Limbrecht, Kerstin, Steffen Walter, Stefanie Rukavina& Harald C. Traue (2012) On the test-retest reliability of facial emotion recognition with the „Pictures of Facial Affect". GSTF Journal of Law and Social Sciences (JLSS). *1(2)* 19-22

On the test-retest reliability of facial emotion recognition with the „Pictures of Facial Affect Ulm"

Kerstin Limbrecht, Stefanie, Rukavina, Steffen Walter & Harald C. Traue

***Abstract*—As the abilities in facial emotion recognition continue to provide relevant clinical information as well as organizational decision makers the need for valid and reliable instruments to assess the success of standardized systems for measuring facial emotion recognition is increasingly important.**

In this study 59 participants underwent two administrations of the "Pictures of Facial Affect-Ulm" for measuring facial emotion recognition abilities with an average time interval of 7.06 days (SD = 2.91). Results revealed that PFA-U provides high test retest-reliability that can be used for defining pre- and post-intervention scores in a wide range of application areas (e.g. performance in facial emotion recognition for patients with neurological or mental disorders in therapeutic setting).

***Index Terms*—facial emotion recognition, test retest-reliability, pre- and post-intervention scores**

I. Introduction

IN many areas of emotion research, images of emotional facial expressions are used to stimulate emotions, test emotion recognition abilities, or define and describe emotion categories. When using facial emotional stimuli for experimental studies, the quality of the results is directly related to the validity and reliability of the emotion stimuli used. Although facial emotion recognition seems to belong to the general abilities of human beings, individual differences can be detected. For some factors, e.g. age, gender, personality or status general differences in emotion recognition can be observed [1-5]. Some factors actually influence social and professional life quite heavily. E.g. [6] could find a correlation between extraversion, transformational leadership and emotion recognition abilities for experts and management. [7] detected advantages in emotion recognition for people of lower status. Patients with mental or neurological disorders like anxiety disorder, depression or apoplexy show specific problems in some areas of emotion recognition, which often cause particular problems in social interaction [8-11]. These before mentioned descriptions clarify the relevance of emotion recognition in daily life. Tests for proving facial emotion recognition could not only help to detect deficits, but could also be used for training purposes or evaluation. It is imaginable that picture material can be used to train appropriate reactions in an interaction. To ensure these intentions, validity and reliability of the picture material has to be guaranteed.

One of the most prominent features to guarantee a well-established test with excellent psychometric qualities is test retest-reliability. Test retest-reliability or repeatability concerns the variation in measurements taken by the same instrument and population under the same conditions. Differences in values can be monitored, e.g. changes in facial emotion recognition. Often a "critical difference" is defined. Values that are smaller than this critical difference can be ascribed by the success of a certain (clinical) treatment. Training procedures are e.g. used in the field of autism spectrum disorders. Autism is defined by impairments in social

Manuscript submitted May 17th, 2012.

This research was supported by grants from the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

Corresponding author: Kerstin Limbrecht, University of Ulm, Medical Psychology, Frauensteige 6, 89075 Ulm, Germany. Phone: (0049)731/500-61921. E-mail: kerstin.limbrecht@uni-ulm.de

interaction, impairments in communication, and restricted interests and repetitive behavior [12]. Studies show (e.g. [13]) that adults with autism are seemingly able to compensate for these deficits by using effortful cognitive strategies based on learned associations and prototypical references to label emotional expressions. It is quite possible that this fact can be used for specific training procedures enabling people with autism to assign different emotions in their interaction partner more easily. This may lead to more successful social interactions and improve quality of life for autistic patients.

This study reports on the test retest-reliability for the PFA-U – a standardized and picture set for measuring facial emotion recognition abilities. PFA-U was used twice in a non-treatment setting to detect possible differences in performance. It was intended to achieve an accordance of at least 70%.

