Beyond Snapshots: Validation of a Continuous Frustration Assessment in a Simulator and Real-World Setting

E. Bosch[1], S. Bohmann[1], U. Slivsek[1,2], K. Ihme[1]

[1] German Aerospace Center, Institute for Transportation Systems, Braunschweig, Germany, firstname.lastname@dlr.de

[2]University of Ljubljana, MEi:CogSci, Ljubljana, Slovenia

Abstract

Scale-based questionnaires are frequently used to assess complex psychological states such as emotions; however, these scales are often utilized for single-instance reporting and as such do not capture the complete dynamics of emotion occurrence and changes. This study aimed to compare a continuous after-study measurement of subjectively experienced frustration, to frustration ratings given on a 5-point Likert scale reported after each condition. Data was collected in a high-fidelity driving simulator and in a real-world study with an automated driving vehicle. We found that the during-study Likert-Scale ratings correlate highly with the mean after-study continuous frustration ratings in both the simulator and real-world setting. The results indicate that the after-study continuous rating is a viable alternative to the during-study Likert-scale when measuring frustration.

Introduction

Traditional emotion questionnaires employ single-instance questionnaires due to the need of multi-item questionnaires to ensure inter-item reliability [22]. However, when investigating time-resolved indicators of emotion, such as physiological data, a time-resolved subjective rating is necessary in order to research how dynamics of emotion occur and how changes in subjective experience and physiological changes are interrelated. Continuous ratings given during an experiment can disrupt the natural progression of emotions and reveal the objective of emotion induction. Therefore, one viable alternative for a continuous scubjective emotion rating is a post-hoc (post study) and single-item assessment of an emotion. One attempt of continuous emotion measurement is the affect rating dial first used by Levenson and Gottman [19]. To receive a continuous emotion rating while couples were interacting, the couples were video-taped during their interaction. Subsequently, they returned to the lab separately and provided a continuous positive-negative emotion rating post-hoc. Other studies have also used continuous post-hoc measurements by recording participants and collecting their rating afterwards [9, 16, 18, 20, 24, 27]. In this paper, we compare such a post-hoc continuous measurement to a during-study 5-point Likert scale frustration rating. Allowing participants to self-rate their emotions after the study circumvents some of the challenges associated with continuous emotion annotation as described in [21].

This study is set in the context of measuring subjectively experienced frustration in fully automated driving. Frustration is especially interesting in the context of automated driving, as the experience of frustration can inhibit the acceptance of new mobility concepts such as automated driving [10, 25]. It is, therefore, highly interesting to understand how frustration can be recognized in this context [5]. For this, traditional subjective ratings ask for participant’s emotion ratings once after every condition. However, to acquire highly time-resolved information about emotional responses to different events within an experimental condition and to connect it to possible changes in acquired sensor data, it can be helpful to obtain a time-resolved subjective rating. To see whether relationships between single-instance and post-hoc continuous frustration ratings differ depending on the context, we collected data in a high-fidelity driving simulator and a real automated driving car on a test track. Based on previous research [4], we used an in-vehicle interface to induce frustration. We then explored how well both ratings correlated. For this correlation, different metrics of the continuous frustration rating can be interesting. For example, [26] found that due to a duration neglect, the maximum and end pain ratings during a colonoscopy can be better predictors for an overall experience rating given after the procedure than the mean rating. We therefore compared the during-study Likert-scale frustration not only to mean, but also combined maximum and last-minute values of the after-study continuous rating.

Methods

Summary

Study 1 was conducted at our institute's high-fidelity driving simulator with 50 participants. Frustration was induced by interaction with an in-vehicle user interface. Subjective frustration ratings were collected after each drive on a 5-point-scale and after all drives as a continuous (i.e., highly time-resolved) rating. To test whether the results of the simulator study could be replicated in a real-world setting, we designed Study 2 as close as possible to Study 1 in a real car on a test track with 23 participants. Every participant experienced baseline and frustrating drives which were all driven on the same test track. Subjective frustration ratings were, again, collected after each drive on a 5-point-scale and as a continuous rating after all drives as manipulation check. The participants were brought to a test track before the start of the study, which took about 20 min. The participants were different from the ones in Study 1. Results of this study's camera and EEG data results are published under [6].

