PSYCHOPHYSIOLOGY

Psychophysiology, 54 (2017), 12–23. Wiley Periodicals, Inc. Printed in the USA.
Copyright © 2016 Society for Psychophysiological Research
DOI: 10.1111/psyp.12676

Reliability of surface facial electromyography

URSULA HESS,[a] RUBEN ARSLAN,[b] HEIDI MAUERSBERGER,[a] CHRISTOPHE BLAISON,[a] MICHAEL DUFNER,[c] JAAP J. A. DENISSEN,[d] AND MATTHIAS ZIEGLER[a]

[a]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany
[b]Department of Psychology, Georg August Universität Göttingen, Göttingen, Germany
[c]Department of Psychology, Universität Leipzig, Leipzig, Germany
[d]Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands

Abstract

Data from two studies were used to estimate the reliability of facial EMG when used to index facial mimicry (Study 1) or affective reactions to pictorial stimuli (Study 2). Results for individual muscle sites varied between muscles and depending on data treatment. For difference scores, acceptable internal consistencies were found only for corrugator supercilii, and test-retest reliabilities were low. For contrast measures describing patterns of reactions to stimuli, such as high zygomaticus major combined with low corrugator supercilii, acceptable internal consistencies were found for facial reactions to smiling faces and positive affective reactions to affiliative images (Study 2). Facial reactions to negative emotions (Study 1) and facial reactions to power and somewhat less to achievement imagery (Study 2) showed unsatisfactory internal consistencies. For contrast measures, good temporal stability over 24 months (Study 1) and 15 months (Study 2), respectively, was obtained. In Study 1, the effect of method factors such as mode of presentation was more reliable than the emotion effect. Overall, people's facial reactions to affective stimuli seem to be influenced by a variety of factors other than the emotion-eliciting element per se, which resulted in biased internal consistency estimates. However, the influence of these factors in turn seemed to be stable over time.

Descriptors: Facial EMG, Reliability

Electromyography (EMG) is a measure of the electrical activity that is generated during muscle contraction, which is directly related to the force produced by the muscle (Lawrence & DeLuca, 1983). Specifically, striated muscles consist of groups of bundles composed of individual muscle fibers. EMG records the changes in electrical potential that result from the conduction of action potentials along these muscle fibers.

Surface facial EMG has been used for the assessment of affective states in a large number of contexts. The use of facial EMG for this purpose can be traced to early research by Schwartz and colleagues (Schwartz, Fair, Salt, Mandel, & Klerman, 1976) who used it to show that nondepressed individuals more consistently react with facial expressions to imagery and specifically show more consistent happy expressions during happiness imagery. Starting with an article by Cacioppo, Petty, Losch, and Kim (1986), who asserted that "facial EMG activity differentiated both valence and intensity of the affective reaction," facial EMG has become a generally accepted index of affective reactions to a variety of visual (e.g., Davis, Rahman, Smith, & Burns, 1995; Larsen, Norris, & Cacioppo, 2003), auditory (e.g., Dimberg, 1990), gustatory (e.g., S. Hu et al., 1999), and olfactory (e.g., Jäncke & Kaufmann, 1994) emotional stimuli. It has been employed to assess reactions to emotional faces (e.g., Dimberg, 1982; Dimberg & Ohman, 1996), human (e.g., Hess & Bourgeois, 2010) or virtual (e.g., Mojzisch et al., 2006) interaction partners, nicotine (e.g., Robinson, Cinciripini, Carter, Lam, & Wetter, 2007), and other drugs (e.g., Newton, Khalsa-Denison, & Gawin, 1997). Furthermore, it has been used as an index of attitudes toward others (e.g., Brown, Bradley, & Lang, 2006; Dambrun, Després, & Guimond, 2003) and oneself (e.g., Buck, Hillman, Evans, & Janelle, 2004) in adults as well as in children (e.g., Armstrong, Hutchinson, Laing, & Jinks, 2007) using supra- as well as subliminal stimuli (e.g., Arndt, Allen, & Greenberg, 2001). For certain questions, facial EMG measures of affect have been found to be more effective and revealing than self-report measures, making this method especially attractive (e.g., Dufner, Arslan, Hagemeyer, Schönbrodt, & Denissen, 2015; Hazlett & Hazlett, 1999; Vanman, Paul, Ito, & Miller, 1997). As such, facial EMG is a widely used tool.

The present article focuses on the use of EMG to assess facial activity in psychology from a psychometric perspective. In particular, we will focus exclusively on issues of reliability. As such, readers who are interested in basic methodological aspects of this procedure such as suitable electrode dimensions, interelectrode distances, crosstalk, and the like are referred to relevant articles and chapters that cover these aspects (e.g., Fridlund & Cacioppo, 1986; Hess, 2009; Tassinary, Hess, & Carcoba, 2012). Likewise, the

This research was supported by a grant from the German Research Foundation allocated to JJAD (DE-1662/2-1) and a grant from the PPP Program of the DAAD (50774769) to UH.

Address correspondence to: Ursula Hess or Matthias Ziegler, Department of Psychology, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany. E-mail: ursula.hess@hu-berlin.de

theoretical implications of the data used here are of subordinate importance because the focus is on different methods of reliability estimation.

In many of the studies cited above—starting with the classic study by Schwarz and colleagues (1976)—there is an implicit assumption that the measured reactions are typical for the participant in that the same participant would react the same way on a different occasion or in real life at least within a reasonable time span. That is, the implicit assumption is that facial EMG reactions to affective stimuli have adequate test-retest reliability.

Further, in a typical psychological experiment, participants will react to a series of stimuli of the same type, such as emotional pictures, videos, or sounds of a specific valence. Accordingly, reactions to different stimuli of the same type—for example, reactions to disgust-eliciting images—are typically averaged for each muscle site prior to data analysis (see Larsen et al., 2003). A variant are studies that use facial EMG to assess reactions to facial expressions, for example, the study of facial mimicry, the imitation of the facial behavior of others (see Hess & Fischer, 2013). In this case, it may be of interest to assess not the reactions of individual muscles but rather a contrast that describes a pattern of reactions corresponding to the imitated expressions. For anger imitation, this may be a contrast between corrugator supercilii reactions (which are activated in frowns and hence may be expected to increase during anger imitation) and zygomaticus major or orbicularis oculi reactions (which are deactivated in frowns and hence may be expected to decrease during anger imitation). In this case, the EMG data for items of the same class—for example, reactions to anger faces—will be averaged prior to data analysis (e.g., Hess & Blairy, 2001). Yet, the basic assumption underlying the calculation of such an average is that the individual items share considerable construct variance, that is, that they are internally consistent.

Interestingly, guidelines (Fridlund & Cacioppo, 1986), handbook articles (Tassinary & Cacioppo, 2000; Tassinary et al., 2012), or validation studies (Larsen et al., 2003) have focused on technical issues or aspects of the underlying neuroanatomy to evaluate the usefulness of facial EMG, but did not consider the reliability of this approach. This is the goal of the present research. Specifically, we reanalyzed data from two studies, one measuring emotional mimicry (Mauersberger, Blaison, Kafetsios, Kessler, & Hess, 2015) and one measuring affective reactions to pictorial social stimuli (Dufner et al., 2015), to assess both internal consistency and test-retest reliability for these measures. Before describing these studies in detail, some relevant issues regarding the concept of reliability in this context need to be discussed.

Reliability

There is a broad range of definitions of reliability. Already in 1947, Cronbach (1947, p. 1) stated: "The literature of testing contains many discussions of test reliability. Each year, new formulations are offered, and new procedures for estimating reliability are championed. There appears to have developed no universally accepted procedure . . ."

Cronbach's own alpha coefficient (Cronbach, 1951) is a widely used reliability estimator that seemed to overcome the issues of defining and estimating test score reliability. However, there is also ample critique regarding Cronbach's alpha (e.g., Ziegler, Kemper, & Kruyen, 2014; Ziegler, Poropat, & Mell, 2014; Zinbarg, Revelle, Yovel, & Li, 2005), and Cronbach himself in later years expressed severe concerns regarding his coefficient as the sole estimate of test score reliability (Cronbach & Shavelson, 2004).

