
Severe	2	2.8%	2	2.8%
Total	21		21	

Although reductionists were able to classify the emotion type satisfactorily, the classification reliability of intensity was low to moderate. Subsequent analyses were conducted on combined ratings for intensity in order to provide a better understanding of rating patterns. Table 8 and Table 9 summarize the results for the collapsed accuracy scores, with further description provided below.

Table 8. Summary of rescored agreement for experienced and novice reductionists—Type accuracy.

Source of Ratings	Experienced Raw Accuracy	Experienced Practical Accuracy	Novice Raw Accuracy	Novice Practical Accuracy	Inter-Group Agreement
Original Agreement	65.5%	70.8%	65.5%	70.8%	79.1%
Severe & Marked Subset	78.4%	85.3%	76.5%	85.3%	85.2%
Severe Only Subset	83.3%	80.0%	81.7%	80.0%	100%

Table 9. Summary of rescored agreement for experienced and novice reductionists—Intensity accuracy.

Source of Ratings	Experienced Raw Accuracy	Experienced Practical Accuracy	Novice Raw Accuracy	Novice Practical Accuracy	Inter-Group Agreement
Original Agreement	44.7%	54.2%	47.2%	50.0%	63.8%
Present or Ambiguous	73.8%	81.9%	71.3%	76.3%	83.3%
Severe & Marked Subset	63.2%	73.5%	57.8%	61.8%	56.4%
Severe Only Subset	100.0%	100.0%	81.7%	90.0%	88.7%

Given that the data show that ambiguity in lower-intensity cases is the primary source of disagreement, data were collapsed based on intensity level. The initial four categories were collapsed into two: Emotion Present (Marked and Severe cases) and Emotion Ambiguous (Neutral and Slight Cases). The reductionists' initial ratings were rescored. For a case that was initially classified as Neutral or Slight, a rating of either Neutral or Slight would be considered correct. Marked and Severe cases were also rescored such that either Marked or Severe would be scored as correct. When accuracy was recalculated on the rescored data, experienced reductionists showed practical agreement with the initial classification on 81.9% of cases (59 of 72; 74% raw agreement) and novice reductionists showed agreement on 76.3% of cases (55 of 72; 71% raw agreement). Inter-group agreement was 83.3% for the rescored intensity ratings.

Based on these results, accuracy for emotion reaction coding was recalculated for the subset of cases that were classified as Emotion Present (cases classified as Marked or Severe). Accuracy scores for cases classified as Emotion Present (34 cases total) increased to 85% for both experienced and novice reductionists (78% experienced and 77% novice raw agreement). Inter-group agreement was also 85% for this subset. Agreement for intensity in cases recoded as Emotion Present was 74% (63% raw agreement) for experienced and 62% (58% raw agreement) for novice reductionists. Inter-group agreement was 56.4%.

Agreement for emotion reaction cases classified only as Severe (10 cases) was 80% for both experienced and novice reductionists (83% and 81.7% raw agreement, respectively). Inter-group reaction agreement was 100%. Furthermore, the agreement for emotion present increased to 100% for experienced and 90% for novice reductionists with inter-group agreement at 88.7%.

CHAPTER 4. DISCUSSION

The intent of this project was to develop a protocol for reductionists to determine the emotional content of video segments sampled from an NDS that could be quickly implemented with little training. The results suggest that raters are reliable at categorizing emotions that are marked or pronounced, and that this is more reliable than the rating of emotion intensity alone. Practically, this means that with little training, reductionists are able to reliably assess the emotional content that is most likely to impact driving performance. The FACS system provides extensive training on identifying subtle changes in emotion state, and the limited training used here is likely the explanation for the disagreement for low-intensity emotion cases. Although accurate ratings for intensity of emotion were only observed in the highest intensity levels, the results are positive overall. Considering that high emotional involvement is more likely to have an impact on driving performance, the protocol should allow reductionists to reliably and accurately assess the emotional content of NDS video segments.

