IDENTIFICATION OF EMOTIONAL FACIAL EXPRESSIONS IN A LAB AND OVER THE INTERNET

M. VESKER a, D. BAHN b, F. DEGÉ c, C. KAUSCHKE b, G. SCHWARZER a

a Justus-Liebig University Giessen, 10F Otto-Behaghel-Strasse, Giessen, 35394, Germany
b Philipps-University Marburg, 16 Pilgrimstein, Marburg, 35032, Germany
c Max Planck Institute for Empirical Aesthetics, 14 Grueneburgweg, Frankfurt am Main, 60322, Germany

Abstract
Collecting data over the internet is an approach that allows researchers to vastly expand the possible sample sizes of their studies, and enables the study of populations that may otherwise be difficult to access. However, to ensure that data collected over the internet is of the same level of quality as data collected in a lab, the comparability of internet-collected data with lab-collected data must first be assessed for individual areas of research and experimental approaches. To answer the question of whether internet data collection is suitable for experiments involving facial expressions, we conducted a deliberately difficult facial emotion-identification experiment where participants completed the same task either under supervision in our lab, or at an unsupervised location over the internet. Stimuli consisted of sad faces that participants were asked to identify as resembling either anger, fear, or disgust. Regardless of belonging to either the group tested in the lab or over the internet, participants showed highly similar response distributions, while differences between the groups were non-significant and of very low magnitude. We can therefore conclude from our findings that internet data collection is a viable method for experiments requiring the identification of emotional facial expressions, being able to produce similar results to those which can be obtained in a lab.

Keywords: internet, online, verification, faces, emotion, identification.

Introduction

With the emergence of the internet, researchers immediately seized upon the potential for internet-based remote data collection as a tool of enormous potential. Data gathering over the internet can give researchers access to sample sizes in the hundreds and even thousands with relative ease, while allowing for data gathering from participants of very specialized populations. Despite these apparent advantages, questions were soon raised about the compatibility of data collected remotely over the internet with standard laboratory experiments. Were participants truly
performing similarly regardless of the administration method, or could the lower
degree of control over the testing environment in the case of internet experiments
produce different results?

This question of the comparability of data collected in the lab and over the internet
is crucial if one wishes to conduct an experiment over the internet while maintaining compatibility with existing lab-based experimental literature in a given area of research. Thus far, several reviews of this topic broadly seem to indicate that internet data collection is a viable tool for psychological experiments (Birnbaum, 2004; Coles, Cook, & Blake, 2007; Hewson, 2015; Krantz & Dalal, 2000). However, several studies (even those that ultimately conclude in favor of internet and in-lab testing interchangeability) do show differences with respect to some measures.

In the realm of behavioral experiments, Linnman et al. (2006) found that response times were significantly slower on the web compared to the in-lab condition while attempting to replicate the Stroop effect in an internet experiment. Germine and colleagues (2012) meanwhile found age- and sex-matched subject showed significantly lower accuracy scores in the internet versions of the VPAM (Verbal Paired Associates Memory) and FDS (Forward Digit Span) tasks relative to participants tested in the lab. Thus, it is necessary to individually verify the suitability of internet testing for various areas of experimental psychology. To collect data over the internet without prior verification that this data possesses similar qualities to data collected in the lab could result in doubts being cast on the resulting findings in terms of fitting into the existing literature.

Experimental tasks involving social aspects such as emotion recognition from human faces may be particularly influenced by aspects of the testing environment such as the presence of others in the room, the familiarity of the testing location, etc. Thus, we were interested in whether such an experiment could produce truly similar results, regardless of being conducted under supervision in a laboratory, or in the comfort of one’s home.

Faces expressing negative emotions are also especially interesting due to the links between attentional biases towards such expressions and disorders such as depression and various anxiety disorders (Gotlib, Krasnoperova, Yue, & Joorman, 2004; Mogg & Bradley, 2002; 2004). Some studies have also demonstrated results which show promise in using tasks involving negative face perception as tools for potential identification of patients at risk of depression (Joormann, Talbot, & Gotlib, 2007), and even as a therapeutic technique for the treatment of social anxiety disorder (Schmidt, Richey, Buckner, & Timpano, 2009). Being able to provide such therapies remotely over the internet could greatly increase access for patients, but as noted previously, one must first verify the comparability of emotion perception abilities for human faces when assessed remotely from uncontrolled locations versus the lab.