Participants

Fifty participants recruited through the institute's participant pool took part in Study 1. Previous studies with similar scope and settings had comparable sample sizes [13, 14, 31]. In total, nine participants were excluded from data analyses, due to motion sickness (2), data saving problems (3), and missing data (4). The n = 41 participants included in the analyses were aged 20 to 59 years (y) (M = 31.54 y, SD = 12.46 y, 12 female, 29 male). Twenty-two participants recruited through the institute's participant pool took part in Study 2. The decision to recruit twenty-two participants was based on the tradeoff of measuring as many participants as possible within a feasible time of availability of the research car and the test track. One participant had to end the experiment early (for urgent private reasons). The n = 21 participants included in the analyses were aged 23 to 58 years (y) (M = 41.71 y, SD = 10.34 y, 5 female, 16 male). As reimbursement for their time, all participants received 5 € per commenced half hour for their participation.

Set-Up

Study 1 was conducted in a driving simulator virtual reality lab with 360° full view. The participants sat in a realistic vehicle mock-up. Study 2 was conducted in our institute's test vehicle on a test track (comparable to SAE Level 4, 28). The participant sat in the driver seat and did not engage in any driving task. A security driver was present at all times on the co-driver seat with access to break and throttle. The car drove with a maximum speed of 30 km/h on a track of roughly 1.6km. In both studies, the UI was displayed on a tablet (Microsoft Surface Pro 7, 12.3') that was attached over the center console of the car.

Stimuli

The participants read a story to immerse into the setting before all drives that told them they were driving to a business meeting. Participants then solved a task on the in-car user interface displayed on the tablet. The participants were told to receive a 2 € reward upon successful completion of their task. In the baseline condition ('*Baseline*'), the participants were asked to visit a website, which could be accomplished easily. They were then asked to press the one button that appeared in different places of the UI. They were told to have no time pressure and to interact with the UI as natural as possible. In the first frustration condition (*'Frust1'*):, the participants received a call from their 'boss', who told them that they were urgently needed for another, more important meeting and needed to turn around immediately to arrive on time. The participants then had to change the destination of the navigation system. Through ambiguous naming of buttons, unclear icons, and unintuitive paths, this was hard to achieve within 7 min. In the second automation condition (*'Frust2')*, a 'boss' called and asked the participant to very urgently join an online conference with clients. Again, the UI was so difficult to understand that it was hard to reach the goal of joining the online conference within the given time.

Measures

To assess the participant's frustration levels, the participants rated their frustration in two different frustration scales. One was an emotion questionnaire that was filled in after every drive ('during-study Likert-scale frustration rating'). It first asked four distraction questions about gaze behavior in line with the cover story (see Supplementary Materials for the exact questions). Afterwards, the participants rated an emotion scale based on the German version of the positive and negative affect scale 'PANAS' [17]. It has a reliability of Raykovs ρ = 0.93 [7] and is a commonly used method to acquire participant's emotions (see, for example, 2, 12, 30). The translated emotions words used were 'active', 'distressed', 'interested', 'excited', 'upset', 'scared', 'inspired', 'proud', 'enthusiastic', 'ashamed', 'alert', 'nervous', 'determined', 'attentive', 'jittery', 'afraid' (from the original PANAS) and 'insecure', 'frustrated', 'angry', 'sad', 'surprised', 'relaxed' (our own addition) were rated on a 5-point scale from 'not at all' to 'extremely'. We decided to acquire this broad emotion spectrum to hide that we were trying to induce frustration and also added emotion words similar to frustration to test for a latent frustration construct by factor analysis.