Traditionally, four conceptualizations of reliability have been distinguished (Cronbach, 1947): (1) coefficient of stability, (2) coefficient of stability and equivalence, (3) coefficient of equivalence, and (4) theoretical self-correlation. Especially equivalence and stability, called internal consistency and test-retest reliability today, have become the methods of choice during the last decades. Both methods operationalize measurement error, and thus the source of inconsistency, in different ways. In the following, the core ideas behind these two reliability estimates with a focus on potential problems for scores derived from EMG will be outlined.

Different Estimators, Different Problems

Internal consistency (Cronbach's alpha). Cronbach's alpha is probably the most widely used reliability estimator. However, some very strict assumptions have to be met for the estimate to be accurate. In particular, for alpha to be a true estimate of the systematic variance, the items need to be tau-equivalent (Osburn, 2000). This means that all items measure the same construct, and the relationship between item and construct (e.g., factor loading) is equivalent across items. Especially the first prerequisite is critical because it assumes that items are unidimensional.

With regard to typical EMG data, the assumption of tau-equivalence is problematic. In most experiments using EMG, the items used are supposed to elicit a specific muscle reaction as an index of an affective reaction (e.g., zygomaticus major activity in response to positive stimuli or in response to a happy face). This can be regarded as the manifestation of the construct intended to be measured. Cronbach's alpha assumes, however, that this is the only source of variation in all items except for random measurement error. However, many paradigms use items that systematically vary an additional methodological aspect (the pleasant stimuli may show a mix of erotic content, cute animals, or beautiful vistas; the faces may be male or female, etc.). This also has a systematic impact on the reaction and thus adds variance that is not due to the construct to be measured (the affective reaction) but to the nonemotional variation of the stimuli. These variance sources will increase the shared variance between those items that also share the same method manipulation. Moreover, the different method sources might have systematic relationships as well, which will affect the correlations between items with different method manipulations (e.g., stimuli with an erotic content versus cute animals for heterosexual female versus male participants). The formula for alpha is blind to the source of the variation; that is, all types of shared variance will be treated alike. Thereby, alpha can be distorted, and, depending on the size of the method effects and the direction of their interrelationships, alpha can be too large or too small. This means that, for typical EMG experiments, Cronbach's alpha is unlikely to be an appropriate reliability estimate.

Test-retest reliability. Another approach to estimating reliability is what Cronbach (1947) called stability. The idea is that a reliable measure should yield scores that have the same ranking when administered twice. Theoretically, this is a very straightforward way to estimate reliability. Moreover, it has been shown that test-retest reliabilities are more important than internal consistencies when it comes to a test score's test-criterion correlation (McCrae, Kurtz, Yamagata, & Terracciano, 2011). However, test-retest reliability estimates also come with a price. One issue regards the appropriate time between measurements. Emotional reactions do not necessarily have to be consistent across situations or stable across time. For example, participants' mood can influence their reactions

to facial expressions (Moody, McIntosh, Mann, & Weisser, 2007). In general, according to Forgas's (1995) "affect infusion model," perceivers' information-processing strategies are affected by their mood leading potentially to different perceptions of the same expressive stimuli (cf. Hess & Hareli, 2015). Studies using the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) suggest both stability and variability of mood. Mood, however, is only one of many factors that may intervene.

Another issue regards the way that measurement error is operationalized in test-retest reliability. If the measured trait is indeed stable across time, all changes within the rank orders between the two measurement points are considered to reflect error. Such changes can indeed be the result of random error and thus measurement error. However, such changes can also be the result of differential practice or repetition effects. Such effects do not have to be the same for all participants. Thus, when someone sees the same emotion-eliciting stimulus (e.g., a picture of a cute cat) for a second time, they may conceivably react with less but also with more intensity. This would change the rank order but is not a random error. Yet, if the time gap is long enough and the stimuli not too salient, participants may well react as if they saw the stimulus for the first time.

Omega w (Ω_w). A different way of conceptualizing reliability was proposed by McDonald (1999) and further developed by Hancock and Mueller (2001). This approach uses results from structural equation modeling and allows the estimation of the reliability of every latent source causing variance in items. The estimator, also referred to as weighted McDonald's omega or construct reliability, has been widely advocated as an alternative to Cronbach's alpha (e.g., Ziegler & Brunner, 2016; Zinbarg et al., 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006). The idea behind Ω_w is to use the loadings from a latent variable to the items and thereby the influence of exactly this latent variable on each of its indicator items. Moreover, other possible variance sources can be modeled so that the loadings become purified indicators of the latent interindividual differences that are manifested in the score. Thus, the problems with regard to Cronbach's alpha can be successfully dealt with. Another advantage is that the use of structural equation modeling brings some inherent benefits. First, structural equation modeling provides a test for the assumed measurement model underlying the score derived from the items. Second, the analytical framework provides different estimators allowing the analysis of data, which does not follow a multivariate normal distribution using robust estimators such as the robust maximum likelihood estimator (e.g., Satorra & Bentler, 2001). One disadvantage is that, for structural equation models, much larger sample sizes than typically used in EMG studies are needed.

In order to estimate Ω_w, one needs the standardized loadings of the latent variable on all its indicators. The squared loading for each item is regarded as the variance explained by the latent variable. The formula for Ω_w is based on the ratio of the amount of variance explained by the latent variable to the amount of variance not explained. This ratio is summed up across all items and represents the nominator in the formula. Yet, if the items contain more systematic variance than unsystematic variance, this ratio will go asymptotically toward infinity. To avoid this, the formula has a denominator in which the sum of the ratios just explained is added to 1. Thus, if the nominator is 1 (as much systematic as unsystematic variance), the denominator will be 2, and the corresponding reliability estimate will be .5. The larger the variance explained by the latent variable is in relation to the unexplained variance, the greater the ratio will become, and the reliability estimate will increasingly approach the value of 1.

Aims of the Present Research

The present research aims to assess Cronbach's alpha, test-retest correlations, and Ω_w for EMG measures using the contrast approach. In Study 1, a mimicry study, four emotional expressions (anger, sadness, happiness, and disgust) were shown in order to elicit mimicry. Within each specific emotional display, stimuli also varied in terms of the presentation mode (see below). That is, the problem described above regarding stimuli that are influenced by more than one latent variable is present. Here, variance is caused by the specific emotion the stimulus presents but also the presentation mode. Thus, it will be possible to compare the influence of these circumstances on Cronbach's alpha and on Ω_w. An estimate of test-retest reliability is available for a subset of the participants. However, because this subset of participants was rather small ($n = 38$), no latent variable model for the stability of the measure was calculated; instead, we will only present test-retest correlations and intraclass correlations.

We considered three different measurement approaches. First, one way to analyze facial EMG is to focus on the pattern of expressive muscle movement. Thus, for example, positive affect is usually signaled by both an increase in activity of zygomaticus major and a decrease in corrugator supercilii activity (Larsen et al., 2003). As such, a contrast between these two muscles can be calculated to assess this pattern. If, as in our case, orbicularis oculi is also measured, it should be combined with zygomaticus major as these muscles both index smiling. This can be done directly, as we do here, or indirectly via either a contrast (if orbicularis oculi is also considered) or a post hoc test comparing zygomaticus major and corrugator supercilii activity, as these procedures are equivalent.

Alternatively, one can compare the reactions of single muscle sites across conditions. In this case, one would compare the activity of, for example, zygomaticus major across elicitation conditions with the assumption of higher zygomaticus major activity in situations that elicit positive versus negative affect.

In addition, there are different ways to express trial means while controlling for baseline differences. Specifically, even though the theoretical baseline activity for a relaxed muscle would be zero (Hess, 2009), this is not a realistic value. As baseline values differ between participants, they need to be controlled for. Two typical ways of controlling for baseline differences are to take the difference between trial mean and baseline mean (a difference score) or to express the trial mean as a percentage of the baseline (percent score). The former is frequently transformed as the resulting distribution tends to be nonnormal. We used the z transformation recommended for within-subject designs (Bush, Hess, & Wolford, 1993). The z-transformed score also has the advantage to control for idiosyncratic measurement differences such as posed by reduced sensitivity due to idiosyncratic variation in muscle anatomy.