There was little to no difference between experienced and novice ratings in the present iteration, as evidenced by the similar inter-group agreement and similar agreement with the initial classifications. This could be due to the nature of the protocol; it was largely driven by innate ability and less by experience or training with the protocol itself. Furthermore, although experienced reductionists had involvement with earlier iterations of emotion ratings, they did not have any prior experience with the current protocol. This may have limited the differences between the two groups. Overall, inter-group agreements were higher than agreements with the initial classification. This suggests that the protocol may have improved consistency among the ratings and may improve future classifications. Still, for analyses where there did appear to be a difference between experienced and novice reductionists, experienced reductionists showed higher accuracy and reliability.

While the general results were positive, there is room for improvement. Although the ratings were significantly different from chance, the general state of agreement was moderate to good without rescoring the results. This was especially the case for raw agreement and for ratings of emotion intensity. The most accurate ratings were observed for practical agreement, which requires multiple ratings to achieve a consensus among raters. While it is likely that cases will be reviewed by more than one reductionist for quality assurance purposes, improving overall accuracy will be beneficial for future iterations of the protocol. Based on these results, future analyses should be conducted using extreme or pronounced cases. Improved agreement may be observed in future iterations of the protocol as discussed below.

LIMITATIONS AND FUTURE SUGGESTIONS

There are several important limitations to this protocol. The primary limitation is that there was no ground truth testing set; the accuracy scores were computed as best estimates based on existing ratings. While naturalistic data collection can capture the state of the driver, it does make it difficult to determine the exact presence of emotion. A future test of the protocol should use means to collect a specific “ground truth” sample that is verified. Methods could include developing a set of categories collected from a study that has captured audio inside the vehicle, which would allow for the use of conversation content to distinguish between emotion types.

The second limitation is sample size. Implementing the protocol at the present scale was deemed an initial test at development. Further analysis of future implementations is necessary to determine if the results seen here will be generalizable in future studies, as well as initial comparison to a ground truth subset to determine the validity of this method.

The level of reductionists' training could be considered a limitation. The intent was to determine the emotional content of each case; however, reductionists were not explicitly trained in the FACS system or any other similar system. While this may make the specific categorizations suspect, it should not detract from the protocol itself. The specific emotional content is not meant for interpretation; rather, the goal of this protocol was to develop a middle ground when compared to these more-intensive rating systems in order to relate emotion experienced by a driver to driver safety and distraction potential.

The results suggested that the largest source of error was for cases in which slight emotion was exhibited, regardless of the type of emotion. In order to avoid restricting options, adjusting ratings, as was done here, will continue to alleviate any ambiguity for those rating cases with low intensity. This should not be problematic, as the focus will be on higher intensity situations that can be studied for their relationship to SCE risk.

The protocol implemented here gave instructions using only text descriptions. Reductionists' ratings may become even more accurate if an image depicting a broad spectrum of specific facial features, similar to Figure 1, is added to the protocol to enhance written descriptions. While the largest source of disagreement observed here was for lower-intensity emotional content, there were some cases in which the emotions were categorized incorrectly (e.g., cases categorized as Happy were rated as Angry). Figure 1 is reprinted from Pollak & Kistler (2002), in which it was shown to improve emotion categorization for young children. Although data reductionists are a different target population, this image or a similar one may still be useful to reduce ambiguity between some facial expressions.