II. Material and Methods

A. General Methods

A new set of FACS-based pictures for the six emotions (happiness, sadness, disgust, anger, surprise, and fear) called "Pictures of Facial Affect Ulm" (PFA-U) was created and tested in advance [14]. Therefore, a total of 2,810 pictures of 48 people out of five different perspectives were taken under standardized conditions (see Fig. 1). Frontal pictures were evaluated first in order to select the expression with the highest recognition rate for each person. Recognition results varied between 71% for fear and 99% for happiness, but were quite homogeneous within the distinct emotional categories. 96 pictures (two emotional pictures per person) with recognition rates above 55% were chosen for further analysis and final selection. For the final image set consisting of 48 people showing one emotional facial expression, naturalness, attractiveness and intensity were concerned to create a highly standardized picture for documenting the ability of facial emotion recognition[15-16]. Pictures can be embedded in the FEEL test software [17] to ensure a standardized testing of human emotion recognition ability. This new picture set can be used for future studies in terms of emotion management, and especially emotion recognition and stimulation. The additionally available perspectives offer the possibility for a more detailed analysis of recognition processes.

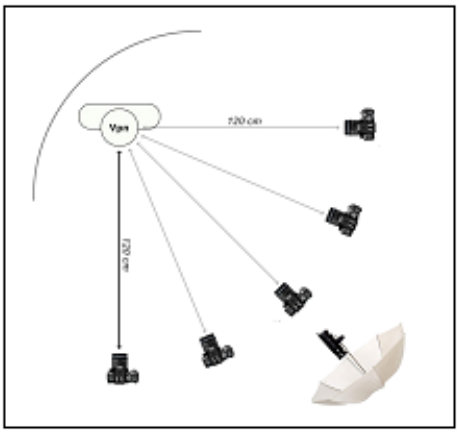

Fig. 1. Setup for stimulus development out of five perspectives

B. Participants

A total of 59 psychology students (average age = 23.55; SD = 4.38) participated in this study. All participants confirmed in writing that their participation was voluntary in accordance with the ethical guidelines (ethics commission 245/08_UBB/se). 10.45% of the sample was male. Participants received a credit of one hour's participation.

C. Hypotheses

It was hypothesized that facial emotion recognition can be defined as a basic human ability. To ensure that pre- and posttest scores measure influences of treatment solely, accordance rates between test 1 and test 2 should be eventually high. Backwards this means that no significant differences in emotion recognition performance were expected for test 1 and test 2.

D. Procedure

Participants were invited to do the test online. Therefore environment factors very carefully introduced to maximize standardization. Students were asked to use the same computer for both tests. Subjects looked at the 48 portrait pictures in the following manner: in a randomized order every frontal emotional picture was displayed on a computer screen. Picture material was scaled, that means height was constant for all participants apart from monitor size. Subjects were asked to rate the same 48 emotional pictures at two different points in time (average time interval = 7.05 days; SD = 2.91). Each participant created an individual code to a) ensure anonymity and b) allow for attribution for the two testing times. For the second testing it was not communicated to the participants that they have seen the same picture material in advance.

III. Results

To assess the reliability of this picture set two methods were intended to use: paired Wilcoxon-tests (due to the missing normal distribution) and correlations. Wilcoxon-tests were conducted to determine the difference in mean responses on emotional level for measurement 1 and measurement 2.

Therefore average emotion recognition rates were calculated (see table 1) on the basis of correct labeling. Happiness was recognized best in both test situations, fear was recognized worse. The sequence in correct emotion labeling was identical for both testing times.

Tab. 1. Performance in emotion recognition task for measurement 1 and measurement 2 (in percent)

	Measurement 1	Measurement 2
Fear	79.79	82.47
Sadness	84.31	86.34
Disgust	89.44	90.79
Surprise	93.21	95.81
Anger	94.54	96.34
Happiness	99.57	98.35
Average recognition rate	90.14	91.68

Wilcoxon-tests revealed no significant differences between the two testing times for fear ($Z = -1.159$; $p = .246$), anger ($Z = -1.249$; $p = .212$), disgust ($Z = -0.300$; $p = .764$), happiness ($Z = -1.869$; $p = .062$), sadness ($Z = -0.461$; $p = .645$), and surprise ($Z = -1.1550$; $p = .121$) as well as for the average recognition rate ($Z = -1.415$; $p = .157$).