Therefore, using the data from measurement Point 1, reliability for contrast measures as well as single muscle EMGs was estimated. This was done both for within-subject z-transformed difference scores and for untransformed scores expressed as a percentage of baseline level.

Within Study 2, an emotion elicitation study, a more complete test-retest design was realized. Stimuli were images relating to affiliation, achievement, and power motives. After approximately 15 months, participants returned to the laboratory (79% retention) and underwent the same procedure, allowing us to compute a test-

retest correlation as well as a latent test-retest correlation using structural equation modeling. Thus, this data set will allow the comparison of all three reliability estimates.

Study 1

Method

Participants. A total of 162 healthy participants (113 women) were recruited via the participant database at the Humboldt-Universität zu Berlin and participated individually. They received either course credit (42 psychology students) or a small gift as compensation for their participation. Due to several reasons (excessive EMG artifacts during more than one third of the trials, equipment malfunction, or retrospective indications that participants did not adhere to the instructions), data from 30 participants were excluded from analyses. Thus, data of 132 participants (93 women) with a mean age of 26.0 years ($SD = 5.2$ years) were included in the analyses. For a subset of 38 participants (25 women) with a mean age at Time 1 of 24.6 ($SD = 4.2$), the same procedure was repeated at least 24 month after Time 1. The study was carried out in accordance with the guidelines of the Declaration of Helsinki and approved by the Institutional Ethics Committee. Participants were aware that they had the right to terminate participation at any time and that their responses were confidential.

Stimulus Material. Facial expressions were taken from a set of spontaneous facial expressions similar to those that occur during social situations, the Assessment of Contextualized Emotions–Faces (see Hess, Kafetsios, Mauersberger, Blaison, & Kessler, 2016), which consists of a series of photos with a central figure showing four emotional expressions (sadness, happiness, disgust, anger) either by one person (individual condition) or by a central person surrounded by two others who showed either the same emotion (congruent condition) or a neutral expression (incongruent condition). The set consists of a total of 144 stimuli (six male and six female actor groups, four emotions, three types of presentation). Of importance for the current analyses are not the different stimuli per se but rather the fact that different stimulus features exist and the associated variance needs to be modeled.

A Latin square design was used to create 12 parallel orders of 48 stimuli including each central figure of the six male and six female groups either in a congruent (with two friends expressing the same emotion), incongruent (with two friends showing a neutral face), or individual (without the two friends) presentation type for each emotion.

Procedure. Upon arrival at the laboratory, participants were informed about the experimental procedure[1] and signed a consent form. Participants reclined in a comfortable chair while physiological sensors were attached. The experimenter then left the room, monitored the experiment via a video camera, and explained the instructions presented on screen to the participants via microphone. Subsequently, participants watched a 5-min relaxing video; a baseline period for the EMG measures was recorded during the last 3 min of the video. Following this, participants saw one random order of one of the 12 versions of the preselected 48 stimuli on a 520 × 325 mm screen. They were instructed to rate the intensity of the central person's emotion expressions on each of the following 7-point scales anchored with 1 = *not at all* and 7 = *very much*: sadness, happiness, disgust, anger, calm, fear, and surprise, while facial EMG was recorded to assess mimicry. Expressions were presented for 6 s before the rating scales appeared. Finally, participants were fully debriefed.

Facial EMG. Facial EMG was measured at the corrugator supercilii (frown), orbicularis oculi (wrinkles around the eyes), the levator labii superioris (lifting the upper lip in disgust), and the zygomaticus major (lifting the corners of the mouth in a smile) sites on the left side of the face using bipolar placements of EasyCap 4-mm Ag/AgCl miniature surface electrodes filled with SignaGel (Parker Laboratories Inc.). The skin was cleansed with lemon prep peeling and 70% alcohol. Electrodes were placed according to the guidelines published in *Psychophysiology* (Fridlund & Cacioppo, 1986). Raw EMG data were recorded and sampled using a BioLab Acquisition software and a Mindware BioNex BioPotential Amplifier with a 50 Hz notch filter at 1000 Hz. The signals were band-pass filtered between 30 and 300 Hz.

Artifact control and data preparation. The EMG data were offline rectified and smoothed. The video records for each trial and each participant were visually inspected for nonstimulus-related artifacts (e.g., movements such as yawning, coughing, or sneezing) that could disrupt the EMG measures. Periods corresponding to such artifacts were selectively eliminated and excluded from further analyses. Data from stimuli with artifacts lasting longer than one third of the entire stimulus presentation were set missing and thus rejected entirely from further analyses. We then computed within-subject and muscle site z-transformed difference scores for each participant, each muscle, and each trial, and also calculated a contrast index for each emotion based on the scores. For sadness and anger mimicry, we calculated the difference between the muscle activity of corrugator supercilii and the mean muscle activity of orbicularis oculi and zygomaticus major. For happiness mimicry, we subtracted the muscle activity of corrugator supercilii from the mean muscle activity of orbicularis oculi and zygomaticus major. Finally, for disgust mimicry, we calculated the difference between the muscle activity of levator labii superioris and the muscle activity of zygomaticus major. These indices describe the pattern of reactions to the facial expressions. In addition, we calculated for each muscle the trial score as a percentage of the baseline.

Statistical Analysis. All analyses were conducted using R (R Development Core Team, 2015) and the R packages lavaan (Rosseel, 2012), semPlot (Epskamp, 2013), and psych (Revelle, 2014). The complete R code can be found in the online supporting information. In both studies, Cronbach's alpha was estimated based on all items referring to the same emotion.

Within Study 1, the EMG-based scores refer to the emotions anger, disgust, happiness, and sadness. The 12 stimuli for each emotion can be grouped into three method groups (presentation type: stimulus face presented individually, stimulus face presented with bystanders who show a congruent expression, stimulus face presented with bystanders who show an incongruent expression) with four items each. Thus, for the structural equation modeling framework, the model represented in Figure 1 was specified. Figure 1 shows the specific model for anger mimicry. The 12 items are all loaded by a general latent variable representing the emotion. Moreover, the items are grouped according to their additional method variance source, which is also modeled as a latent variable (for a

1. As the current study focuses on the reliability of EMG measures, additional data collected for the original research paradigm will not be discussed in the present context.

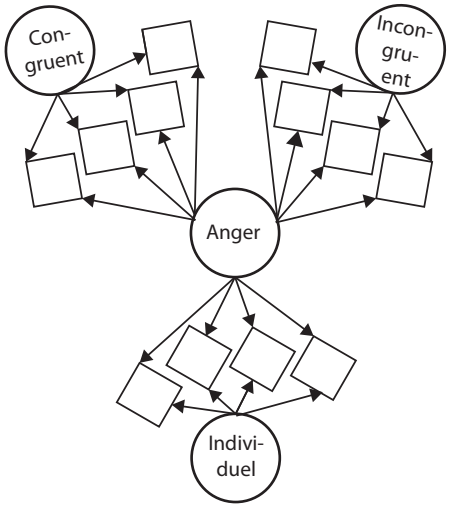

Figure 1. Structural equation model for anger. Covariances between latent method variables are not displayed.

more differentiated discussion of modeling method effects, see Eid et al., 2008). In order to identify each measurement model, latent variances for the first stimulus were always fixed to 1. The latent variables representing method variance were allowed to correlate (not displayed in Figure 1).

This model was specified for each of the four emotional mimicry reactions, once for the contrast score and again for each of the individual muscle sites represented in the contrast score. A maximum likelihood estimator was used. Model fit was based on the recommendations by Hu and Bentler (1999) and Beauducel and Wittmann (2005). Therefore, a global model test in the form of a χ^2 test was conducted. Moreover, the comparative fit index (CFI), the root mean squared error of approximation (RMSEA), and the standardized root mean square residual (SRMR) were used for model evaluation. According to the commonly accepted cutoff criteria, we applied the following rules: CFI $\geq$ .95, RMSEA $<$ .08, in combination with SRMR $<$ .11. However, those cutoffs were derived from simulation studies assuming average loadings that are most likely too high for the purpose of EMG-based scores (around. 75). For instances such as this, Heene, Hilbert, Draxler, Ziegler, and Bühner (2011) recommend carefully investigating sources of misfit. Moreover, these authors suggest modeling the sources of misfit and then testing the parameters in question. Therefore, in case the specified models did not fit the data as judged by the criteria just stated, misfit was investigated and modeled. All estimated construct reliabilities were based on models that were in accordance with the set cutoffs. In order to estimate Ω_w, an R function was programmed, which can be found in the supporting information. This function uses the output generated with the lavaan package during model testing and estimates construct reliabilities for all latent variables in the model.