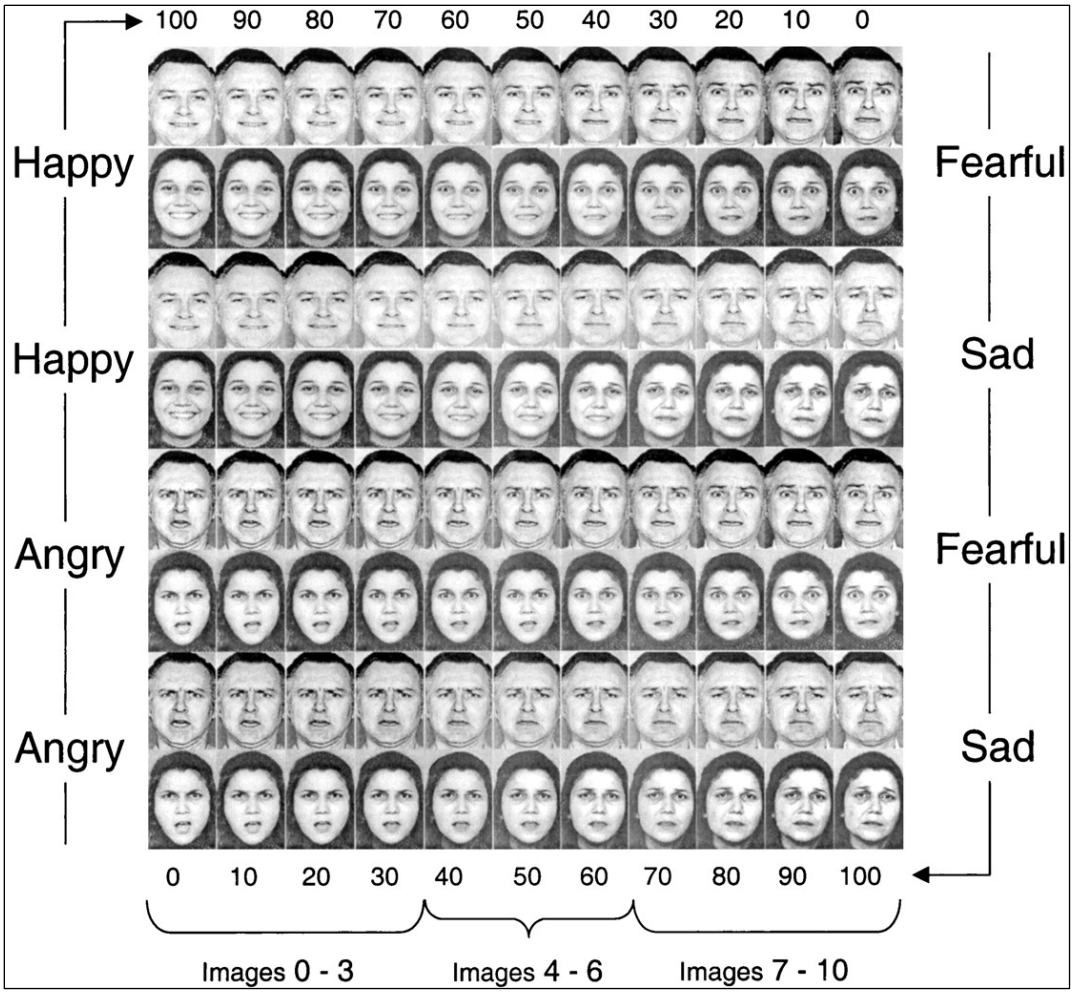


Figure 1. Chart. Examples of individuals expressing multiple emotions (Pollak & Kistler, 2002).

To conclude, the initial implementation of the protocol was considered successful and can continue to be improved in the future. The following procedures should be implemented in future iterations of the protocol:

- Reductionists should still be given the full rating scale for their initial ratings of emotion reaction and emotion intensity, and this should be used to calculate reliability between reductionists.
- Each case should be viewed by multiple reductionists, and the mode of the ratings should be used to classify the case for emotion reaction and emotion intensity.
- It may not be feasible for six reductionists to view each case in future studies. However, in the event that only two reductionists are able to view each case, cases in which those two reductionists show disagreement should be decided by a third, independent reductionist.
- Once reliability is established, categories can be collapsed into Emotion Present and Emotion Ambiguous categories as described above. These collapsed ratings should serve

as a reliable estimate to determine if extreme or pronounced emotion is present.

- Subsequent analyses comparing emotion state to SCE risk should utilize the collapsed ratings of extreme or pronounced emotion.

Future applications of the protocol are expected to yield insight into the SCE risk of cell phone conversation while driving. Furthermore, with continued refinement, the protocol has the potential to be developed for use as a generic and standardized means of classifying the emotions experienced by drivers, not only within naturalistic driving studies but also in driving studies employing other methods, including simulation.