The correlation coefficient due to Spearman between the two individual performances was unfortunately not sufficient for this context. The explanatory power of this measurement was constricted by varying basic populations. As subjects could attend online, some pictures (less than 5%) were not displayed correctly. Therefore accordance between the two testing times would have been underestimated by calculating correlations. For example, the correlation coefficient for happiness (detected nearly up to 100% for both measurements) was only $r = -.055$. As it was intended to include only "real pairs" – means stimuli that were correctly displayed for both testing times, correct accordance between measurement 1 and 2 was calculated individually for each participant. Incorrect accordance is defined as the decision for the same incorrect emotional label for both testing times and was not considered here. The number of correct and consistent responses was added and divided by the number of stimulus pairs (means the same stimulus rated two times). Accordance rates are: 73.24% for fear, 82.09% for sadness, 84.98% for disgust, 91.33% for anger, 92.01% for surprise, and 98.13% for happiness. As a good test retest-reliability is defined by correlations of at least $r = .7$ [18], percentage agreements of at least 70% are highly sufficient.

IV. Discussion

Three potential problems have to be concerned when applying the test retest-reliability: recall, time, and reactivity [19]. A recall problem can potentially arise when participants are tested within a too short time interval. Subjects may recall their responses out of memory and decide based on this recall. This would influence consistency of the instrument. Similarly, a time problem can arise if the subjects are administered the (emotion recognition) task within too long time interval. Differences in responses may be due to intrinsic changes and not inconsistencies in the instrument. At last, reactivity may arise when participants are asked to repeat the test for several times. Subjects get sensitized for the test and learn to respond as presumably expected. In this study only a short time interval was used to test conservatively possible learning effects. If significant improvements in emotion recognition performance had occurred these would have been attributed to recall effects. But results of this study indicate a) no significant differences and b) high agreement rates for both testing times. The PFA-U was found to have highly sufficient test retest-reliability rates.

The current study makes an important contribution in that it provides information about the reliability of a standardized picture set for measuring emotion recognition performance. So far, this was not reported for other picture sets used in science. The results suggest some clinical implications. PFA-U can be used to detect improvements in facial emotion recognition in therapeutic settings. For example, training effects for patients with autism spectrum disorders can be tracked.

One limitation of the study is that divergent basic populations occurred due to the adoption of an online testing method. In future studies it should be assured that correlation coefficients can be produced. Studies with longer time intervals and more testing times would complete the first impression of test retest-reliability in facial emotion recognition presented here. Nonetheless, the results found in this study demonstrate a high degree of test retest-reliability.

V. Conclusion

It was intended to estimate test retest-reliability for a standardized picture set for analysis facial emotion recognition performance – the "Pictures of Facial Affect-Ulm" (PFA-U). In this study high accordance rates could be detected between the two measurements representing a high test retest-reliability. PFA-U can be used for monitoring progresses in e.g. clinical or therapeutic setting in which facial emotion recognition performance is intended to be improved.

References

[1] J. S. Anastasi, and M. G. Rhodes, "Evidence for an own-age bias in face recognition," *North American Journal of Psychology,* vol. 8, no. 2, pp. 237-252, 2006.

[2] M. Batty, and M. J. Taylor, "Early processing of the six basic facial emotional expressions," *Brain Res Cogn Brain Res,* vol. 17, no. 3, pp. 613-20, Oct, 2003.

[3] E. B. McClure, "A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents," *Psychol Bull,* vol. 126, no. 3, pp. 424-53, May, 2000.

[4] S. W. Chan, G. M. Goodwin, and C. J. Harmer, "Highly neurotic never-depressed students have negative biases in information processing," *Psychol Med,* vol. 37, no. 9, pp. 1281-91, Sep, 2007.

[5] M. W. Kraus, and D. Keltner, "Signs of socioeconomic status: a thin-slicing approach," *Psychol Sci,* vol. 20, no. 1, pp. 99-106, Jan, 2009.