Results and Discussion

Table 1 includes model fits for all tested models. As can be seen, none of the models fitted the data without adjustment. Table 2 shows estimates for Cronbach's alpha, Ω_w, and test-retest correlations. A first consideration has to be what can be considered an acceptable level of internal consistency. Generally speaking, reliabilities for facial EMG measures might more reasonably be compared to interrater reliabilities for facial expression coding than to internal consistencies for questionnaire items. For behavioral coding, interrater agreements of .70 are frequently considered adequate (LeBreton & Senter, 2007). The Cronbach's alphas generally fell well below this criterion for all measures. In fact, some alphas were close to zero even after negatively loading items had been eliminated. Yet, this can be expected, as the stimuli were presented in three rather different modes and the shared variance related to presentation mode should bias estimates of alpha.

Notably, the alphas for the untransformed single muscle EMG scores expressed as a percentage of baseline were much higher than for any *z*-transformed data or the contrast scores. This is to be expected: When trial means are expressed as percentage of baseline in a ratio, this baseline is common to all resulting item scores, and hence there is a partial autocorrelation between those scores.[2] When partial autocorrelations occur in psychometric analyses, one way to address the issue prior to reliability testing is to subtract the common element, in this case the baseline, from the trial mean. This is a standard procedure when estimating item-total discriminations (Guilford, 1954). As such, the alphas obtained for the baseline percentage scores cannot be considered good estimates.

Notably, this autocorrelation problem potentially also distorts the omegas reported below. The omegas are estimated based on the loadings within a structural equation model. The loadings in turn are estimated based on the items' correlations and variances. The former, as we noted, are potentially inflated due to autocorrelations. These inflations will distort the loadings and, hence, the omegas. It can further be assumed that the strongest source of variance in all items would "grab" the common variance because of the autocorrelation and would hence be inflated.

Within some of the models using the baseline percentage scores, the correlations between the method factors were extremely high and therefore sometimes had to be fixed to unity. In those models, method variance could be overestimated. In other models, those correlations had to be fixed to zero or method effects could not even be modeled. Consequently, the latent variable representing the emotion will be strongest, and its effects and thus its reliability could be overestimated. In different contexts and samples, these effects may occur for different muscles. Overall, the alphas as well as the omegas for the trial means expressed as percent baseline have to be treated with caution.[3]

When construct reliabilities are considered, the impact of presentation mode (i.e., the context in which the facial expression was shown) becomes clear. Notably, the construct reliabilities for the presentation mode were quite substantial, but as Table 2 shows, they differed hugely between presentation modes, emotions, and measures. For the contrast measure, congruent and individual presentation yielded reliable method factors in each emotion, whereas the incongruent presentation factor was only reliable in combination with anger expressions. For the individual muscles, this pattern was generally obtained as well, but there are exceptions.

2. The occurrence of this bias as well as its magnitude depends on the level and the variance of individual differences in the baseline, for example, as a result of electrode placement or skin factors. If these individual differences are small and/or invariant, the size of the autocorrelation and thus its biasing influence is limited.

3. In Study 1, a number of items loaded negatively on the main construct. These negative loadings were usually minor. Both setting such item loadings to zero and reversing their loadings would inflate omega w post hoc. We opted to reverse their loadings when estimating omega w. This may have led to slightly inflated omega w estimates for the anger model, but should have affected the other models only minimally (see supporting information).

Table 1. *Model Fits for All Tested Models in Studies 1 and 2*

	χ^2	*df*	*p*	CFI	RMSEA	SRMR
			Study 1: Contrast measures			
Anger	30.64	39	.838	1.00	0.00	0.05
Anger fitted	40.73	44	.613	1.00	0.00	0.06
Disgust	Did not converge					
Disgust fitted	42.56	42	.447	0.99	0.01	0.07
Happiness	Did not converge					
Happiness fitted	50.24	46	.309	0.52	0.03	0.08
Sadness	46.28	39	.197	0.84	0.04	0.06
Sadness fitted	40.07	40	.467	1.00	0.01	0.06
			Study 1: Individual muscle sites, difference scores			
Anger corrugator supercilii	48.20	39	.148	0.90	0.04	0.07
Anger corrugator supercilii modified	49.01	41	.183	0.91	0.04	0.07
Happiness corrugator supercilii	71.39	39	.001	0.78	0.08	0.07
Happiness corrugator supercilii modified	41.88	39	.347	0.98	0.02	0.05
Sadness corrugator supercilii	Did not converge					
Sadness corrugator supercilii modified	46.62	39	.188	0.94	0.04	0.06
Anger zygomaticus major	25.56	39	.952	1.00	0.00	0.05
Anger zygomaticus major modified	35.49	45	.844	1.00	0.00	0.06
Happiness zygomaticus major	26.88	39	.929	1.00	0.00	0.05
Happiness zygomaticus major modified	34.56	44	.845	1.00	0.00	0.06
Sadness zygomaticus major	30.14	39	.845	1.00	0.00	0.05
Sadness zygomaticus major modified	45.54	44	.408	0.90	0.02	0.06
Anger orbicularis oculi	21.02	39	.992	1.00	0.00	0.04
Anger orbicularis oculi modified	21.34	40	.993	1.00	0.00	0.04
Happiness orbicularis oculi	14.05	39	1.000	1.00	0.00	0.04
Happiness orbicularis oculi modified	14.06	40	1.000	1.00	0.00	0.04
Sadness orbicularis oculi	52.29	42	.133	0.73	0.04	0.07
Sadness orbicularis oculi modified	41.30	41	.458	0.99	0.01	0.06
Disgust zygomaticus major	Did not converge					
Disgust zygomaticus major modified	35.09	45	.856	1.00	0.00	0.06
Disgust levator labii superioris	34.55	39	.673	1.00	0.00	0.05
Disgust levator labii superioris modified	45.35	45	.457	0.99	0.01	0.06
			Study 1: Individual muscle sites, percent baseline scores			
Anger corrugator supercilii	121.26	39	<.001	0.94	0.13	0.04
Anger corrugator supercilii modified	121.35	40	<.001	0.94	0.12	0.04
Happiness corrugator supercilii	126.09	39	<.001	0.92	0.13	0.06
Happiness corrugator supercilii modified	135.95	43	<.001	0.92	0.13	0.06
Sadness corrugator supercilii	Did not converge					
Sadness corrugator supercilii modified	124.53	41	<.001	0.94	0.12	0.04
Anger zygomaticus major	136.88	39	<.001	0.89	0.14	0.09
Anger zygomaticus major modified	97.35	39	<.001	0.93	0.11	0.08
Happiness zygomaticus major	187.45	39	<.001	0.87	0.17	0.16
Happiness zygomaticus major modified	151.64	40	<.001	0.90	0.15	0.16
Sadness zygomaticus major	125.91	39	<.001	0.88	0.13	0.06
Sadness zygomaticus major modified	175.71	51	<.001	0.83	0.14	0.09
Anger orbicularis oculi	119.84	39	<.001	0.87	0.13	0.07
Anger orbicularis oculi modified	106.15	40	<.001	0.89	0.11	0.07
Happiness orbicularis oculi	78.13	39	<.001	0.94	0.09	0.06
Happiness orbicularis oculi modified	117.59	40	<.001	0.86	0.12	0.06
Sadness orbicularis oculi	87.64	39	<.001	0.93	0.10	0.07
Sadness orbicularis oculi modified	121.61	45	<.001	0.89	0.11	0.06
Disgust zygomaticus major	114.11	39	<.001	0.87	0.12	0.10
Disgust zygomaticus major modified	120.94	42	<.001	0.86	0.12	0.09
Disgust levator labii superioris	123.48	39	<.001	0.94	0.13	0.06
Disgust levator labii superioris modified	139.62	42	<.001	0.93	0.13	0.06
			Study 2			
Affiliation W1	6.04	2	.05	0.96	0.100	0.036
Affiliation W1 fitted	0.02	1	.88	1	0	0.002
Power W1	0.02	2	.99	1	0	0.003
Power W1 fitted	0.37	3	.95	1	0	0.012
Achievement W1	14.40	2	.001	0.78	0.176	0.060
Achievement W1 fitted	2.29	1	.13	0.98	0.80	0.019
Affiliation W2 fitted	0.06	1	.815	1	0	0.003
Power W2*	3.72	2	.156	0.92	0.071	0.035
Achievement fitted W2	0.00	1	.974	1	0	0.001

Note. W = wave.
*It was not possible to replicate the ad hoc adjustments from W1 for power.