APPENDIX A. PROTOCOL GIVEN TO REDUCTIONISTS

The Emotional Inter-rater Test questions that you will answer are below. All questions will be on the same page (page 1). Please review and utilize the definitions located in the Appendix of this document in order to better understand driver emotion and emotion intensity.

1. Emotion: **Rate the driver's emotional state during the 6 second (or 6,000 timestamp) time period from the START of the event**: Please refer to **Appendix A** for descriptions of different emotional states:
    - Neutral/No Emotion Shown
    - Happy
    - Angry/Frustrated/Impatient
    - Surprised

2. EmotionConfidence: **How confident are you in your choice for the driver's emotional state during the 6 second (or 6,000 timestamp) time period from the START of the event?**
    - 5 = Extremely Confident – you are 90%+ certain you are correct
    - 4 = Very Confident – you are 70-89% certain you are correct
    - 3 = Somewhat Confident – you are 40-69% confident you are correct
    - 2 = Slightly Confident – you are11-39% confident you are correct
    - 1 = Not Confident At All – you are 10% and below confident you are correct (thus, you are 90% confident that you are incorrect)

3. EmotionalIntensity: **Rate the intensity of the driver's emotional state during the 6 second (or 6,000 timestamp) time period from the START of the event**: Please refer to **Appendix A** for descriptions of different emotional intensities:
    - Neutral/No Emotion Shown
    - Slight (Emotion Somewhat Shown)
    - Marked or Pronounced (Emotion Very Much Shown)
    - Severe (Emotion Extremely Shown)

4. EmotionalIntensityConfidence: **How confident are you in your choice for the intensity of the driver's emotional state during the 6 second (or 6,000 timestamp) time period from the START of the event?**
    - 5 = Extremely Confident – you are 90%+ certain you are correct
    - 4 = Very Confident – you are 70-89% certain you are correct
    - 3 = Somewhat Confident – you are 40-69% confident you are correct
    - 2 = Slightly Confident – you are11-39% confident you are correct
    - 1 = Not Confident At All – you are 0-10% confident you are correct (thus, you are 90%+ confident that you are incorrect)

    Raters Comments. Please include any comments you may have or should note about the video. This is a free text response in which we would like you to provide any insightful information about the video.

Driver Emotion and Emotion Intensity Definitions

Driver Emotion Reaction Definitions

Emotion	Operational Definition
Neutral/No Emotion Shown	• The driver has a straight face, does not smile or laugh, does not gesture
Happy	• The driver smiles or laughs • The driver gestures in excitement
Angry/Frustrated	• The driver lowers/squeezes eyebrows, wrinkling forehead • The driver clenches his/her teeth • The driver yells (opens mouth wide with eyebrows lowered) • The driver gestures in anger/frustration • The driver raises his/her upper lip or tightens lips
Surprised	• The driver's eyebrows raise • The driver's mouth opens • The driver gestures

Driver Emotion Intensity Reduction Definitions

Intensity of Emotion	Operational Definition
Neutral/No Emotion Shown	• The driver has a straight face, does not smile or laugh, does not gesture • Note, will always be selected is Neutral/No Emotion is selected above
Slight (Emotion Somewhat Shown)	• The driver no longer has a straight face • However, no gesturing or head movement is observed
Marked or Pronounced (Emotion Very Much Shown)	• The driver no longer has a straight face • The driver gestures one time in a reserved manner • The driver moves his head one time
Severe (Emotion Extremely Shown)	• The driver has wide eyes and a wide open mouth • The driver is screaming • The driver gestures wildly, or the driver moves his head frequently

REFERENCES

Angell, L., Auflick, J., Austria, P. A., Kochhar, D., Tijerina, L., Biever, W., . . . Kiger, S. (2006). *Driver workload metrics task 2 final report* (Report No. DOT HS 810 635). Washington, DC: National Highway Traffic Safety Administration.

Atchley, P., & Dressel, J. (2004). Conversation limits the functional field of view. *Human Factors, 46*(4), 664-673.