The second frustration rating ('after-study continuous frustration rating') was obtained after all drives. For this, the participants watched the videos that were recorded during all drives of the whole scene (the participant's face was not visible) and rated their frustration with a joystick on a level from 0 to 100%. This rating was given continuously, i.e. the participant always held the joystick in the position that corresponded to their frustration level as experienced in the situation shown in the video. The joystick was moveable only in one direction and automatically returned to zero-position when not touched. The participants saw a visual feedback of their current rating, which was presented next to the video. They were asked to move the joystick according to the frustration level that they felt in the situation shown in the presented video. By this, a continuous frustration rating for each drive and each participant was collected. We decided for this continuous measure in addition to the common method of questionnaires to receive a subjective rating not only once per drive, but for every timepoint during the drive. In the Simulator study, the time between the last drive and giving the after-study continuous frustration rating was about 10 min, in the Real-world study it was about 45 min.

Procedure

All participants arrived and filled in an informed consent and a data privacy statement. Before the start of Study 1, participants were informed of potential risks of driving in simulators (e.g., the experience of simulator sickness) according to the simulator safety concept. In Study 2, Participants were brought to a test track, which took about 20min. Before the start of the study, the participants were informed about potential risks of driving in an automated vehicle on a test track with safety driver (e.g., the experience of motion sickness) according to the vehicle safety concept. The participants were informed that they could take a break or abort their participation at any time. All participants provided written informed consent to take part in the study and the video recording. The participants were told the cover story that the study investigated differences in gaze behavior between manual and automated driving modes. This was done to conceal the true aim of frustration induction and enable natural emergence of emotions. To reduce effects that came from unfamiliarity, all participants experienced automated driving scenarios before the start of the experiment until they said to be adapted to the simulator or the automated riving car, respectively. This took five minutes on average. After the all drives, the participants were informed about the true goal of the experiment (evoking frustration) and the necessity to conceal this goal with a cover story. They then gave the continuous frustration rating for all drives. The whole procedure took 2 hours on average. The collected data was handled and saved in line with the European General Data Protection Regulation. A project-internal ethics committee reviewed and approved the study.

Experimental Design

This data collection was part of a larger study as described in [5, 6]. Therefore, Participants also drove manual driving modes in Study 1. In a 2 (driving mode: automated vs. manual) x 2 (frustration induction: frustration vs. baseline) within-subject design, each participant experienced six drives in total. Three of these were driven by the participants themselves (manual driving mode) and in three the car drove automatically (automated driving mode).

Both driving modes consisted of one baseline drive and two frustration-inducing experimental drives each. The order of the drives was balanced by a balanced Latin square design for all participants, which means that every condition was driven in every position, and also the order of the drives was balanced (see for example 15). The same was true for Study 2, where only automated driving conditions existed (see [5]).

Data Analysis

As our factor-analytical approaches did neither reveal a fitting measurement model for negative affect nor for a latent frustration construct, we correlated the 'frustrated' item ratings after each condition ('during-study Likert-scale frustration rating') with the after-study continuous rating's mean of each respective condition. This was done by spearman rank correlation due to the ordinal nature of the during-study Likert-Scale. Considering a heuristic perspective to the experience of affective episodes [11, 26] we first fitted an ordinal logistic regression model (Model 1) with the predictor variable 'mean after-study continuous frustration'. We then extended this model to include a linear combination of the peak and the mean of the last minute of the continuous frustration rating as predictors in Model 2. We report both models' pseudo-r-squared values and used a likelihood ratio test to compare Model 1 and Model 2. In Model 3, we fitted an ordinal logistic regression model with the predictor variable 'Peak-end value of after-study continuous frustration' only and then compared it to Model 2 by likelihood ratio.

Results

Figure 1 shows a comparison of during- and after-study frustration ratings. Over both the simulator and real world drives, the Spearman's rank correlation coefficient of emotion scale rating per drive and mean continuous frustration rating per drive was 0.57, which is a high correlation according to Cohen [8]. The correlation was as high when only considering the simulator setting and 0.69 when only considering the real-world study (see Table 1).

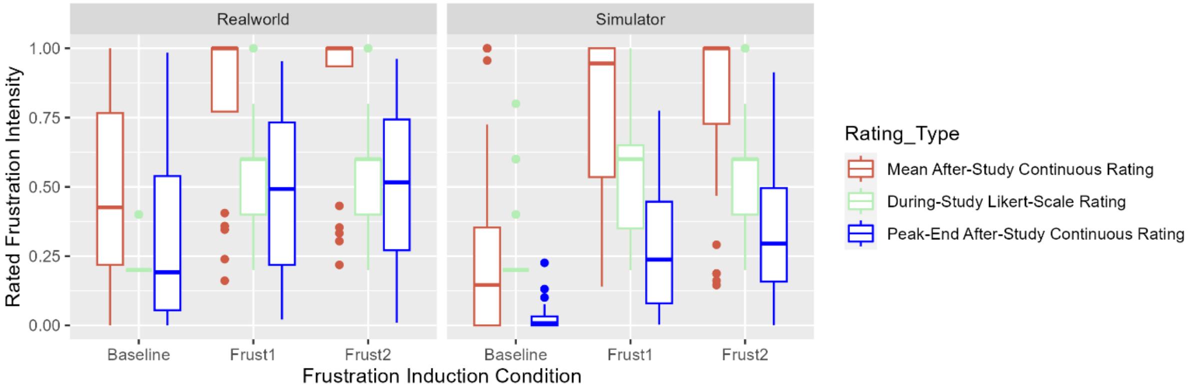

Figure 1: Comparison of during-study Likert-Scale rating and after-study continuous rating by using the mean and peak-end rating of the continuous rating. Likert Scale of Emotion Scale divided by 5 to have a comparable axis.

Table 1: Spearman's rank correlations of after-study continuous frustration rating with during-study Likert-Scale frustration ratings.

Setting	Mean of After-study continuous rating with During-study Likert-Scale rating	Peak-end value of After-study continuous rating with During-study Likert-Scale rating
Both settings	0.57	0.53
Simulator	0.57	0.54
Real-world	0.69	0.63

The ordinal logistic regression model with only the predictor variable ‘mean continuous after-study frustration’ (Model 1) yielded a 33.4% explanation of the variance in the data based on the Cragg and Uhler's pseudo R-squared [23]. Model 2 that added the peak and last-minute-mean values of the continuous after-study rating (‘Peak-end value’), yielded a pseudo R-squared value of 34.6%. A likelihood ratio test was performed to compare Model 1 and Model 2. The test statistic was $LF = 3.20$, and the p-value was $p = .07$. Since the p-value is higher than 0.05 and the increase in explained variance is minimal, we do not reject the more parsimonious Model 1. Model 3 with only the predictor variable ‘Peak-end value’ yielded a pseudo R-squared value of 21.0% and a likelihood ratio test to compare Model 2 and 3 resulted in a test statistic of $LF = 32.29$ with a $p < .001$. We therefore accept the hypothesis that a model including ‘continuous after-study frustration’ is significantly better than a model that only contains ‘peak-end value’. Overall, we therefore prefer the most parsimonious model, using only the mean of ‘continuous after-study frustration’ as predictor, as it achieves no worse goodness of fit than model with additional predictors.

Discussion

In this study, we aimed to compare subjective frustration ratings given on a 5-point Likert Scale after every drive (‘during-study Likert-scale frustration rating’) to a continuous frustration rating given after all drives (‘after-study continuous frustration rating’). We did this comparison in the setup-up of a high-fidelity driving simulator and a real-world study with an automated driving car on a test track. As a result, we found that the ratings given after every drive correlate highly with the continuous frustration rating given after all drives in both set-ups. Previous research has compared a continuous emotion rating to a partner’s emotion rating [20], to emotions expected by induction methods [18] or not compared it to another rating [16, 24, 27]. These findings suggest that a post-hoc continuous rating can be used in studies where a higher time-resolution of a subjective rating is necessary. The higher mean after frustration rating compared to the during study frustration rating might occur because the emotion of frustration became more salient for the subjects as a result of the instruction.

One disadvantage of single-item measurements compared to multi-item measurements of a latent construct is that they are more prone to measurement error and therefore have a lower reliability in many cases [1]. Considering that low reliability reduces the correlation with other variables, the rather strong correlation we found between the continuous rating and the frustration item indicate that the post-hoc rating is a viable alternative to the after-drive frustration scale. Applying heuristics that have previously been found to yield a better fit than the mean of a continuously given rating [11, 26] did not result in a meaningfully improved model fit in comparison to only taking the mean post-study continuous frustration rating as predictor for the during-study Likert-scale rating. Using only the peak-end heuristic as proposed by [26] resulted in a significantly worse model fit than using mean and the peak-end values as predictors. This indicates that the peak-end value does not improve the model fit that can be achieved by only using the mean after-study continuous rating as predictors. [26]’s heuristics do not seem to apply to the relationship between after-study continuous rating and during-study Likert-scale rating in our study.

One limitation of our study is that the Likert scale frustration after each drive was measured by a single frustration item, so that measurement error cannot be considered. Probably due to small sample size and skewed indicators, longitudinal confirmatory factor analysis models resulted either in bad model fits or estimation problems like negative error variances. On the other hand, our struggles to fit a model dovetail with reports of structural ambiguity of the German version of PANAS e.g. by [29]. Future research could induce frustration in a larger study sample, for example in an online study, and do a similar comparison of after-drive and after whole study ratings. We encourage researchers to factor analyze indicators of emotional constructs, particularly in German language. A limitation of the comparison to [26]’s heuristics of memory bias is that in our study, the continuous rating was given later than the single-item scale rating. This is opposed to [26]’s study design and might explain why adding the peak and end ratings did not improve the model fit.

Conclusion

This study set out to compare whether a continuous frustration rating given after a study yields results comparable to a 5-point Likert scale rating given after every experimental condition. Our experiments confirmed that the correlation between the two ratings is high. This suggests that memory effects that might bias the rating after all drives can be neglected in future studies and, when in need of a continuous rating, this is a viable alternative to the during-study Likert-Rating in future studies. Further research using a multidimensional during-study frustration rating and more participants is needed.

Acknowledgements

This work was partly supported by the project F-RELACS which receives funding from the German Federal Ministry of Education and Science in the program KMU-innovativ under the grant 16SV7930.

Data availability

This manuscript's data will be made publicly available after acceptance under [3].

1 References

[1] Allen, M. S., Iliescu, D., and Greiff, S. 2022. Single Item Measures in Psychological Science. *European Journal of Psychological Assessment* 38, 1, 1–5.

[2] Barańczuk, U. 2018. Emotion regulation mediates the effects of temperament traits and posttraumatic stress disorder symptoms on affect in motor vehicle accident survivors. *Transportation research part F: traffic psychology and behaviour* 58, 528–535.

[3] Bosch, E. 2023. *Continuous vs. Single-Instance Frustration Rating Dataset. DOI*=10.17605/OSF.IO/MVRJK.

[4] Bosch, E., Ihme, K., Drewitz, U., Jipp, M., and Oehl, M. 2020. Why drivers are frustrated: results from a diary study and focus groups. *European Transport Research Review* 12, 1, 52.

[5] Bosch, E., Käthner, D., Jipp, M., Drewitz, U., and Ihme, K. 2023. Fifty shades of frustration: Intra- and interindividual variances in expressing frustration. *Transportation research part F: traffic psychology and behaviour* 94, 436–452.

[6] Bosch, E., Klosterkamp, M., Guevara, A., Kaethner, D., Bendixen, A., and Ihme, K. 2022. Multimodal Estimation of Frustrative Driving Situations Using a Latent Variable Model. In *2022 13th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 11–16. DOI=10.1109/CogInfoCom55841.2022.10081636.

[7] Breyer, B. and Bluemke, M. 2016. Deutsche version der positive and negative affect schedule PANAS (GESIS panel).

[8] Cohen, J. 1988. *Statistical power analysis for the behavioral sciences, 2nd edn. Á/L*. Erbaum Press, Hillsdale, NJ, USA.

[9] Cowie, R., McKeown, G., and Douglas-Cowie, E. 2012. Tracing Emotion. *International Journal of Synthetic Emotions* 3, 1, 1–17.

[10] Ferreri, N. R. and Mayhorn, C. B. 2022. Identifying and understanding individual differences in frustration with technology. *Theoretical Issues in Ergonomics Science*, 1–19.

[11] Fredrickson, B. L. and Kahneman, D. 1993. Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology* 65, 1, 45–55.

[12] Frison, A.-K., Wintersberger, P., and Riener, A. 2019. Resurrecting the ghost in the shell: A need-centered development approach for optimizing user experience in highly automated vehicles. *Transportation research part F: traffic psychology and behaviour* 65, 439–456.

[13] Hoque, M. and Picard, R. W. 2011. Acted vs. natural frustration and delight: Many people smile in natural frustration.

[14] Ihme, K., Unni, A., Zhang, M., Rieger, J. W., and Jipp, M. 2018. Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy. *Frontiers in human neuroscience* 12, 327.

[15] Kim, B. G. and Stein, H. H. 2009. A spreadsheet program for making a balanced Latin square design. *Revista Colombiana de Ciencias Pecuarias* 22, 4, 591–596.

[16] Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., and Zafeiriou, S. 2019. Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *Int J Comput Vis* 127, 6-7, 907–929.

[17] Krohne, H. W., Egloff, B., Kohlmann, C.-W., and Tausch, A. 1996. Untersuchungen mit einer deutschen Version der" Positive and negative Affect Schedule"(PANAS). *Diagnostica-Gottingen-* 42, 139–156.

[18] Laurans, G., Desmet, P. M. A., and Hekkert, P. 2009. The emotion slider: A self-report device for the continuous measurement of emotion. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–6. DOI=10.1109/ACII.2009.5349539.

[19] Levenson, R. W. and Gottman, J. M. 1983. Marital interaction: physiological linkage and affective exchange. *Journal of personality and social psychology* 45, 3, 587–597.

[20] Levenson, R. W. and Ruef, A. M. 1992. Empathy: A physiological substrate. *Journal of personality and social psychology* 63, 2, 234–246.

[21] Metallinou, A. and Narayanan, S. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8. DOI=10.1109/FG.2013.6553804.

[22] Moosbrugger, H. and Kelava, A. 2020. *Test theory and questionnaire construction*. Berlin: Springer.

[23] Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. *biometrika* 78, 3, 691–692.

[24] Ong, D. C., Wu, Z., Tan, Z.-X., Reddan, M., Kahhale, I., Mattek, A., and Zaki, J. 2019. Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset. *IEEE transactions on affective computing* 12, 3, 579–594.

[25] Partala, T. and Saari, T. 2015. Understanding the most influential user experiences in successful and unsuccessful technology adoptions. *Computers in Human Behavior* 53, 381–395.

[26] Redelmeier, D. A. and Kahneman, D. 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 66, 1, 3–8.

[27] Ruef, A. M. and Levenson, R. W. 2007. Continuous measurement of emotion. *Handbook of emotion elicitation and assessment*, 286–297.

[28] SAE International. 2014. *Automated Driving Levels of Driving Automation are Defined in New SAE International Standard J3016*. SAE International Troy, MI.

[29] Seib-Pfeifer, L.-E., Pugnaghi, G., Beauducel, A., and Leue, A. 2017. On the replication of factor structures of the Positive and Negative Affect Schedule (PANAS). *Personality and Individual Differences* 107, 201–207.

[30] Zhang, M., Ihme, K., and Drewitz, U. 2019. Discriminating drivers' emotions through the dimension of power: evidence from facial infrared thermography and peripheral physiological measurements. *Transportation research part F: traffic psychology and behaviour* 63, 135–143.

[31] Zhang, M., Ihme, K., Drewitz, U., and Jipp, M. 2021. Understanding the Multidimensional and Dynamic Nature of Facial Expressions Based on Indicators for Appraisal Components as Basis for Measuring Drivers' Fear. *Frontiers in Psychology* 12, 622433.