Table 2. *Cronbach's Alphas, Construct Reliabilities, and Test-Retest Reliability for Study 1*

Stimuli	Cronbach's alpha	Ω_w Emotion	Ω_w Congruent	Ω_w Incongruent	Ω_w Individual	Test-retest correlations ($N = 38$)	ICC
			Reactions to happiness expressions				
Contrast	0.05	0.74	0.91	0.35	0.65	0.51	0.66
Difference scores							
Corrugator supercilii	0.23	0.91	1.00	0.63	0.60	0.32	0.49
Zygomaticus major	0.10[1]	0.67	0.53	0.43	0.99	0.38	0.55
Orbicularis oculi	0.03[1]	0.47	0.03	0.38	0.45	0.23	0.38
Percent baseline							
Corrugator supercilii	0.94	0.31	0.95	0.15	0.36	0.34	0.50
Zygomaticus major	0.92	0.98	0.97	0.96	0.61	0.88	0.93
Orbicularis oculi	0.89	0.45	0.93	0.47	0.87	0.18	0.29
			Reactions to sadness expressions				
Contrast	0.38	0.49	0.79	0.29	0.71	0.46	0.64
Difference scores							
Corrugator supercilii	0.42	0.82	0.59	1.00	0.58	0.25	0.40
Zygomaticus major	0.02[1]	0.53	0.99	0.93	0.24	0.06	0.12
Orbicularis oculi	0.07	0.68	0.37	0.31	0.25	0.04	0.07
Percent baseline							
Corrugator supercilii	0.94	0.35	0.29	0.44	0.97	0.32	0.48
Zygomaticus major	0.90	0.91	–	–	–	0.15	0.25
Orbicularis oculi	0.91	0.59	0.54	0.86	0.90	0.25	0.27
			Reactions to anger expressions				
Contrast	0.11	0.45	0.86	0.95	0.94	0.35	0.48
Difference scores							
Corrugator supercilii	0.07	0.85	0.99	0.78	0.94	0.09	0.17
Zygomaticus major	0.04[1]	0.52	0.94	0.79	0.86	0.12	0.22
Orbicularis oculi	0.05[1]	0.45	0.35	0.22	0.35	0.06	0.12
Percent baseline							
Corrugator supercilii	0.95	0.89	0.79	0.71	0.66	0.40	0.56
Zygomaticus major	0.90	0.91	0.92	0.80	0.80	0.61	0.69
Orbicularis oculi	0.88	0.91	0.32	0.51	0.29	0.17	0.21
			Reactions to disgust expressions				
Contrast	0.03[1]	0.62	0.89	0.30	0.97	0.17	0.27
Difference scores							
Levator labii superioris	0.01[1]	0.51	1.00	0.84	1.00	0.32	0.48
Zygomaticus major	0.03[1]	0.56	0.98	0.69	0.96	0.16	0.28
Percent baseline							
Levator labii superioris	0.86	0.48	0.94	0.72	0.55	−0.07	−0.08
Zygomaticus major	0.83	0.93	0.88	0.97	1.00	0.43	0.43

Note. ICC = intraclass correlations.
[1]One to three items were negatively correlated with the total scale, and their factor loading had to be fixed to zero ad hoc.

By and large, these findings suggest that presentation mode can have a reliable impact on facial reactions to facial expressions. As the presentation mode in this case varied the social context of the expression in that either an individual alone or congruent versus incongruent others were shown, this suggests that facial EMG measures show overall good internal consistency for social context effects.

However, of principal interest here are the construct reliabilities for the emotional mimicry factors controlling for these method effects, that is, whether facial EMG measures show internal consistency with regard to the emotion that they are supposed to index when method factors are controlled for. When considering individual muscle reactions, Table 2 shows that, for the *z*-transformed data across all emotions for which corrugator supercilii is part of the indexed facial reaction, the reliabilities for the corrugator supercilii reactions were above .80 and thus highly reliable for a behavioral measure. Thus, when assessing reactions to facial stimuli, corrugator supercillii reactions were indeed internally consistent. However, none of the other muscle sites reached the criterion level of .70 even though zygomaticus major reactions to happy faces and orbicularis oculi to sad faces came close. By contrast, for the percent baseline measure construct reliability was generally low, but reliability for zygomaticus major was high. The construct reliabilities for all three indexed muscles for reactions to angry faces were high as well. However, as mentioned above, these estimates should be treated with caution.

When combining muscle sites for a contrast measure, Table 2 shows that construct reliability was highest for happiness, suggesting that the pattern of facial reactions to happy faces is adequately reliable. Facial reactions to all other emotions, however, were again relatively unreliable.

These findings suggest that, when showing a series of facial expressions of anger, sadness, or disgust, reaction reliability estimates vary—both when indexed by individual muscles or by a pattern of muscle activity and depending on which type of score is chosen. This may occur for different reasons. One of the reasons is that, even though we controlled for variance due to presentation mode, the design had in fact other methodological aspects that varied between items. Thus, different participants saw different actors, who can also be grouped into men and women. The low internal consistency suggests that the same emotion expression elicits somewhat different reactions in observers, depending on who shows the emotion expression. The theoretical question is whether this is a reason for concern.

The answer depends on the goal of the study and the underlying assumptions. As noted above, when items represent more than one construct (for example, gender as it interacts with emotion expression, see Becker, Kenrick, Neuberg, Blackwell, & Smith, 2007; Hess, Adams, & Kleck, 2009), then a lack of internal consistency may be expected. In this case, it would actually make little sense to demand high internal consistencies. In essence, for the present study this means that facial reactions to facial expressions or emotional mimicry reactions are influenced not only by the emotion that is reacted to, but also by the actor who shows the emotion. In fact, studies on emotion perception show that the same identical expressions will be evaluated differently depending on who shows the expression (Wiggers, 1982).

The second question relates to the issue of whether facial mimicry responses are stable over time. In order to address this question, we combined the EMG data across presentation modes and calculated both test-retest correlations and intraclass correlations for the 38 participants for whom data for two time points (at least 24 months apart) were available (see Table 2). Waters, Williamson, Bernard, Blouin, and Faulstich (1987) assessed test-retest correlations over a period of 2 weeks for a variety of measures including frontalis EMG for several stress-related tasks. Test-retest reliability for frontalis EMG varied from $r = -.28$ for a habituation task to $r = .40$ for stress imagery. One reason for the low and even negative correlations in that study may have been the differential practice effects mentioned above, which were likely strong given the short time gap.

As can be seen in Table 2, the test-retest correlations for the contrast measure for all emotions except disgust compare favorably with the highest of the values found by Waters et al. (1987), suggesting that patterns of facial mimicry assessed with EMG are relatively stable responses over time.

The issue is somewhat more complex when individual muscle reactions are considered. Test-retest correlations for individual muscle sites were generally lower than for the contrast measure and varied considerably between site and emotions. For the percent baseline score for zygomaticus major, high test-retest correlations can be observed for reactions to happy and angry faces, but otherwise correlations were low and varied considerably between site and emotions as well. A likely reason for the finding that the test-retest reliabilities of the contrast measure were somewhat superior is that aggregating across different muscles maximized the shared and thus reliable variance. This would suggest that using a contrast measure is preferable, especially when the EMG data are to serve as a predictor measure.