To our knowledge, the only experiment which attempted to verify the suitability of internet data collection for the identification of emotions in a part of the face was the “Reading the Mind in the Eyes” experiment carried out by Germine et al. (2012) as part of their battery of tests. This task, based on the paradigm of Baron-Cohen et al. (2001), involves participants being asked to identify emotions having only seen the eye-region of a target face stimulus. Although Germine et al. (2012)
did not find significant differences between the data of subject who performed this
task in the lab versus online, the specificity of the task with regards to the eye-
region, as well the lack of other studies which could confirm their findings make it
difficult to conclusively say that internet testing is universally suitable for emo-
tion-identification experiments involving whole faces. To further clarify this ques-
tion, we conducted an emotion identification experiment where participants were
shown complete faces, and were asked to identify the emotion out of three possible
choices (fear, anger, and disgust). The task was made deliberately difficult not only
by the possibility of three responses instead of the usual two, but even more so by
showing participants only sad faces, effectively forcing participants to choose the
second most obvious expression, given that sadness was not available as a response.
By using a deliberately difficult task, our goal was to create a situation where even
slight influences of effects such as the presence of an experimenter in the lab con-
dition would produce noticeable differences in our data if they were present.
Negative faces were chosen for this task due our aforementioned interest in their
clinical importance, as well the greater variety of primary emotions possessing a
clear negative emotional valence (sadness, anger, fear, and disgust) relative to posi-
tive faces. This crowding of basic emotions in the negative emotional valence cat-
egory naturally produces greater rates of mistakenly identifying one negative facial
expression for another (Calvo & Lundqvist, 2008; Loughead, Gur, Elliot, & Gur,
2008; Palermo & Coltheart, 2004). This, in turn, may allow for easier perception of
negative emotions as “compound” emotions bearing resemblance to more than one
emotion at once (Du, Tao, & Martinez, 2014), a necessary component of our method.

Method

Participants

154 undergraduate student participants (45 male, 121 female, mean age = 30.3
years) were recruited for the internet-testing condition through an email bulletin
at the University of Marburg. Twenty undergraduate student participants (10 ma-
le, 10 female, mean age = 23.6 years) were recruited for the in-lab testing condition
through an email bulletin at the University of Giessen.

Stimuli

Our stimuli consisted of 56 images of human faces (frontal view, 31 female, 25
male) from the Caucasian category of the Tarr Lab face database (see Figure 1).

The faces were selected from a larger set of 92 faces which were categorized as
sad faces in the Tarr data base, and wore no glasses or headgear. In a preliminary
test with 4 adult participants we presented the 92 faces and asked participants to
indicate which of the faces had expressions resembling other negative emotions in

---

1 Images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of
Psychology, Carnegie Mellon University, http://www.tarrlab.org/ Funding provided by NSF award 0339122
addition to sadness. Fifty-six of the 92 faces were judged as expressing fear, anger, or disgust in addition to sadness by at least 2 of the 4 participants in this preliminary test, and these faces were selected as the stimuli faces for our main experiment.

Procedure

Stimuli faces were presented on a computer monitor in a browser window using the Ilias internet platform. Participants were asked to identify 56 sad faces as expressing the emotions of anger, disgust, or fear by clicking on one of three text labels for those emotions presented below each face on the screen. Thus, our participants were forced to identify not the primary emotion (sadness), but rather the additional/lesser emotional characteristics of each face, increasing the task difficulty, and allowing for a highly sensitive comparison that would not suffer from ceiling effects.

For the internet-testing condition, participants completed the task through the Ilias internet platform at the University of Marburg from their personal computers at uncontrolled locations. For the in-lab testing conditions, participants completed the same task at the University of Giessen during supervised individual appointments (again using the Ilias platform) on a laboratory-provided laptop computer with a 15.6 inch screen.