Blanco, M., Hickman, J. S., Olson, R. L., Bocanegra, J. L., Hanowski, R. J., Nakata, A., . . . Bowman, D. (in press). *Investigating critical incidents, driver restart period, sleep quantity, and crash countermeasures in commercial operations using naturalistic data collection* (Contract No. DTFH61-01-C-00049, Task Order # 23). Washington, DC: Federal Motor Carrier Safety Administration.

Briggs, G. F., Hole, G. J., & Land, M. F. (2011). Emotionally involving telephone conversations lead to driver error and visual tunnelling. *Transportation Research Part F: Traffic Psychology and Behaviour, 14*(4), 313-323. doi: 10.1016/j.trf.2011.02.004

Cai, H., & Lin, Y. (2011). Modeling of operators' emotion and task performance in a virtual driving environment. *International Journal of Human-Computer Studies, 69*(9), 571-586.

Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. *Accident Analysis & Prevention, 40*(4), 1282-1293. doi: 10.1016/j.aap.2008.01.009

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

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology, 58*(2), 342-353.

Ekman, P., Friesen, W. V., Irwin, W., & Rosenberg, E. (2003). Introduction to EMFACS 2012 [Web page]. Retrieved from http://face-and-emotion.com/dataface/facs/emfacs_intro_authors.html

Fitch, G. M., & Hanowski, R. J. (2011, September 5-7). *The risk of a safety-critical event associated with mobile device use as a function of driving task demands.* Paper presented at the 2nd International Conference on Driver Distraction and Inattention, Gothenburg, Sweden.

Fitch, G. M., Soccolich, S. A., Guo, F., McClafferty, J., Fang, Y., Olson, R. L., . . . Dingus, T. A. (2013). *The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk* (Report No. DOT HS 811 757). Washington, DC: National Highway Traffic Safety Administration.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 65*(5), 4. doi:10.1037/h0031619

Grimm, M., Kroschel, K., Harris, H., Nass, C., Schuller, B., Rigoll, G., & Moosmayr, T. (2007). On the necessity and feasibility of detecting a driver's emotional state while driving. In A. R. Paiva, R. Prada, & R. Picard (Eds.), *Affective computing and intelligent interaction* (Vol. 4738, pp. 126-138). Berlin Heidelberg: Springer.

Hager, J. C. (2003a). Description of Facial Action Coding System (FACS) [Web page]. Retrieved from http://web.archive.org/web/20150303032329/http://face-and-emotion.com/dataface/facs/description.jsp

Hager, J. C. (2003b). EMFACS -- Scoring for emotion with FACS [Web page]. Retrieved from http://web.archive.org/web/20150303032329/http://face-and-emotion.com/dataface/facs/emfacs.jsp

Hager, J. C. (2003c). Emotion and facial expression [Web page]. Retrieved from http://web.archive.org/web/20150303032329/http://face-and-emotion.com/dataface/emotion/expression.jsp

Hager, J. C. (2003d). How to read the AU sections. Retrieved from http://web.archive.org/web/20150303032329/http://face-and-emotion.com/dataface/facs/manual/AU_sections.html

Hager, J. C. (2003e). NSF report - Facial expression understanding [Web page]. Retrieved from http://web.archive.org/web/20150303032329/http://www.face-and-emotion.com/dataface/nsfrept/psychology.html

Hanowski, R. J., Blanco, M., Nakata, A., Hickman, J. S., Schaudt, W. A., Fumero, M. C., . . . Madison, P. (2008). *The Drowsy Driver Warning System Field Operational Test: Data collection methods final report* (DOT HS 810 035). Washington, DC: National Highway Traffic Safety Administration.

Hickman, J. S., & Hanowski, R. J. (2012). An assessment of commercial motor vehicle driver distraction using naturalistic driving data. *Traffic Injury Prevention, 13*(6), 566-574.doi: 10.1080/15389588.2012.683841

Hickman, J. S., Hanowski, R. J., & Bocanegra, J. (2010). *Distraction in commercial trucks and buses: Assessing prevalence and risk in conjunction with crashes and near-crashes* (Report No. FMCSA-RRR-10-049). Washington, DC: Federal Motor Carrier Safety Administration.