In sum, Study 1 suggests that the high internal consistency of the means expressed as percent of baseline is inflated due to autocorrelation. The lower internal consistency of the alternative methods can be explained by the observation that reactions are influenced by factors other than the emotion shown by the target. It has to be noted here that the occurrence of unintended variance is also a problem in widely used questionnaires (Ziegler, Poropat, & Mell, 2014) and as such is not a reason to advocate against using this measure. Importantly, the facial mimicry reactions are, with the exception of disgust mimicry, acceptably stable over time when contrast measures are used.

Study 2

Method

Participants. In the context of a larger study on the transition from student life to work, we recruited (209, 66% women) students with a mean age of 27.48 ($SD = 3.07$) years, who were in the process of submitting their final theses, from universities in and around Berlin. We aimed for a representative selection of study domains, but excluded psychology students from participation to ensure that participants would be unfamiliar with the tests used. Students of social sciences were slightly overrepresented. Participants agreed to be contacted again once they had left university and were rewarded with 120 € upon completion of the second wave. The study was carried out in accordance with the guidelines of the Declaration of Helsinki and approved by the Institutional Ethics Committee. Participants were aware that they had the right to terminate participation at any time and that their responses were confidential.

Stimulus Material. The images were taken from various sources and grouped according to their motive content. Four images each represented one of three motives, affiliation (images of a smiling elderly couple, hand-holding children, a family on a trip, and friends laughing together), achievement (images of a rock climber, a student holding up her diploma, a runner finishing first, and a basketball player dunking), and power (an image of a gavel, a superhero, a politician waving to a crowd, and a crime boss).[4] In an independent online study with 26 (65% women, mean age = 23.04, $SD = 4.98$) raters, we found that these stimuli represented the desired motive content (for details, see Dufner et al., 2015, Appendix C).

Procedure. The introductory procedure was largely identical to Study 1. Participants were told that the electrodes measured skin conductance and only debriefed after the second wave.

Participants saw the stimuli as part of a series of tasks that also included games and implicit association tests. Within the task, image order was random. The tasks were presented at the center of a computer screen (400 mm × 260 mm) approximately 80 cm in front of the participants. For each image, participants first saw a white fixation cross on a black background for 1 s, then the image for 4 s, and then three rating scales. Participants had to rate each picture on three 5-point scales, anchored with 1 = *do not agree at all* and 5 = *agree totally*, to which extent they felt arousal, positive, and negative emotions.

Facial EMG. We recorded the zygomaticus major and corrugator supercilii activity on the left side of the face using bipolar placements of 4 mm Ag/AgCl miniature surface electrodes filled with electrode gel. Electrodes were placed according to the guidelines published by *Psychophysiology* (Fridlund & Cacioppo, 1986). The skin was cleansed with lemon prep peeling and 70% alcohol.

Raw EMG data were recorded and sampled using a digital Psychlab amplifier with a 50 Hz notch filter at 1000 Hz. Offline, the signals were band-pass filtered between 30 and 300 Hz, rectified, and *z*-standardized within in each person.

For each image and muscle, the reaction from 1 s after stimulus presentation until the end of stimulus presentation was averaged and baseline corrected. This window was chosen because it was used in the published article as well. We then subtracted the baseline-corrected average for corrugator supercilii response from the baseline-corrected average for the zygomaticus major response. The resulting contrast scores were then used as factor indicators.

4. At Wave 2, we also presented participants with a random selection of pictures that were taken from a large a pool of motive-relevant pictures. However, as the presented pictures varied across participants, EMG reactions to these cues were not considered in the current analyses.

Table 3. *Cronbach's Alphas, Construct Reliabilities, and Test-Retest Reliability for Study 2*

	Affiliation		Power		Achievement	
	W1	W2	W1	W2	W1	W2
n	209	166	209	165	209	165
Cronbach's alpha	0.62	0.72	0.15	0.40	0.52	0.46
Ω_w motive	0.66	0.88	0.73[1]	0.63	0.52	0.57
Test-retest correlations	0.50		0.15		0.28	
Latent test-retest correlations	0.73		−0.04		0.70	

Note. W = wave.
[1]One item was negatively correlated with the total scale, and its factor loading had to be fixed to zero ad hoc. In W2, the same item was correlated positively with the scale and left in.

Statistical Analysis. The analyses were conducted using the same statistical packages as in Study 1. We used full information maximum likelihood, which allowed us to use all cases for the retest analyses. We used robust Huber-White standard errors because the EMG scores were not normally distributed. Unlike in Study 1, there were no method factors to consider, so each motive was simply modeled as a latent variable with the four image scores as indicators.

Results and Discussion

Similar to Study 1, the model fits were inadequate before ad hoc adjustments were carried out (see Table 1). Adjustments required correlating two item residuals in both the affiliation and the achievement model. In the case of power, one item insignificantly loaded in the opposite direction of the prediction and its factor loading was fixed to zero (see also Footnote 2). The complete analyses, including both code and results, can be found in the supporting information.

As Table 3 shows, Cronbach's alphas were all below the criterion of .70. Construct reliabilities using Ω_w were generally higher and for affiliation came close to or exceeded the criterion. This increase is probably due to the lower prescriptiveness of Ω_w with respect to required uniformness of the stimuli. The Cronbach's alpha for power was very low because one item actually correlated negatively with the item total. For the assessment of Ω_w, this problem was corrected, and consequently internal consistency is much higher when assessed in this manner.

In all, the internal reliability of the measures as estimated by Ω_w was only adequate for the measurement of affiliation. The Ω_w values for achievement were lower than criterion but still above .50. Given that only four items were used, it can be argued that an internal consistency above .50 is already acceptable. This argument has indeed been made for questionnaire items for which internal consistency is usually expected to be higher than the .70 used as criterion for behavioral measures. However, even for self-report questionnaire items, internal consistencies of $< .70$ are often considered acceptable if scales consist of very few items (Hahn, Gottschling, & Spinath, 2012).

With respect to power, however, the problem may be with the selection of stimuli, rather than in the EMG measure. This notion is also supported by the observation that the fitted structural equation model for power did not replicate for Wave 2.

To assess latent retest correlations, we first built the measurement models separately for each wave for each motive. Achievement and affiliation had similarly strong latent rank-order stabilities. The findings suggest that, when appropriate items are used, affective reactions to pictorial stimuli as indexed by an EMG contrast measure are very stable over time.

General Discussion

The present research had the goal to assess both internal consistencies and test-retest reliabilities for facial EMG measures using two examples: facial reactions to facial expressions (Study 1) and facial reactions indicative of positive affective reactions to motive-relevant images (Study 2). The results provide a somewhat complex picture. Cronbach's alphas were generally low except for data expressed as a percentage of baseline, but in the latter case alpha is likely to be inflated due to autocorrelation. Another important exception was the measurement of affiliation in Study 2, which showed good reliability.

More relevant, however, are the construct reliability estimates—that is, the reliability estimated for the emotion measurement score. On the level of individual muscles, only *z*-transformed corrugator supercilii reactions (Study 1) were found to have consistently acceptable construct reliability estimates. High values for zygomaticus major when expressed as percentage of baseline were also found, but the calculation of omega needs to be treated with caution due to partial autocorrelation between the items when calculated in this way. In both studies, when using contrast measures, only some reactions (reactions to smiling faces and positive affective reactions to affiliative images) were found to have internal consistencies that compare well with reliability estimates for scores derived from behavioral measures.

In Study 1, facial reactions to negative emotions and, in Study 2, facial reactions to power and somewhat less to achievement imagery showed unsatisfactory internal consistencies. The images shown in both studies varied in a number of characteristics and especially in the specific content. For example, even though all faces in Study 1 expressed the same emotion, the faces themselves varied. Also, even though all images in Study 2 depicted a specific type of event associated with a specific motive, the actual events shown differed considerably. That is, the specific context for each image seemed to have a rather large impact on the facial reactions. In Study 1, one context, the type of presentation (i.e., whether or not others who showed the same expression were shown on the picture), could be explicitly modeled. The results showed a more consistent impact of context than of emotion expression, for two of the three contexts.