[6] R. S. Rubin, D. C. Munz, and W. H. Bommer, "Leading from within: The effects of emotion recognition and personality on transformational leadership behavior," *Acad Manage J*, vol. 48, pp. 845-858, 2005.

[7] M. W. Kraus, S. Cote, and D. Keltner, "Social class, contextualism, and empathic accuracy," *Psychol Sci*, vol. 21, no. 11, pp. 1716-23, Nov, 2010.

[8] A. L. Bouhuys, E. Geerts, and P. P. Mersch, "Relationship between perception of facial emotions and anxiety in clinical depression: does anxiety-related perception predict persistence of depression?," *J Affect Disord*, vol. 43, no. 3, pp. 213-23, May, 1997.

[9] G. Sachs, D. Steger-Wuchse, I. Kryspin-Exner *et al.*, "Facial recognition deficits and cognition in schizophrenia," *Schizophr Res*, vol. 68, no. 1, pp. 27-35, May 1, 2004.

[10] M. Braun, H. C. Traue, S. Frisch *et al.*, "Emotion recognition in stroke patients with left and right hemispheric lesion: results with a new instrument-the FEEL Test," *Brain Cogn*, vol. 58, no. 2, pp. 193-201, Jul, 2005.

[11] H. Kessler, M. Schwarze, S. Filipic *et al.*, "Alexithymia and facial emotion recognition in patients with eating disorders," *Int J Eat Disord*, vol. 39, no. 3, pp. 245-51, Apr, 2006.

[12] G. B. Hall, H. Szechtman, and C. Nahmias, "Enhanced salience and emotion recognition in Autism: a PET study," *Am J Psychiatry*, vol. 160, no. 8, pp. 1439-41, Aug, 2003.

[13] H. Saß, H.-U. Wittchen, M. Zaudig *et al.*, *DSM-IV-TR – Diagnostisches und Statistisches Manual Psychischer Störungen – Textrevision*, Göttingen: Hogrefe, 2003.

[14] K. Limbrecht, H. Hoffmann, S. Walter *et al.*, "Pictures of Facial Affect-Ulm (PFA-U): A new FACS-based set of pictures for basic emotions."

[15] K. Limbrecht, S. Rukavina, A. Scheck *et al.*, "The influence of naturalness, attractiveness and intensity on facial emotion recognition," *Psychology Research*, 2012.

[16] K. Limbrecht, S. Rukavina, A. Scheck *et al.*, "The role of naturalness, attractiveness and intensity in facial emotion recognition.," in Annual International Conference on Cognitive and Behavioral Psychology (CBP 2012), Singapore, 2012, pp. 91-96.

[17] H. Kessler, P. Bayerl, R. M. Deighton *et al.*, "Facially expressed emotion labeling (FEEL): PC-gestützer Test zur Emotionserkennung.," *Verhaltenstherapie und Verhaltensmedizin*, vol. 23, no. 3, pp. 297-306, 2002.

[18] M. Amelang, and W. Zielinski, *Psychologische Diagnostik und Intervention*, Berling: Springer, 1994.

[19] J. C. Nunnally, *Psychometric Theory*, New York: McGraw-Hill, 1978.

Kerstin Limbrecht is a researcher at the University of Ulm. After her diploma in psychology in 2003 she started to work at the Medical Psychology section in 2009. Her research interests include facial emotion recognition and human-computer interaction. She is currently working on her PhD. Kerstin Limbrecht is member of ISRE.

Stefanie Rukavina is a researcher at the University of Ulm. After her diploma in biology in 2009 she started to work at the Medical Psychology section. Her research interests include gender differences and human-computer interaction. She is currently working on her PhD.

Dr. Steffen Walter is a scientist at the Emotion Laboratory of the University of Ulm, Germany. He studied at the University of Constance with the specialization of neuropsychology. His dissertation was in the field of psychotherapy process research at the University of Ulm titled: "Therapeutic conversation and emotion/abstraction patterns of patients". His research focus is affective computing and companion technology. In this regard, he focuses on trans-situational experiments. A further area of research is automatic multimodal pain recognition.

Prof. Dr. Harald C. Traue is a professor for Medical Psychology at the University of Ulm.