Results

Selection Percentage Analysis

To examine the possible effects of data-collection methods on our participants’ selection of one of the three emotion categories, a $2 \times 3$ (Condition [lab, internet] $\times$ Emotion [anger, fear, disgust]) item-based repeated-measures ANOVA was performed on the selection percentages across all subjects of each emotion for each face (i.e. what percentage of participants selected each of the three possible emotion categories for each face, see Figure 2). The order of stimuli presentation was
Face Perception in a Lab and over the Internet

Figure 2

Raw Selection Percentages for All 56 Individual Faces by the Selected Emotion Percentages for Fear (Top), Anger (Middle), and Disgust (Bottom) under Each Testing Condition

Note. Box-A highlights an example of a tied leading selection for the internet condition, Box-B highlights an example of a single leading selection in lab conditions, while Box-C highlights an example of a mismatch in selections between the lab and internet conditions. Box-D shows an example of a face stimulus with a matching leading emotion selection across both conditions, but with varying margins between the percentages of the most commonly selected, and the second most commonly selected emotion.

added as a co-variate. The focus of our analysis was the interaction between the two main factors of Condition and Emotion, which would be an indicator if the testing condition influenced the resulting selection of emotion by our participants. The results revealed no significant interaction of condition and emotion, $F(2, 108) = 0.128, p = .88, \eta^2 = .002$, with a calculated power of 0.946 according to a post-hoc analysis in G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007), indicating no differences across testing conditions. No other factors or interactions showed any significant or noteworthy results, indicating similar selection percentages for all faces, regardless of testing condition (see Figure 3).

Matching the Most Common Selections of Emotion Categories across Testing Conditions

Next, we examined how closely the final selection of one of the emotion categories for each face matched across our two conditions (internet and lab). In other words, how often were participants in the two testing conditions in agreement with each other about which emotion was expressed by each face. To answer this question we first had to score each face stimulus in terms of which emotion emerged as
the most commonly selected one across all participants in a given test condition (internet, lab). However, given the subjective nature of the task, we anticipated that some of the faces may produce situations where more than just one emotion would emerge as the preferred choice across participants. Therefore, in terms of absolute agreement in the choice of one of the three emotions, a clear choice for each face stimulus was counted only if the emotion with the highest selection percentage among participants was at least 5% higher than the next closest selection. If the two highest selections for a stimulus were less than 5% apart, the two choices were counted as a tie. For example, in the lab condition, face number 15 (see Figure 2, Box B-solid grey lines) was determined by 60% of participants to be displaying fear, while another 30% selected anger, and the remaining 10% selected disgust, fear was therefore determined to be the leading selection amongst those participants. Meanwhile, face number 6 was described in the internet condition as showing fear by 48% of participants, anger by 46% of participants, and disgust by the remaining 6% of participants, the leading selection among participants was thus determined to be tied between fear and anger (see Figure 2, Box A-dotted black lines). A clear mismatch between the lab and internet conditions was registered for any face for which there was no match between the emotions selected as either the leading choices, or tied for the lead choice. For example, face 35 had disgust emerged as the most common selection in the lab condition, while in the internet condition anger emerged as the most commonly selected emotion at 54%, leading to a mismatch classification for this face between the two conditions (see Figure 2, Box C).

This comparison showed a total of only 6 mismatches among our 56 stimuli (faces 7, 10, 35, 42, 44, and 48), giving a disagreement rate of just 10.7%.

### Selection Certainty across Participants in Each Condition

Finally, we examined whether the testing condition would influence the degree of certainty or agreement among subjects’ selections even if the highest selected emotion was matched overall between the two conditions. For instance, for face

---

**Table 1**

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Selection Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>37.9, 39.6</td>
</tr>
<tr>
<td>Disgust</td>
<td>23.0, 24.5</td>
</tr>
<tr>
<td>Fear</td>
<td>39.0, 35.9</td>
</tr>
</tbody>
</table>

**Figure 3**

Selection Percentages for Each Emotion by the Data-Collection Method for Each of the 56 Faces

*Note.* Error bars represent standard error.
number 39 (see Figure 2, Box D) 65% of participants in the lab condition selected disgust as the emotion on display and 30% selected anger, while in the internet condition only 59% of participants selected disgust and 33% selected anger. Thus, disgust would be the most common selection under both conditions. However, the margin between the selection of fear and anger for this face still differed between the two conditions, which could still show an influence of the testing condition on the degree of certainty in participants’ choices. We therefore analysed this degree of certainty among our participants in the 50 matching selections across both testing conditions. For this comparison we used the differences between the selection percentages of the highest selected emotion and the second-most commonly selected emotion (for each stimulus within each condition). A t-test showed a high degree of similarity between the two conditions in this comparison, $t(98) = 0.297$, $p = .767$, (see Figure 4), and thus the certainty of our participants in their selection choices was very similar under both testing conditions.

**Discussion**

Utilizing internet-based data collection for psychological research can provide obvious benefits in terms of the sheer volume of possible subject participation in psychological experiments, as well as improving access to potential diagnostic and therapeutic techniques involving face perception. The crucial downside of this approach is the lack of control over the testing environment compared to the lab. Experimenters can of course provide their participants with instructions, but have not real way of insuring that the environment for each subject contains no sensory distractions that may impact the experimental performance. Various domains of

**Figure 4**

*Average Differences in the Selection Percentages between the First and Second Most Common Choices*

![Graph showing average differences in selection percentages between the first and second most common choices for Lab and Online conditions.]

*Note.* Error bars represent standard error.
psychological research may be differently impacted by this lack of environmental control (Germine et al., 2012), and thus experimental approaches must first be verified for suitability under these conditions. Tasks involving the identification of human facial emotions are inherently social in nature, and thus may be especially sensitive to differences in the testing environment, particularly the presence of other persons. We therefore decided to perform a short verification experiment to determine if data collected over the internet for such tasks could be used as a viable substitute or complement for lab-collected data. Our results demonstrated the suitability of data collection over the internet in tasks involving the identification of facial expressions by showing no effects of testing method when comparing data collected over the internet to that collected in the lab.

Up to this point, the only related published evidence concerning a part of the face that we are aware of came from the work of Germine et al. (2012). The authors asked subjects to perform the “Reading the Mind in the Eyes” task (Baron-Cohen et al., 2001) in the lab, while others performed the same task over the internet. However, while the eyes do convey a lot of emotional information, the mouth can also play an important role in the identification of certain emotions such as disgust and anger, and thus it is necessary to conduct an internet-lab comparison also for whole faces as in the present study (Bassili, 1979; Kohler et al., 2004; Yuki, Maddux, Masuda, 2007).

We therefore decided to conduct an experiment using a more difficult 3-choice emotion identification task using whole faces. The increased difficulty would also produce data that would be more sensitive to differences resulting from subtle influences in the internet- or lab testing environment, and help avoid ceiling effects. This increased sensitivity would make it easier to detect differences between the testing conditions, and thus would make a more convincing argument for the interchangeability of testing methods should no differences be found as in the study by Germine at al. (2012). Additionally, our usage of whole faces would increase the generalizability of our results to other whole-face paradigms.

In agreement with the findings of Germine at al. (2012), the analysis of the response distribution in our study showed that the testing condition had no significant effect on the participants’ selections for each stimulus, with the interaction between condition and emotion-selection being extremely weak and showing no significance. In terms of the most common responses to each face, our results showed that despite the task difficulty, participants were able to reach a very high agreement level between the two testing modalities, disagreeing only in 6 cases out of 56. Our final analysis examined whether the testing condition had any influence on the degree of agreement or selection-certainty among participants within each individual condition for the 50 stimuli where participants agreed on the final selection in both conditions. We averaged differences between the selection percentages of the first most common, and second most common choices for each stimulus, and compared those averages between the two conditions. Once again, we found no effect of testing condition, indicating that participants demonstrated highly similar levels of certainty in their selections across both testing methods. Our results thus demonstrate that even in a difficult emotion-identification task where the
increased noise of uncontrolled testing environments could theoretically influence the outcome of the experiment, participants tested over the internet still demonstrated highly similar outcomes compared to those tested under controlled laboratory conditions.

Our study was limited in scope to the agreement in emotion-selection between participants, and did not measure response times. An extension of this study based on measuring response times would be necessary to investigate whether the choice of internet- or lab-based data collection might influence the processing speeds of such mental tasks. Furthermore, although the ability to collect larger data sets is one of the primary benefits of online data collection, a larger sample collected in the lab from a population more closely matched to the online sample would be a useful additional aspect of further studies examining the comparability of data collected in the lab versus online.

Conclusions

In summary, we conclude from our study that collecting data over the internet for tasks requiring participants to identify emotional facial expressions yields highly similar data in terms of the participants' selections compared to data collected under laboratory conditions. Thus, internet-based data collection can serve as a useful method for expanding the scope of experiments focused on emotional faces, as well as diagnostic and therapeutic approaches to depression and anxiety disorders.

References


Michael Vesker — research fellow, Department of Developmental Psychology, Justus-Liebig University (Giessen, Germany), D.Sc.
Research area: developmental psychology.
E-mail: michael.vesker@psychol.uni-giessen.de

Daniela Bahn — research fellow, Department of Clinical Linguistics, Philipps-University (Marburg, Germany), M.Sc.
Research area: clinical linguistics.
E-mail: daniela.bahn@uni-marburg.de
Franziska Degé — research fellow, Music Department, Max Planck Institute for Empirical Aesthetics (Frankfurt am Main, Germany), D.Sc.  
Research area: musical abilities’ development.  
E-mail: franziska.dege@ae.mpg.de

Christina Kauschke — head of department, Department of Clinical Linguistics, Philipps-University (Marburg, Germany), D.Sc., professor.  
Research area: clinical linguistics.  
E-mail: kauschke@staff.uni-marburg.de

Gudrun Schwarzer — head of department, Department of Developmental Psychology, Justus-Liebig University (Giessen, Germany), D.Sc., professor.  
Research area: developmental psychology.  
E-mail: gudrun.schwarzer@psychol.uni-giessen.de
Идентификация эмоциональных выражений лица в лабораторных условиях и через Интернет

М. Вескер*, Д. Бан*, Ф. Дэжэ*, К. Каушкэ*, Г. Шварцер*

*Гисенский университет имени Юстуса Либиха, 10F Otto-Behaghel-Strasse, Giessen, 35394, Germany
*Марбургский университет, 16 Pilgrimstein, Marburg, 35032, Germany
*Институт эмпирической эстетики Макса Планка, 14 Grueneburgweg, Frankfurt am Main, 60322, Germany

Резюме

Возможность сбора данных посредством сети Интернет позволяет исследователям значительно расширить выборку данных для своих исследований, а также изучать популяции, доступ к которым в ином случае может быть затруднен. Чтобы удостовериться в том, что качество данных, собранных в сети Интернет, соответствует качеству данных, собранных в лабораторных условиях, необходимо сопоставить полученные обеими путями данные для отдельных областей исследований и экспериментальных подходов. Чтобы проверить, сопоставимы ли данные, собранные через Интернет, с данными, собранными в лаборатории, для исследований, связанных с восприятием выражений лиц, мы провели намеренно усложнённый эксперимент по идентификации эмоций, в ходе которого испытуемые выполняли одну и ту же задачу в одной из двух групп: либо под наблюдением в нашей лаборатории, либо в неконтролируемых условиях посредством Интернета. Стимулы представляли собой фотографии людей, изображающих эмоцию печаль/грусть, которые испытуемые должны были идентифицировать как изображающие либо гнев, либо страх, либо отвращение. Вне зависимости от способа тестирования, испытуемые показали очень схожие результаты между группами. Различия между испытуемыми, протестированными в лабораторных условиях, и испытуемыми, принявшими участие в исследовании посредством сети Интернет, были не значимы и показывали очень малую величину. Таким образом, мы можем сделать вывод, что сбор данных посредством Интернета является эффективным для экспериментов, требующих идентификации эмоций, и позволяет получить результаты, аналогичные тем, которые могут быть получены в лабораторных условиях.

Ключевые слова: Интернет, онлайн, верификация, лица, эмоции, идентификация.

Вескер Михаил — научный сотрудник, кафедра психологии развития, Гисенский университет имени Юстуса Либиха (Гисен, Германия), доктор наук.
Сфера научных интересов: психология развития.
Контакты: michael.vesker@psychol.uni-giessen.de

Бан Даниэла — научный сотрудник, кафедра клинической лингвистики, Марбургский университет имени Филиппа (Марбург, Германия), магистр наук.
Сфера научных интересов: клиническая лингвистика.
Контакты: daniela.bahn@uni-marburg.de
Дэжэ Франзиска — научный сотрудник, кафедра музыки, Институт эмпирической эстетики Макса Планка (Франкфурт, Германия), доктор наук.
Сфера научных интересов: развитие музыкальных способностей.
Контакты: franziska.dege@ae.mpig.de

Каушкэ Кристина — глава кафедры, кафедра клинической лингвистики, Марбургский университет имени Филиппа (Марбург, Германия), доктор наук, профессор.
Сфера научных интересов: клиническая лингвистика.
Контакты: kauschke@staff.uni-marburg.de

Шварцер Гудрун — глава кафедры, кафедра психологии развития, Гисенский университет имени Юстуса Либиха (Гисен, Германия), доктор наук, профессор.
Сфера научных интересов: психология развития.
Контакты: gudrun.schwarzer@psychol.uni-giessen.de