These findings show that, even when pretests suggest that images are judged to be equivalent in terms of their context, this may not ensure that the context and other specifics do not influence the facial expressions shown in response. While in psychophysiological research it has long been a tradition to control for aspects of

images such as luminosity and contrast, the specific content domain has often been neglected as long as all images had the same valence or represented the same emotion. The present data suggest that it may be problematic to consider all items in a series of images as equivalent.

Constructs such as affiliation or anger expressions are inherently complex. The investigator's guide for the facial action coding system (Ekman, Friesen, & Hager, 2002) specifies several prototypes for each emotion, including more than 30 prototype variants for anger expressions alone. These variants combine different muscles with different intensities and hence result in visibly different expressions, which then may well result in different facial mimicry reactions. In addition, the facial morphology of the expresser (Wiggers, 1982) impacts on the perception of these expressions as do facial morphological traits such as dominance, masculinity, maturity, or affiliation, which in turn tend to be confounded by gender (Becker et al., 2007; Hess et al., 2009; Marsh, Adams, & Kleck, 2005). As such, it may not be realistic to expect uniform reactions to these heterogeneous stimuli. Similar arguments can be made with regard to content domains such as affiliation or power. A wide variety of possible contents can reflect these motives, and each of these contents likely elicits specific facial reactions as well.

Yet, when it comes to the stability of these complex reactions, we found much more reason for optimism. In Study 1, the test-retest reliability over a 2-year period was very acceptable not only for happiness but also for anger and sadness expressions when the contrast measure was used. This suggests that these reactions reflect a stable tendency over time. That is, even though people's facial reactions to the facial expressions of others are seemingly influenced by a variety of factors, which results in low internal consistency, these factors in turn seem to be quite stable over time.

An exception to this relative satisfactory stability of mimicry reactions constituted reactions to disgust. It should be noted that, in the vast majority of studies on facial mimicry, only reactions to happy and angry or sometimes alternatively sad expressions are measured. Of the few studies that included disgust, even fewer found significant patterns indicating disgust mimicry on the group level (for a review, see Hess & Fischer, 2013). This was also the case in Study 1. Across all participants, the effect for disgust mimicry was only marginally significant; further, the tendency to mimic disgust at all was positively related to neuroticism (Mauersberger et al., 2015). As such, it may not be surprising that reactions to this expression are not stable.

Similarly, in Study 2, test-retest reliabilities over a 15-month period were quite high for a behavioral measure at least for affiliation and achievement images, suggesting that people's affective reaction to such images is indeed stable over time. Interestingly, power images were not only internally inconsistent but also had negligible test-retest reliability. This suggests that, unlike reactions to images with affiliative or achievement content, the participants' reactions to the power content changed over this time period. As the participants were in a transition phase from the university to the business world, this may reflect a change in perception of the problematic item as much as a lack of reliability. Given that only four images each were used for the motives, it is difficult to draw strong conclusions from these findings and a replication is certainly needed. The findings suggest that, especially when affective reactions to pictorial stimuli are measured via EMG, much attention has to be paid to the specifics of the content of these stimuli.

The present data also point to some interesting issues with regard to reliability coefficients. As the data show, high internal consistency and high test-retest reliability are not the same. In most cases, test-retest reliabilities were higher than internal consistencies, but there were exceptions. This points to the fact that neither is a prerequisite for the other and that even measures with low internal consistency can be stable over time. In fact, in other domains such as measures of implicit motives, similar observations have been made (Schultheiss, Liening, & Schad, 2008). Thus, it is not appropriate to conclude that a measure with low internal consistency cannot be used to predict behavior at later times. This is especially relevant in contexts where facial EMG is used as a predictor or to describe specific behavioral styles.

Moreover, the present analyses portray different indices to estimate reliability. Of all these indices, Cronbach's alpha entails the strictest prerequisites (tau-equivalence) while yielding the overall poorest results at the same time. Importantly, even when Cronbach's alpha was very high as in the case of the single muscle data expressed as percentage of baseline, the internal consistency did not derive as much from the common emotion construct as from common method effects in combination with a statistical artifact. Thus, we strongly encourage the use of alternative coefficients, such as omega w or, when possible, test-retest correlation.

In sum, of the individual muscle sites, only z-transformed corrugator supercilii and zygomaticus major when expressed as percentage of baseline showed good estimated construct reliabilities. Facial reactions to happy faces as well as positive affective reactions to affiliative images (and to a lesser degree to achievement images) can be measured with adequate reliability both in terms of construct reliability and test-rest reliability when contrast measures are used. Only the latter is the case for facial reactions to anger and sadness. Facial reactions to disgust expressions and power images by contrast were not reliable in either sense. For Study 1, the findings suggest that the specific context in which an image is shown or specific content elements of the image strongly influence the facial reactions to pictorial stimuli, but that these reactions in turn tend to be stable over time. For Study 2, it seems that affiliation works quite well and that achievement can be improved if one controls for manifest unreliability, but that power stimuli are more problematic. Overall, more research is needed to gain insight into the methodological factors that affect the reliability of EMG-derived individual difference indices.

References

Armstrong, J. E., Hutchinson, I., Laing, D. G., & Jinks, A. L. (2007). Facial electromyography: Responses of children to odor and taste stimuli. *Chemical Senses*, *32*, 611–621. doi: 10.1093/chemse/bjm029

Arndt, J., Allen, J. J. B., & Greenberg, J. (2001). Traces of terror: Subliminal death primes and facial electromyographic indices of affect. *Motivation and Emotion*, *25*, 253–277. doi: 10.1023/A:1012276524327

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*(1), 41–75.

Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K. C., & Smith, D. M. (2007). The confounded nature of angry men and happy women. *Journal of Personality and Social Psychology*, *92*, 179–190. doi: 10.1037/0022-3514.92.2.179

Brown, L. M., Bradley, M. M., & Lang, P. J. (2006). Affective reactions to pictures of ingroup and outgroup members. *Biological Psychology*, *71*, 303–311.

Buck, S. M., Hillman, C. H., Evans, E. M., & Janelle, C. M. (2004). Emotional responses to pictures of oneself in healthy college age females. *Motivation and Emotion*, *28*, 279–295. doi: 10.1023/B:MOEM.0000040155.79452.23

Bush, L. K., Hess, U., & Wolford, G. (1993). Transformations for within-subject designs: A Monte Carlo investigation. *Psychological Bulletin*, *113*, 566–579. doi: 10.1037/0033-2909.113.3.566

Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can discriminate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, *50*, 260–268. doi: 10.1037/0022-3514.50.2.260

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*, 1–16. doi: 10.1007/BF02289289

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi: 10.1007/BF02310555

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. doi: 10.1177/0013164404266386

Dambrun, M., Desprès, G., & Guimond, S. (2003). On the multifaceted nature of prejudice: Psychophysiology responses to ingroup and outgroup ethnic stimuli. *Current Research in Social Psychology*, *8*, 200–204.

Davis, W. J., Rahman, M. A., Smith, L. J., & Burns, A. (1995). Properties of human affect induced by static color slides (IAPS): Dimensional, categorical and electromyographic analysis. *Biological Psychology*, *41*, 229–253. doi: 10.1016/0301-0511(95)05141-4

Dimberg, U. (1982). Facial reactions to facial expressions. *Psychophysiology*, *19*, 643–647. doi: 1469-8986.1982.tb02516.x

Dimberg, U. (1990). Perceived unpleasantness and facial reactions to auditory stimuli. *Scandinavian Journal of Psychology*, *31*, 70–75. doi: j.1467-9450.1990.tb00804.x

Dimberg, U., & Ohman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion*, *20*, 149–182. doi: 10.1007/BF02253869

Dufner, M., Arslan, R. C., Hagemeyer, B., Schönbrodt, F. D., & Denissen, J. J. (2015). Affective contingencies in the affiliative domain: Physiological assessment, associations with the affiliation motive, and prediction of behavior. *Journal of Personality and Social Psychology*, *109*, 662–676. doi: 10.1037/pspp0000025

Eid, M., Nussbeck, F., Geiser, C., Cole, D., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*, 230–253. doi: 10.1037/a0013219

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system investigator's guide*. Salt Lake City, UT: Research Nexus.

Epskamp, S. (2013). semPlot: Path diagrams and visual analysis of various SEM packages' output (Version R package version 1.0.1). Retrieved from http://CRAN.R-project.org/package=semPlot

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, *117*, 39–66. doi: 10.1037/0033-2909.117.1.39

Fridlund, A. J., & Cacioppo, J. T. (1986). Guidelines for human electromyographic research. *Psychophysiology*., *23*, 567–589. doi: 10.1111/j.1469-8986.1986.tb00676.x

Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality—Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality*, *46*, 355–359. doi: 10.1016/j.jrp.2012.03.008

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. D. Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—Festschrift in honor of Karl Jöreskog* (pp. 195–216): Lincolnwood, IL: Scientific Software International, Inc.

Hazlett, R. L., & Hazlett, S. Y. (1999). Emotional response to television commercials: Facial EMG vs. self-report. *Journal of Advertising Research*, *39*, 7–23.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*, 319–336. doi: 10.1037/a0024917

Hess, U. (2009). Facial EMG. In E. Harmon-Jones & J. S. Beer (Eds.), *Methods in social neuroscience* (pp. 70–91). New York, NY: Guilford Press.

Hess, U., Adams, R. B., Jr., & Kleck, R. E. (2009). The face is not an empty canvas: How facial expressions interact with facial appearance. *Philosophical Transactions of the Royal Society London B*, *364*, 3497–3504. doi: 10.1098/rstb.2009.0165

Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology*, *40*, 129–141. doi: 10.1016/S0167-8760(00)00161-6

Hess, U., & Bourgeois, P. (2010). You smile—I smile: Emotion expression in social interaction. *Biological Psychology*, *84*, 514–520. doi: 10.1016/j.biopsycho.2009.11.001

Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review*, *17*, 142–157. doi: 10.1177/1088868312472607

Hess, U., & Hareli, S. (2015). The role of social context for the interpretation of emotional facial expressions. In M. K. Mandal & A. Awasthi (Eds.), *Understanding facial expressions in communication* (pp. 119–141). New York, NY: Springer.

Hess, U., Kafetsios, K., Mauersberger, H., Blaison, C., & Kessler, C.-L. (2016). Signal and noise in the perception of facial emotion expressions: From labs to life. *Personality and Social Psychology Bulletin*, *42*(8), 1092–1110.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi: 10.1080/10705519909540118

Hu, S., Player, K. A., Mcchesney, K. A., Dalistan, M. D., Tyner, C. A., & Scozzafava, J. E. (1999). Facial EMG as an indicator of palatability in humans. *Physiology & Behavior*, *68*, 31–35. doi: 10.1016/S0031-9384(99)00143-2

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Jäncke, L., & Kaufmann, N. (1994). Facial EMG responses to odors in solitude and with an audience. *Chemical Senses*, *19*, 99–111. doi: 10.1093/chemse/19.2.99

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*, 776–785. doi: 10.1111/1469-8986.00078

Lawrence, J. H., & DeLuca, C. J. (1983). Myoelectric signal versus force relationship in different human muscles. *Journal of Applied Physiology*, *54*, 1653–1659.

LeBreton, J. M., & Senter, J. L. (2007). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*. *11*, 815–852. doi: 10.1177/1094428106296642

Marsh, A. A., Adams, R. B., Jr., & Kleck, R. E. (2005). Why do fear and anger look the way they do? Form and social function in facial expressions. *Personality and Social Psychological Bulletin*, *31*, 73–86. doi: 10.1177/0146167204271306

Mauersberger, H., Blaison, C., Kafetsios, K., Kessler, C.-L., & Hess, U. (2015). Individual differences in emotional mimicry: Underlying traits and social consequences. *European Journal of Personality*, *29*, 512–529. doi: 10.1002/per.2008

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28–50. doi: 10.1177/1088868310366253

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mojzisch, A., Schilbach, L., Helmert, J. R., Pannasch, S., Velichkovsky, B. M., & Vogeley, K. (2006). The effects of self-involvement on attention, arousal, and facial expression during social interaction with virtual others: A psychophysiological study [Special issue: Theory of mind]. *Social Neuroscience*. *1*, 184–195. doi: 10.1080/17470910600985621

Moody, E. J., McIntosh, D. N., Mann, L. J., & Weisser, K. R. (2007). More than mere mimicry? The influence of emotion on rapid facial reactions to faces. *Emotion*, *7*, 447–457. doi: 10.1037/1528-3542.7.2.447

Newton, T. F., Khalsa-Denison, M. E., & Gawin, F. H. (1997). The face of craving? Facial muscle EMG and reported craving in abstinent and non-abstinent cocaine users. *Psychiatry Research*, *73*, 115–118. doi: 10.1016/S0165-1781(97)00115-7

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*, 343–355. doi: 10.1037/1082-989X.5.3.343

R Development Code Team. (2015). *foreign: Read data stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... R package* (Version 0.8-62). Retrieved from http://CRAN.R-project.org/package=foreign

Revelle, W. (2014). *psych: Procedures for personality and psychological research (Version 1.4.5)*. Northwestern University, Evanston, IL. Retrieved from http://CRAN.R-project.org/package=psych

Robinson, J. D., Cinciripini, P. M., Carter, B. L., Lam, C. Y., & Wetter, D. W. (2007). Facial EMG as an index of affective response to nicotine. *Experimental and Clinical Psychopharmacology*, *15*, 390–399. doi: 10.1037/1064-1297.15.4.390

Rossel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514. doi: 10.1007/BF02296192

Schultheiss, O. C., Liening, S., & Schad, D. J. (2008). The reliability of a picture story exercise measure of implicit motives: Estimates of internal consistency, retest reliability, and ipsative stability. *Journal of Research in Personality*, *42*, 1560–1571. doi: 10.1016/j.jrp.2008.07.008

Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. R., & Klerman, G. L. (1976). Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science*, *192*(4238), 489–491. doi: 10.1126/science.1257786

Tassinary, L. G., & Cacioppo, J. T. (2000). The skeletomotor system: Surface electromyography. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 163–199). Cambridge, England: Cambridge University Press.

Tassinary, L. G., Hess, U., & Carcoba, L. M. (2012). Peripheral physiological measures of psychological constructs. *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics*. (pp. 461–488): Washington, DC: American Psychological Association.

Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology*, *73*, 941–959. doi: 10.1037/0022-3514.73.5.941

Waters, W. F., Williamson, D. A., Bernard, B. A., Blouin, D. C., & Faulstich, M. E. (1987). Test-retest reliability of psychophysiological assessment. *Behaviour Research and Therapy*, *25*(3), 213–221. doi: 10.1016/0005-7967(87)90048-9

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Wiggers, M. (1982). Judgements of facial expressions of emotions predicted from facial behavior. *Journal of Nonverbal Behavior*, *7*, 101–116. doi: 10.1007/BF00986872

Ziegler, M., & Brunner, M. (2016). Test standards and psychometric modeling. In A. A. Lipnevich, F. Preckel, & R. Roberts (Eds.), *Psychosocial skills and school systems in the 21st century (pp. 29–55). Theory, research, and applications*. Göttingen, Germany: Springer International Publishing.

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales—Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*, 185–189. doi: 10.1027/1614-0001/a000148

Ziegler, M., Poropat, A., & Mell, J. (2014). Does the length of a questionnaire matter? *Journal of Individual Differences*, *35*, 250–261. doi: 10.1027/1614-0001/a000147

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133. doi: 10.1007/s11336-003-0974-7

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement*, *30*, 121–144. doi: 10.1177/0146621605278814

(Received September 15, 2015; Accepted April 26, 2016)

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix S1: R code EMG Study 1: Percentage of baseline.

Appendix S2: R code EMG Study 1 (except percentage baseline).

Appendix S3: R code EMG Study 2.