Horrey, W. J., Lesch, M. F., & Garabet, A. (2009). Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions. *Journal of Safety Research, 40*(1), 7-12. doi:10.1016/j.jsr.2008.10.011

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(1), 196-205. doi:10.1518/001872006776412135

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). *The impact of driver inattention on near-crash/crash risk: An analysis using the 100-Car Naturalistic Driving Study data*. Washington, DC: National Highway Traffic Safety Administration.

Klauer, S. G., Guo, F., Sudweeks, J., & Dingus, T. A. (2010). *An analysis of driver inattention using a case-crossover approach on 100-Car data: Final report* (No. DOT HS 811 334). Washington, DC. National Highway Traffic Safety Administration.

Maples, W. C., DeRosier, W., Hoenes, R., Bendure, R., & Moore, S. (2008). The effects of cell phone use on peripheral vision. *Optometry - Journal of the American Optometric Association, 79*(1), 36-42. doi:10.1016/j.optm.2007.04.102

McEvoy, S. P., & Stevenson, M. R. (2009). Measuring exposure to driver distraction. In M. A. Regan, J. D. Lee, & K. L. Young (Eds.), *Driver distraction: Theory, effects, and mitigation*. Boca Raton, Florida: Taylor & Francis Group, LLC.

Mehler, B., Reimer, B., & Dusek, J. A. (2011). *MIT AgeLab Delayed Digit Recall Task (n-back)* (Working Paper 2011-3B). Cambridge, MA: Massachusetts Institute of Technology.

Movellan, J., Frank, M. S., Bartlett, M. S., & Sejnowski, T. (2013). Fully automatic face detection and expression recognition [Web page]. Retrieved from http://mplab.ucsd.edu/grants/project1/research/face-detection.html

National Highway Traffic Safety Administration. (2011). *Traffic safety facts research notes*. Washington, DC: Author.

National Highway Traffic Safety Administration. (2012, February 24). Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices, notice of proposed federal guidelines. *Federal Register 77*(37), 11200.

National Highway Traffic Safety Administration. (2013). *Traffic safety facts research notes: Distracted driving 2011* (No. DOT HS 811 737). Washington, DC: Author.

Neurauter, M., Hankey, J., Schalk, T., & Wallace, G. (2012). Outbound texting. *Transportation Research Record: Journal of the Transportation Research Board, 2321*, 23-30. doi:10.3141/2321-04

Olson, R. L., Hanowski, R. J., Hickman, J. S., & Bocanegra, J. (2009). *Driver distraction in commercial vehicle operations: Final report* (Contract DTMC75-07-D-00006, Task Order 3). Washington, DC: Federal Motor Carrier Safety Administration.

Owens, J. M., McLaughlin, S. B., & Sudweeks, J. (2010). On-road comparison of driving performance measures when using handheld and voice-control interfaces for mobile

phones and portable music players. *SAE International Journal of Passenger Cars - Mechanical Systems, 3*(1), 734-743.

Pollak, S. D., & Kistler, D. J. (2002). Early experience is associated with the development of categorical representations for facial expressions of emotion. *Proceedings of the National Academy of Sciences, 99*(13), 5. doi:10.1073/pnas.142165999

Ranney, T. A., Baldwin, G. H. S., Parmer, E., Martin, J., & Mazzae, E. N. (2011). *Distraction effects of manual number and text entry while driving* (Report No. DOT HS 811 510). Washington, DC: National Highway Traffic Safety Administration.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2012). A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. doi:10.1177/0018720812437274

Sahai, H., & Khurshid, A. (1996). *Statistics in epidemiology: Methods, techniques, and applications*. Boca Raton, Florida: CRC Press.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2013). *Measuring cognitive distraction in the automobile*. Washington, DC: AAA Foundation.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied, 9*(1), 23-32.

World Health Organization. (2009). *Global health risks: Mortality and burden of disease attributable to selected major risks*. Retrieved from http://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf