Emotion

The Nonverbal Emotion Assessment Tool (NEAT): An Intercultural Validation

Matthias Pache, Lena Miketta, Rainer Banse, Milena Elchinova, Taufik Mohammad, and Ursula Hess

Online First Publication, February 12, 2026. https://dx.doi.org/10.1037/emo0001651

CITATION

Pache, M., Miketta, L., Banse, R., Elchinova, M., Mohammad, T., & Hess, U. (2026). The Nonverbal Emotion Assessment Tool (NEAT): An intercultural validation. *Emotion*. Advance online publication. https://dx.doi.org/10.1037/emo0001651

AMERICAN PSYCHOLOGICAL ASSOCIATION

Emotion

© 2026 American Psychological Association
ISSN: 1528-3542

https://doi.org/10.1037/emo0001651

The Nonverbal Emotion Assessment Tool (NEAT): An Intercultural Validation

Matthias Pache[1], Lena Miketta[2], Rainer Banse[2], Milena Elchinova[2], Taufik Mohammad[3], and Ursula Hess[4]

[1] Department of Old Age Psychiatry and Cognitive Disorders, University Hospital Bonn
[2] Department of Psychology, University of Bonn
[3] Department of Psychology, International Islamic University Malaysia
[4] Department of Psychology, Humboldt-University of Berlin

This research aimed to validate a newly developed tool for the assessment of emotions. The “Nonverbal Emotion Assessment Tool” (NEAT) is based on schematic facial expressions of emotions and serves to capture both quantitative and qualitative dimensions of common emotions. Study 1 ($N = 126$) was conducted with primary school children (6–11 years), who matched the emotions represented in vignettes to the emotional facial expressions of the NEAT. Although the children’s recognition rates varied across emotions, they were overall moderately accurate. Older children did not perform substantially better than younger children. Study 2 validated the NEAT scales with adult participants from Germany ($N = 102$), Bulgaria ($N = 116$), and Malaysia ($N = 132$). Cross-country intraclass correlations revealed cultural differences in emotion perception, yet the construct validity was high. Comparisons of the two European samples with the Southeast Asian sample yielded a lower level of agreement across countries than the comparison of the two European samples, suggesting more similarities between the German and Bulgarian samples and stronger differences between the European and the Malaysian sample. Together, these findings provide evidence that the NEAT is a useful and valid tool for the assessment of emotions in child and adult samples from different areas of the world.

Keywords: Nonverbal Emotion Assessment Tool, basic emotions, construct validity of the Nonverbal Emotion Assessment Tool, cross-cultural assessment of emotions

In some contexts, it is appropriate to use assessment tools that do not exclusively depend on verbal labels. This is most often the case in developmental research where participants are not yet able to read, but also in intercultural research, where the exclusive use of written verbal labels or the translation process for verbal labels used in response scales may introduce error. The Nonverbal Emotion Assessment Tool (NEAT) was developed as an easy to administer instrument for emotion research. It is designed to allow both the attribution of emotions to others and to be used to describe one’s own emotions and is based on schematic facial expressions.

The NEAT is based on the observation that some emotion expressions are readily recognized from facial expressions in different cultures (Biehl et al., 1997; Ekman, 1972; Ekman et al., 1987; Izard, 1994), despite overall differences in accuracy (Elfenbein & Ambady, 2002). This also the case for young children (Covic et al., 2020; Saarni, 1999), including children on the autism spectrum who received training in this task (Baron-Cohen et al., 2009; Naumann et al., 2023). As this claim is somewhat controversial, we will briefly discuss the issue below.

The NEAT: An Emotion Scale Based on Facial Expressions

One problem in cross-cultural research is the reliance on specific verbal labels. Even between closely related languages, emotion labels do not translate perfectly (e.g., despite the fact that the two languages are closely related, there is no genuine English term for the German word *Schadenfreude*) and the problem can get even more acute across distinct language families (for overviews, see e.g., Hupka et al., 1999; Ogarkova, 2013, 2016, 2021; Wierzbicka, 1999). This is a serious problem for cross-cultural research that is anchored in language and emotion terms. Relatedly, children do have to learn to master linguistic emotion labels (see e.g., Baron-Cohen et al., 2010; Hoemann et al., 2019; Ridgeway et al., 1985), before they can be tested using the same

Maria Gendron served as action editor.

Matthias Pache https://orcid.org/0000-0002-9993-7076
Lena Miketta https://orcid.org/0000-0001-8002-0738
Rainer Banse https://orcid.org/0000-0002-8817-9994
Taufik Mohammad https://orcid.org/0000-0002-7611-6849
Ursula Hess https://orcid.org/0000-0002-6793-7282

All data, study materials, and the code to reproduce their results are available at https://osf.io/3qpju/?view_only=7b8ad19f4e9b45c48870d80b3fce71b8.

Matthias Pache played a lead role in conceptualization, investigation, methodology, project administration, and data collection, a supporting role in formal analysis, and an equal role in resources and writing–original draft. Lena Miketta played a lead role in data curation, formal analysis, software, validation, and visualization, a supporting role in resources and supervision, and an equal role in writing–original draft. Rainer Banse played a lead role in supervision and a supporting role in conceptualization and methodology. Milena Elchinova played a supporting role in investigation and an equal role in data collection. Taufik Mohammad played a supporting role in data collection. Ursula Hess played an equal role in writing–review and editing and data interpretation.

Correspondence concerning this article should be addressed to Matthias Pache, Department of Old Age Psychiatry and Cognitive Disorders, University Hospital Bonn, Venusberg-Campus 1/NPP (Geb. 80), 53127 Bonn, Germany. Email: mpache@uni-bonn.de

scales as are used for adult samples. The goal of the present research was to develop and validate scales that do not exclusively rely on language labels for responses and that capture a subset of emotions that are typically associated with distinct emotion expressions.

The NEAT uses schematic drawn faces because recognition of emotional expressions in photographs can be biased by race, gender, and age of the poser either through stereotypes (e.g., Bijlstra et al., 2014, 2019; Hugenberg & Bodenhausen, 2003; Hugenberg & Sacco, 2008) or facial appearance (e.g., Becker et al., 2007; Hess et al., 2009). Importantly, children in Western societies and non-Western societies such as Turkey, Japan (Cüceloglu, 1970) or Uganda (Kilbride & Yarczower, 1976) readily recognize facial expressions from schematic drawings. In a direct comparison of facial expressions presented as drawings versus photographs, children were better able to recognize sadness, anger, and fear expressions from schematic drawings than from photographs (Brechet, 2017).

Using pictographic emotion portrayals, the NEAT combines the assessment of qualitatively different emotions with the quantitative differentiation of emotion intensity. The use of scalar ratings which assess the perceived intensity of a range of emotions has been advocated as a better means to capture perception (Matsumoto, 2005) than the still common use of forced choice labels. Notably, the use of scalar ratings allows to capture the perception of mixed emotions (Hess & Kafetsios, 2022; Kafetsios & Hess, 2023). This makes them especially suitable to assess the often subtle differences in cross-cultural emotion perception (Matsumoto, 2005). The NEAT can be used as a scalar emotion profile or to select a single, best fitting expression for labeling, depending on instruction.

To date, the Self-Assessment-Manikin (SAM; Bradley & Lang, 1994) and the AffectButton (Broekens & Brinkman, 2013), itself based on the SAM, are the only language-free methods assessing both emotional intensity and quality. The SAM assesses three dimensions: pleasure–displeasure, arousal, and dominance–submissiveness. Notably, based on these dimensions, discrete emotions can be difficult to distinguish. For example, anger and disgust are both highly negative and high in arousal. Assessing discrete emotions using this approach requires a certain level of abstraction. The AffectButton addresses this issue by presenting a dynamically changing iconic facial expression that changes based on input for the three dimensions. This tool, however, requires a computer or a comparable electronic device.

By contrast, the NEAT is based on a subset of interculturally well-recognized facial expressions (anger, fear, joy, sadness, and surprise) with four intensity levels and thus combines the assessment of qualitatively different emotions with the quantitative differentiation of intensity. The NEAT scales can be presented on a computer screen but also be used without electronic devices, when the NEAT faces are printed on cards. Although the NEAT assumes a categorical approach to emotion, we also explore here to what extent it is able to capture mixed emotions, such as the simultaneous feeling of joy and surprise.

The Recognition of Emotion Expressions Across Cultures

In the field of affective science an ongoing lively debate regards the nature of internal emotional states and their link to expressive behavior (see Hess, 2017; Lindquist et al., 2013, for reviews). The main question in this context is whether facial expressions of emotion actually reflect specific and discrete internal states and to what degree this link is biologically based. Positions range from the notion that there is no meaningful link (Bruner & Tagiuri, 1954; Hassin et al., 2013) to the proposal that expressions are determined by emotion programs (Ekman, 1972). Yet, this discussion notwithstanding, there is good evidence that, as already suggested by Darwin (1872/1998), emotional facial expressions have important communicative value. Notably, as Barrett et al. (2019) pointed out, the simple fact that people attribute a certain state to someone on the basis of their expression does not in any way confirm the presence of the state. It does, however, affect the way the observer reacts to that expressions. Simply put, observers treat facial expressions as informative and act in accordance (Scarantino, 2019; Scarantino et al., 2022). As such, it is important to dispose of instruments that allow us to assess the discrete emotional states that individuals attribute to themselves and others. From this perspective, categorical labels likely capture best how observers process emotional behavior, even though the number of such categories is subject of debate, and some emotions, for example, hope, are not expressed facially. Focusing on emotions that are expressed facially, Ekman (1972) proposed a set of six emotions (anger, disgust, fear, joy, sadness, and surprise; this list has now been reduced to five as surprise was dropped, Ekman & Ekman, 2018). As such, for a scale focusing on the measurement of facially expressed states, the so-called basic emotion categories provide a good starting point.

This leads to a final consideration, specifically, to what degree these signals are shared across cultures. Based on the presumed evolutionary origin of facial expressions as signals of human intentions, the assumption was made that these expressions are universally shared. Darwin (1872/1998) and later Ekman and colleagues (Ekman, 1972; Ekman et al., 1987) as well as Izard (1971) made strong claims that emotion expressions are universally recognized. However, the evidence presented by proponents of this view has been severely criticized (e.g., Fridlund, 1994; Russell, 1991, 1994, 1995).

More recently, these two extreme positions have been abandoned in favor of the view that at least some "universal" expressions exist (e.g., Cowen et al., 2021; Jack et al., 2016; Sauter et al., 2010, 2015) but cultural variation in expressive behavior should not be underestimated (e.g., Barrett et al., 2019; Lindquist et al., 2022; Yang & Wang, 2019). This notion is emphasized by constructionist accounts (e.g., Barrett et al., 2019; Gendron et al., 2018), but also in Elfenbein and Ambady's (2002) dialect theory. The latter argues for a "universal language" for emotion expressions with "regional dialects" that differ subtly from each other. That is, emotions can be expressed in similar yet subtly different ways across cultures. This notion implies that even though considerable overlap in emotion judgments across cultures may exist, systematic variation will also be present (Binetti et al., 2022), but agreement between judges from different cultures is still substantial (Elfenbein et al., 2007).

Using facial expressions does thus not resolve all the challenges of scales using language labels, and in the domain of emotion, cross-cultural variance has been observed with language labels and facial expressions alike (e.g., Lindquist et al., 2022). Yet, using scales like those proposed in this article, based on facial expressions, allows for triangulation and can thus compensate the limitations of language-anchored scales. Notably too, Elfenbein and Ambady (2002) found that decoding agreement is higher for groups which are culturally closer—but these still speak different languages and thus the problem of translating labels arises. For this scenario a scale that does not rely on such labels can prove especially useful.

Overview of the Present Research

We report two studies demonstrating the usefulness of the NEAT. Study 1 addressed its applicability to developmental research and assessed recognition rates in a sample of German primary school children (discrete and mixed emotions). Specifically, this sample had little experience with answering written questionnaires and/or were just learning to read. As such the use of a tool that did not rely on verbal labels as response option was considered beneficial.

In order to assess the applicability of the NEAT for cross-cultural contexts, we recruited Bulgarian, German, and Malaysian adult samples in Study 2. Participants in both studies were asked to classify the emotion portrayed in fictious scenarios using the NEAT. For comparison purpose the vignettes used in Study 2 had previously been rated in Study 2a by a separate German sample on 5-point verbal Likert scales.

Study 1

In Study 1, short vignettes were used to describe an emotional situation. The participants' task was to match a facial expression from the NEAT to the emotion experienced by the protagonist in the vignette. A neutral facial expression was used to represent the absence of an emotion.

Method

Participants

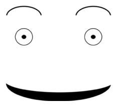
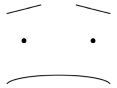
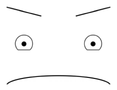
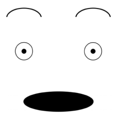
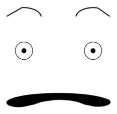
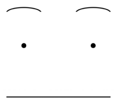
A total of 126 German elementary school children were included in Study 1. Based on Sim and Wright (2005) and Arifin (2024), a minimum of 107 participants is required assuming a minimum acceptable agreement of κ = .40 and expected κ = .70 with a random change agreement of .20 and a power of .80. Ages ranged from 6 to 11 years, with a mean age of 7.87 years (*SD* = 1.30). Three subsamples were formed based on age: youngest (*N* = 35, 13 female, M_{age} = 6.40 years, *SD* = 0.49, range = 6–7 years), middle (*N* = 43, 25 female, M_{age} = 8.15 years, *SD* = 1.03; range = 6–10 years), and oldest (*N* = 48, 32 female, M_{age} = 8.67 years, *SD* = 1.02; range = 7–11 years). Children were recruited from different schools in the city of Cologne. Written consent was obtained from the parent or guardian.

Material

The NEAT. For the pictographic emotion portrayals for the NEAT, only the facial features relevant to emotional expressions were used; images included neither noses nor facial contours. The stimuli were based on the Action Units (in the following: AUs) of the Facial Action Coding System (Ekman et al., 2002). Table 1 shows the images as well as the AUs used; they are also available in the additional online material (https://osf.io/3qpju/overview?view_only=7b8ad19f4e9b45c48870d80b3fce71b8).

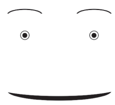
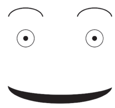
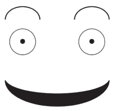
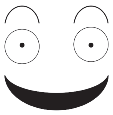
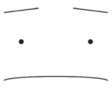
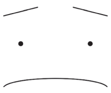
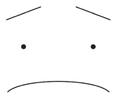
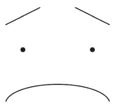
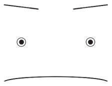
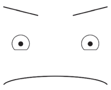
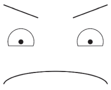
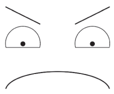
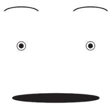
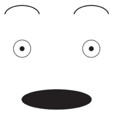
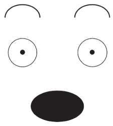
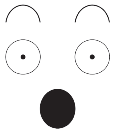
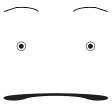
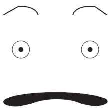
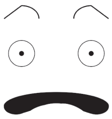
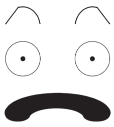
For each emotion portrayed, we developed four different intensity levels ranging from low (Level 1) to maximal intensity (Level 4), as shown in Table 2.

Table 1
AU Represented in the Different Emotion Portrayals and the Final Set of Faces

Emotion	Pictographic emotion portrayal	Action unit
Joy		AU12, lip corner puller, is portrayed by rising edges of the mouth. AU6, cheek raiser, and AU7, lid tightener, are both portrayed in the eyes. (Both eliminated after pretest)
Sadness		AU1, inner brow raiser and AU4, brow lowerer, are both portrayed by eyebrows rising in the middle and lowering on the other edge. AU6, cheek raiser, is portrayed by cheeks going up under the eyes. (Eliminated after pretest) AU15, lip corner depressor, is portrayed by lips lowering on the side.
Anger		AU4, brow lowerer, expressed by brows lowered in the middle. AU5, upper lid raiser, is portrayed by wide opened eyes. AU7, lid tightener, is portrayed by lower eye-lid rising up. AU17, chin raiser, AU23, lip tightener, and AU24, lip pressor, are portrayed by the mouth's two ends going down.
Surprise		AU1, inner brow raiser, and AU2, outer brow raiser, are both portrayed by rising eyebrows. AU5, upper lid raiser, is portrayed by eyes wide open. AU26, jaw drop, and AU27, mouth stretch, are both portrayed by the opening of the mouth.
Fear		AU1, inner brow raiser, AU2, outer brow raiser, and AU4, brow lowerer, are portrayed by rised eyebrows inclined on the edges. AU5, upper lid raiser, is portrayed by wide opened eyes. AU25, lips part, AU26, jaw drop, and AU27, mouth stretch, are portrayed by the open mouth. AU20, lip stretcher, is portrayed by the mouth stretching and lowering on each side.
Neutral		No AUs are portrayed.

Note. AU = action units.

Table 2
Intensity Levels for Each Emotion Portrayal

Depicted emotion	1	2	3	4
Joy				
Sadness				
Anger				
Surprise				
Fear				

Note. Intensity Level 2 is used for Study 1.

The neutral face, signifying no emotion, corresponds to Level 0 (see Table 1). Adding neutral to the four intensity levels resulted in five intensity levels for each emotion.

For Study 1, only intensity Level 2 was used. The more complex differentiated rating scale, which comprises all five intensity levels (0 = *neutral*, 1 = *low intensity* to 4 = *maximum intensity*), was used for Study 2 below.

Vignettes. Twelve vignettes evoking simple emotions were created for Study 1: Two vignettes for each simple emotion assessed here (anger, fear, joy, sadness, surprise) and two vignettes for the neutral condition. An example of a vignette, translated from the German original and portraying fear is "Lukas has done something wrong at school. When he comes home, his father says: 'Your teacher has just called. We need to talk to you.' How does does Lukas feel?"; an example of a vignette portraying sadness is "Laura is celebrating her birthday and none of her friends are coming. How does Laura feel?". Additionally, since the NEAT can be used not only for discrete emotions such as anger, fear, joy, sadness, and surprise, but also to assess complex or mixed emotions, 12 additional vignettes portraying mixed emotions were created for an exploratory substudy. An example of such a vignette, translated from the German original and portraying anger and surprise is "Laura accidentally finds out that her girlfriend has passed on a secret she told her. How does Laura feel?" All vignettes used in Study 1 are available in the additional online material (https://osf.io/3qpju/files/nrwmu?view_only=7b8ad19f4e9b45c48870d80b3fce71b8).

Procedure. Data for Study 1 were collected using pen and paper in a classroom setting. The emotion vignette was presented and participants were asked to match the (Level 2 intensity) emotional face from the NEAT. We used three tasks. The condition "vignette plus emotion label" was adopted from a study by Ekman and Friesen (1971) and aims to ensure that participants understood the intended emotion correctly.

1. Vignette only. For example, children were told: "Lukas is giving a birthday party and all his friends are coming" and were asked to match one or more expressions to the emotion vignette.

2. Vignette plus emotion label. For example, children were told: "Lukas is giving a birthday party and all his friends are coming. He is happy about that" and were asked to match one or more expressions to the emotion vignette plus the corresponding emotion label.

3. Label only. Children were only told the label, for instance, "happy," and asked to match one or more expressions to this label.

The first part of experiment consisted of one of the vignettes tasks, either the vignettes only or the vignettes plus label; this was a between-participants condition. Children were presented with 12 different vignettes with one target emotion, which were the same for all three age groups.[1] These were the same for all children. They then completed the same task for vignettes/labels that referred to two or, in a few instances, three emotions. The specific vignettes and emotions differed between age groups; each group was asked to rate seven vignettes eliciting more than one basic emotion; the vignettes were not the same in all subgroups. All participants completed the label-only task following the vignettes task.

Participants were allowed to take as much time as they wanted for their judgment. The vignettes and verbal labels were always read aloud but were also presented as text in the questionnaire. All participants were told that there were no correct or incorrect responses and that each person could have very different opinions about emotions. Each participant judged the entire set of emotion vignettes for their age group. Participants were explicitly told that the emotion expressions referred to the emotion that the protagonist in the vignette feels.

Statistical Analyses

We used weighted Cohen's κ with Cicchetti–Allison weights as a recognition index. In case of double, triple or quadruple rating, each response was weighted with .5, .33, or .25, respectively, that is, in order to avoid a distortion of the results and an under- or overestimation of the correctness of the answers, in the case of multiple answers we divided by the number of answers given. In 87 cases (5.8%) of all ratings, double ratings were used, triple rating in nine cases (0.6%) and quadruple rating in only three cases (0.20%). Forty-one (32.54%) of the children used multiple rating at least once. The strength of agreement was described using the nomenclature of Landis and Koch (1977).

Results and Discussion

Vignettes With One Target Emotion

To examine the recognition and error rates when presenting vignettes with one target emotion, confusion matrices were calculated (see Table 3). Errors are presented in the off-diagonal entries and correct recognition rates in the main diagonal.

For vignettes only, the overall recognition rate was .66 (range = .47–.97). Recognition rates were highest for joy (97%) and anger (74%). Accuracy using NEAT faces was somewhat weaker for fear (see Table 3). According to Landis and Koch (1977), the range of recognition rates from .47 to .97 indicates a moderate to almost perfect agreement. When excluding fear results, the remaining recognition rates range from substantial to almost perfect. Recognition rates for vignettes with labels were overall slightly higher, at .78 (range = .49–.96). In particular, anger, neutral, sadness, and especially surprise (with label: .77, without label: .49) were better recognized. However, adding a label did not improve fear recognition. Children were best at assigning an expression to a label (.81, range = .61–.96). This was especially the case for anger (label: .95, vignette with label: .89, without label: .74) and surprise (label: .87, vignette with label: .77, without label: .49). However, the neutral expression was somewhat less well-recognized in this condition (label: .64, vignette with label: .72, without label: .60).

Table 3
Emotion Recognition Rates: Cohen's κ Coefficient for Three Different Conditions

	Decoded emotion					
Intended emotion	1	2	3	4	5	6
Vignette only						
1. Joy	**.97**	.00	.01	.00	.00	.01
2. Sadness	.00	**.70**	.04	.02	.05	.02
3. Anger	.01	.10	**.74**	.01	.01	.02
4. Surprise	.06	.02	.00	**.49**	.03	.07
5. Fear	.01	.31	.01	.06	**.47**	.03
6. Neutral	.11	.06	.01	.01	.03	**.60**
Vignette with label						
1. Joy	**.96**	.00	.00	.00	.00	.00
2. Sadness	.00	**.83**	.11	.01	.11	.03
3. Anger	.01	.17	**.89**	.01	.01	.02
4. Surprise	.16	.01	.12	**.77**	.02	.12
5. Fear	.00	.33	.01	.05	**.49**	.04
6. Neutral	.05	.04	.03	.13	.08	**.72**
Label only						
1. Joy	**.96**	.00	.00	.00	.01	.01
2. Sadness	.00	**.83**	.00	.01	.11	.00
3. Anger	.01	.01	**.95**	.00	.00	.01
4. Surprise	.04	.02	.00	**.87**	.03	.00
5. Fear	.02	.24	.00	.02	**.61**	.01
6. Neutral	.06	.04	.02	.03	.11	**.64**

Note. $N = 126$. The hit rates for each emotion are in bold.

These results may reflect differences in the familiarity of the labels or differences in the understanding of the vignettes. We therefore compared recognition rates for the three age groups. Table 4 shows emotion recognition for all emotions in the three conditions as a function of age. Figure 1 shows mean recognition across all emotions as a function of age.

Overall, the youngest group performed quite similarly to the two older groups. In fact, overall performance of the three groups was almost identical. What is different is the performance for certain tasks: For instance, both the middle and the oldest groups had higher recognition rates when labels were provided (with vignettes or alone) than when the vignette was shown alone. For the youngest group the opposite pattern obtained. This suggests that this group may still struggle with verbal emotion labels. The data for individual emotions shown in Table 4 suggests that this is not the case for joy and sadness. This is in line with findings by Michalson and Lewis (1985) that these labels are acquired early. The observation, however, that the youngest group may have struggled with verbal emotion labels stands in contrast with the label superiority effect found in earlier studies: children from preschool to second grade found it easier to rate emotional vignettes on the basis of emotion labels (e.g., "happy," "sad," "surprised") than on the basis of facial expressions on photographs (Camras & Allison, 1985). A label superiority effect was also found in a study where preschool children from 2 to 7 could categorize emotional expressions more easily with

[1] To reduce reliance on strategies that are based on the assumption that there is one face per story so all faces have to be used, two vignettes were used for each of five emotions and two neutral vignettes that presumably did not describe any emotion. With two vignettes per condition, this strategy is less efficient, but the number of vignettes is still adequately low for the youngest age group. In all cases, the children first completed the task with vignettes/labels that referred only to one emotion.

Table 4

Comparison of Emotion Recognition Rates Across Age Groups When Presenting One Target Emotion Vignettes With or Without Verbal Labels or Label Only

Target emotion	Youngest			Middle			Oldest		
	With	Without	Label	With	Without	Label	With	Without	Label
Joy	0.92	0.89	0.91	0.97	1.00	0.96	0.99	0.99	1.00
Sadness	0.78	0.80	0.71	0.92	0.54	0.89	0.78	0.75	0.85
Anger	0.82	0.98	0.93	0.92	0.50	0.90	0.90	0.76	1.00
Surprise	0.52	0.90	0.66	0.83	0.38	0.93	0.90	0.31	0.96
Fear	0.43	0.35	0.38	0.63	0.54	0.83	0.40	0.49	0.58
Neutral	0.64	0.65	0.42	0.71	0.49	0.60	0.79	0.64	0.83

Note. With label: youngest group $N = 23$, middle group $N = 30$, oldest group $N = 28$. Without label: youngest group $N = 12$, middle group $N = 13$, oldest group $N = 20$. Label only: youngest group $N = 35$, middle group $N = 43$, oldest group $N = 48$.

verbal labels than without (Russell & Widen, 2002). In the present study, the youngest children apparently understood the emotional meaning as expressed by the face and were able to match it with the emotional meaning of the vignettes but were confused and not helped when the vignette was accompanied by a label.

Overall, when considering a task in which only one emotion label is relevant, the NEAT was shown to be a useful instrument. Notably, the use of a nonverbal scale made it possible to discern that the youngest age group may not have yet acquired emotion labels fully.

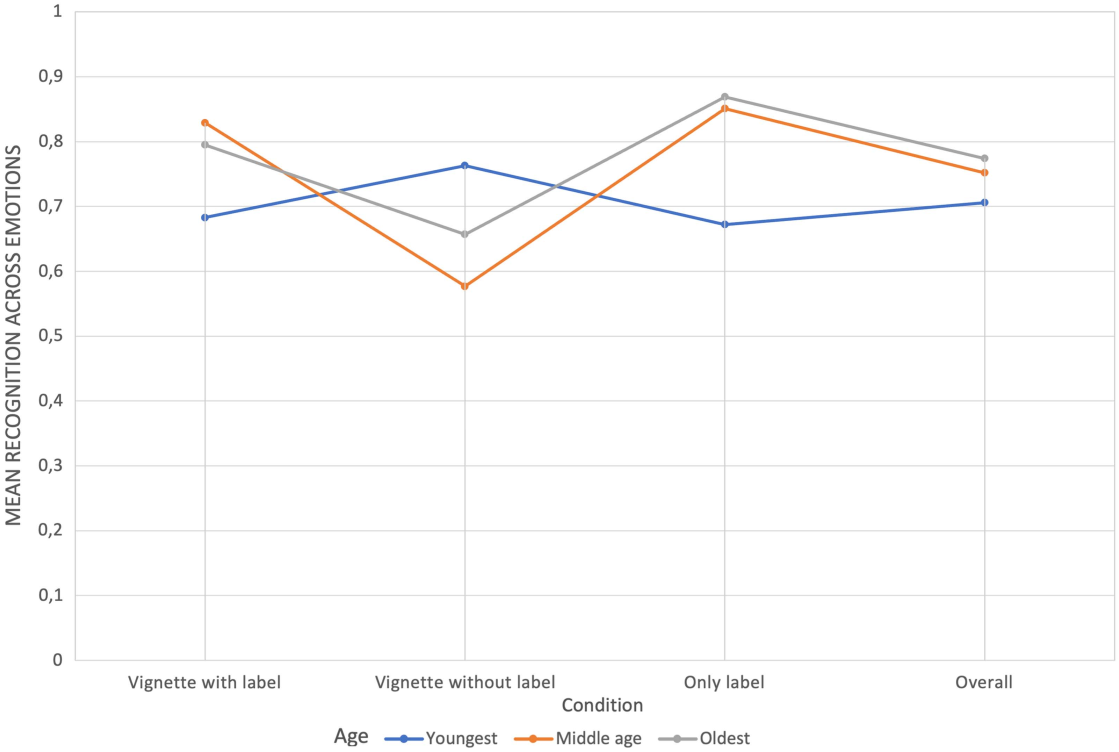
Vignettes With Multiple Target Emotions; With and Without Additional Emotion Label

As can be seen in Table 5, overall performance was relatively low when the vignette (with or without additional label) referred to more than one emotion. However, some combinations were better recognized. Specifically, joy and surprise as well as anger and sadness and joy and sadness were well-recognized. By contrast, the combinations joy and fear and fear and surprise were quite poorly identified, both by the youngest and oldest subgroups. Since different fictitious scenarios were used for the target emotion combinations in each between-condition, these findings can be classified as relatively independent of the vignettes' exact contents.

Comparing age groups across conditions is hampered by the fact that, for this between-groups comparison, Ns are small and not all cells are filled. Nonetheless, the results provide some preliminary evidence. Specifically, for the youngest age group, adding a label to the vignette sometimes helped and sometimes hindered. Overall, across all emotion categories, this group

Figure 1

Mean Recognition Across All Emotions as a Function of Age and Condition



Note. See the online article for the color version of this figure.

Table 5
Comparison of Emotion Recognition Rates When Presenting Multiple Target Emotion Vignettes With or Without Verbal Labels

	Youngest			Middle			Oldest		
Target emotion	With label	Without label	Label	With label	Without label	Label	With label	Without label	Label
Fear, surprise	.31	.45	.35	n.a.	n.a.	n.a.	.38	.38	.40
Fear, anger	.46	.31	.41	.77	.34	.64	n.a.	n.a.	n.a.
Anger, sadness	n.a.	n.a.	n.a.	.88	.46	.76	.70	.35	.57
Anger, surprise	n.a.	n.a.	n.a.	.70	.15	.54	.72	.31	.55
Surprise, sadness	.55	.43	.51	n.a.	n.a.	n.a.	.71	.12	.48
Joy, surprise	n.a.	n.a.	n.a.	.88	.55	.78	.79	.45	.67
Joy, fear	.51	.40	.47	n.a.	n.a.	n.a.	.54	−.13	.27
Joy, sadness	.74	.69	.73	.88	.22	.68	n.a.	n.a.	n.a.
Joy, anger	.68	.58	.65	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Joy (*Schadenfreude*)	n.a.	n.a.	n.a.	.90	.17	.68	n.a.	n.a.	n.a.
Sadness, fear	.53	.77	.62	n.a.	n.a.	n.a.	.73	.45	.63
Surprise, anger, Sadness 1	n.a.	n.a.	n.a.	.70	.27	.57	n.a.	n.a.	n.a.
Surprise, anger, Sadness 2	n.a.	n.a.	n.a.	.85	.15	.61	n.a.	n.a.	n.a.
Total	.54	.52	.53	.82	.29	.66	.65	.27	.51

Note. With label: youngest group $N = 23$, middle group $N = 30$, oldest group $N = 28$. Without label: youngest group $N = 12$, middle group $N = 13$, oldest group $N = 20$. Label only: youngest group $N = 35$, middle group $N = 43$, oldest group $N = 48$. For exploratory purposes, in the middle group, we examined the assessment of a nonbasic emotion (*Schadenfreude*, Vignette 18) in one instance and the combination of three basic emotions in two instances with Vignettes 15 and 20. n.a. = not available.

performed about equally well across conditions with accuracies ranging from .52 to .54.

The picture is quite different for the older age groups. For both older groups adding a label always improved recognition—and sometimes dramatically. Nonetheless, when presented with the label alone, these groups performed less well than when a vignette with label was presented. As such, it is not the case that they base their ratings of the vignette with label solely on the label. Rather it seems that the label helps them to understand the vignettes, whereas either alone is more confusing. These findings also suggest that the labels have a different worth for the younger children.

Overall, these findings support the usefulness of the NEAT in situations where more than one emotion is relevant. However, it should be noted that vignettes referring to mixed emotion situations seem to pose particular challenges especially for the older children. Future research might investigate the concept of mixed emotions from a developmental perspective.

Study 2a: Selection of Vignettes for Study 2

Study 1 included only German children. In Study 2, adult participants from Germany, Bulgaria, and Malaysia were enrolled as examples of different country populations yet not as representatives for a strong intercultural comparison.

The goal of Study 2a was to select the stimulus material for Study 2. The stimuli in Study 2a consisted of 48 emotional vignettes that were intended to evoke a specific emotion, but, unlike in Study 1, with varying intensity. A small German sample was recruited to rate the emotional experiences of the protagonist in a vignette on five different number-based Likert scales. These number-based ratings were also used at a later point to validate the NEAT ratings collected in Study 2.

Method

Participants

A total of 24 German-speaking participants (22 women and two men, with a mean age of 23 years [$SD = 4$, range = 18–32]) were recruited to participate in Study 2a. The majority of the participants were psychology students at the University of Bonn. Individual participants were recruited via social media channels. Students enrolled at the department of psychology received course credit for their participation. Acceptance of the consent form collected via the SosciSurvey survey tool was a prerequisite for participation.

Stimulus Material

A total of 48 vignettes were included in Study 2a, eight evoking five discrete emotions—anger, fear, joy, surprise, sadness—and eight evoking no emotion (neutral condition). Thirteen of the 48 vignettes were adopted from Wingenbach et al. (2019) in slightly modified (e.g., shortened) form. The remaining 35 vignettes were developed by the authors; they were different from the vignettes used in Study 1 and did not contain any emotion labels in any condition. The following is an example: "The neighbors had another party yesterday. The music was so loud that I couldn't sleep at all." For each emotion, we created vignettes varying in intensity levels. Intensity level was determined for each emotion with four different intensity levels ranging from 1 = low arousal to 4 = maximum arousal. All vignettes for Study 2a were written in German. There were two vignettes for each intensity level.

Procedure. The study was conducted online with SosciSurvey (Leiner, 2019). Participants were asked to rate the emotional experience of the protagonist in all emotional and the neutral vignettes on five different verbal Likert scales. These 5-point scales referred to the emotions anger, fear, joy, sadness, and surprise and ranged from 0 (*not at all*) to 4 (*very strongly*), and to the neutral condition. That is, neutrality was indicated by a zero rating across all scales, as a correct identification of the neutral condition required all emotions to be rated as zero/absent. There was no time limit for the judgment task. Each vignette was presented on a separate page in random order.

The selection of vignettes for use in Study 2 was based on two criteria: First, vignettes were considered appropriate if they had higher scores on the scale measuring their target emotion than on other scales. Second, vignettes were selected if they elicited emotions with varying intensities, resulting in meaningful variability in intensity levels.

Statistical Analyses

The results of Study 2a were analyzed descriptively. The mean values of all vignettes on all scales were calculated. The mean values of the 36 best-suited vignettes were then plotted on a graph. This allowed an overview of how the vignettes were scored on the target emotions as well as nontarget emotions and to identify the variability of intensity levels on each scale.

Results and Discussion

Figure 2 shows the mean scores of the 36 best-suited vignettes on each of the five emotion scales; the vignettes are shown in the additional online material (https://osf.io/3qpju/files/jd3tm?view_only=7b8ad19f4e9b45c48870d80b3fce71b8). The 12 vignettes not selected for further use in the study had either too low/high mean scores on the target/nontarget scales or too little variability with regard to the scale range. As can be seen, the selected vignettes were consistently scored higher on the target scale than on any other scale. Nevertheless, the scores of the vignettes were also relatively high on nontarget scales. For example, most vignettes also had moderate or even relatively high values on the surprise scale. One reason for this finding could be that situations that evoke strong emotions usually occur unexpectedly and suddenly. As such, the experience of a strong emotion cannot be clearly decoupled from the experience of surprise. Similar results were found for vignettes intended to elicit negative emotions. For example, a high score on the target emotion sadness was also associated with a moderate score on the affect fear. Despite these limitations, the comparison of the mean values reveals that the target emotion is the emotion that is most strongly evoked by the respective vignette.

The requirement of variability in the intensity levels of each scale was assessed through the standard deviations (*SD*) of the ratings. The surprise scale showed the highest *SD* (1.01), while the sadness scale displayed the lowest *SD* (0.28). The relatively high *SD* for the surprise scale may reflect a broader range of ratings, potentially influenced by a confound between surprise and high general arousal, as high arousal often accompanies the experience of intense emotions. For scales capturing target emotions, the vignettes consistently scored moderately to relatively high (scores of 2–4), resulting in narrower SDs (smallest $SD_{\text{joy}} = 0.20$, largest $SD_{\text{fear}} = 0.56$). As such, the vignettes were deemed adequate for use in Study 2.

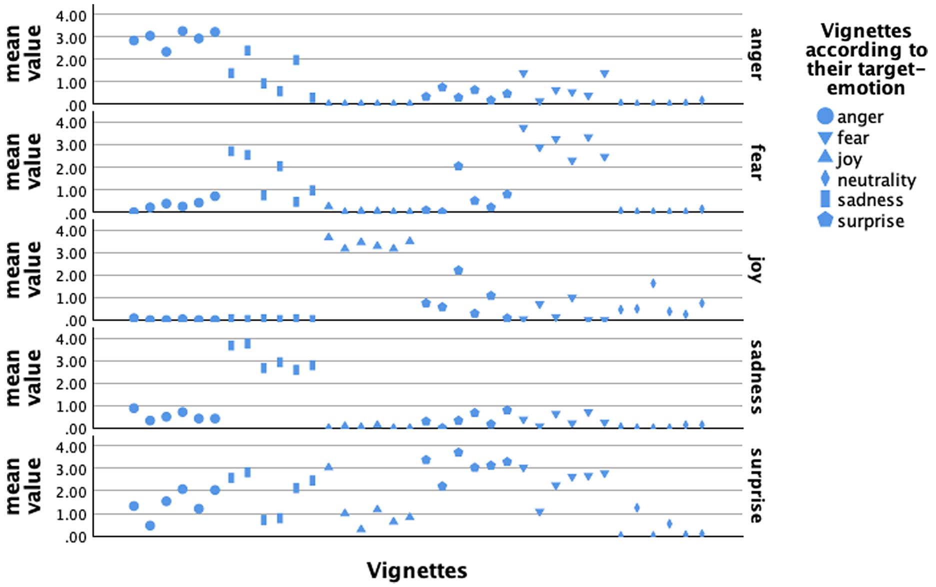
Study 2: Construct Validity and Intercultural Applicability of the NEAT Scales

Study 2 aimed to assess the validity of the NEAT. Specifically, we assessed whether the NEAT consistently captured the theoretically intended construct. For this, the NEAT ratings were compared with the verbal Likert-scale ratings from Study 2a under the assumption that verbal Likert scales are the gold standard, as it were, for assessing emotion attributions.

To examine whether participants from different areas of the world rated the emotion portrayals of the NEAT similarly, participants from three different countries—Germany, Bulgaria, and Malaysia—were asked to rate the 36 vignettes retained from Study 2a with the NEAT.

Figure 2

Mean Values of the 36 Selected Vignettes on the Five Likert Scales



Note. See the online article for the color version of this figure.

This sample of countries allows for both an intra-European and an intercontinental comparison. These intercultural samples are interpreted here as exemplars of different country populations and not as representatives for a strong cross-cultural comparison. Rather, the sample of Study 2 was intended to include people speaking languages from distinct language families (Indo-European and non-Indo-European) and living in different cultural contexts. This was made possible by including participants speaking Bulgarian (Slavic language, Indo-European), German (Germanic language, Indo-European), and Malay (Austronesian language, non-Indo-European).

Method

Participants

The required sample size was determined following Walter et al. (1998) who outline power requisites for studies assessing reliability using the ICC (ρ). Assuming a ρ_0 (minimally acceptable level of concordance) of .5 as well as a ρ_1 (concordance assuming Hypothesis 1) of .8 and an *n* (number of measurement repetitions/raters per participant) of 2, a minimum of 22 participants per group would be required due to the large difference between ρ values.

The sample of German participants included a total of 102 participants (77 women, with a mean age of 28 years [*SD* = 11, range = 18–78]; 47% of the sample was between the ages of 18 and 24; approx. 25% was over 30 years old). Most of the participants were enrolled as psychology students at the University of Bonn at the time of data collection and received course credit for their participation. A minority of participants were recruited through various social media channels.

The Bulgarian sample consisted of 116 participants (95 women with a mean age of 30 years [*SD* = 11, range = 16–56]; 46% were between the ages of 18 and 24 years and approximately 33% were older than 30 years). Participants were recruited through social media channels.

The Malaysian sample consisted of 132 (108 women with a mean age of 28 years [*SD* = 9, range = 18–60]; 52% of the sample was between 18 and 24 years and 21% of the sample were over 30 years old). The recruitment was done via classroom after a lecture session. As reflected in the age range of the participants, the Malaysian sample did not consist of students only. The experimenters also approached family members and colleagues who were willing to take part in the study. Thus, although the background of the Malaysian participants may have varied, including those familiar and not familiar with the cultural background of their German and Bulgarian counterparts, the vast majority of Malaysian participants probably had access to much of the same kind of Western information as their European counterparts.

Stimulus Material and Procedure

The online questionnaire consisted of the 36 vignettes (i.e., six vignettes per five target emotions plus six vignettes for the neutral condition) as well as the NEAT scales. For use with the Bulgarian and Malaysian samples, the original German vignettes were translated into the target languages (see the Bulgarian and Malaysian vignettes in the additional online material, https://osf.io/3qpju/files/c8jf5?view_only=7b8ad19f4e9b45c48870d80b3fce71b8).

As the content of the vignettes was simple and straightforward, we did not use back translation. The fourth author (Bulgarian native speaker) translated the German vignettes into Bulgarian and English and the fifth author (Malay native speaker) translated the English vignettes into Malay. Since the development of the stimulus material was based on a German sample, the Bulgarian and Malaysian vignettes were not locally developed and validated stimuli. Ideally, vignettes should have been created in all three countries and rated by members of all three countries. As such, the present results cannot be interpreted as performance under optimal conditions, but only as a lower threshold for cross-national agreement.

Participants from all samples were instructed to read the vignettes and to rate the quality and intensity of the protagonist's emotion using the NEAT. The participants could use the NEAT as a scalar emotion profile. The vignettes were each presented on a separate page and in random order. As in Study 2a, the participants could proceed at their own pace.

Statistical Analyses

Intraclass correlations (ICCs) between NEAT ratings and verbal scale ratings from Study 2a as well as between the NEAT ratings across countries were calculated to analyze the agreement of the ratings. The ICC was used as a measure of the extent of judgment concordance. Compared with the Pearson correlation, ICC can be conceived as a more stringent concordance measure because it is sensitive to mean level differences between raters, so that not only variance inequality but also inequality of the mean values is detrimental to concordance (Nunnally, 1978; Shrout & Fleiss, 1979).

Testing the construct validity of the NEAT via this concordance measure can be considered a rather conservative procedure, since the agreement in the understanding of the NEAT scales across cultures is based on the premise that the perceived emotions elicited in the protagonist are similar. In the (likely) case of cultural differences in the emotions expected to result from the situations described in the vignettes, the approach underestimates validity.

In addition to the total ICC score, the scores for the individual scales were determined. For example, the ICC for the anger scale was calculated by including all answers on the anger scales across all vignettes. Since the results of the present study indicated that the ICCs for surprise were consistently lower than for the other included emotions, a total score adjusted for surprise was calculated, by eliminating all responses on the surprise scales across all vignettes. To classify the level of the ICC, we followed Cicchetti's (1994) guidelines, according to which an agreement is classified as poor below .40, as fair between .40 and .59, as good between .60 and .74, and as excellent between .75 and 1.00.

We calculated two-way random (absolute agreement) ICCs. Two-way, since each case/setting item was judged by all raters, allowing for differentiation between rater and error variance. Random, since the raters represented a random sample of all possible raters and thus the ICCs were interpreted as a measure of interrater reliability, which indicates the extent to which individual raters from both cultural groups gave consistent ratings for identical scenarios. Absolute agreement was chosen because it is the more conservative method and mean differences between raters were considered part of the error variance, which then reduced reliability. Since the analysis did not include values aggregated across raters (in this case, countries), the ICCs were calculated on the basis of individual assessments rather than aggregated group values at the country level.

Rank correlations were used to compare the relative order of emotion ratings between countries based on the average ratings per vignette. To determine the rank correlations, only the intended emotions were included (i.e., if the vignettes targeted the emotion anger, the anger scores were taken into account, likewise the fear scores were included if the vignette targeted the emotion fear, etc.). The rank correlations were calculated across all emotion vignettes using averaged scores.

Results and Discussion

Descriptive Statistics

Figure 3 shows the mean scores of the six vignettes on their target emotion. Including all four samples (i.e., the three country samples and the data from Study 2a—the verbal Likert scale sample) allows for a direct comparison of the vignette means and their rankings across countries/studies. The results indicate that the mean values were very similar across samples.

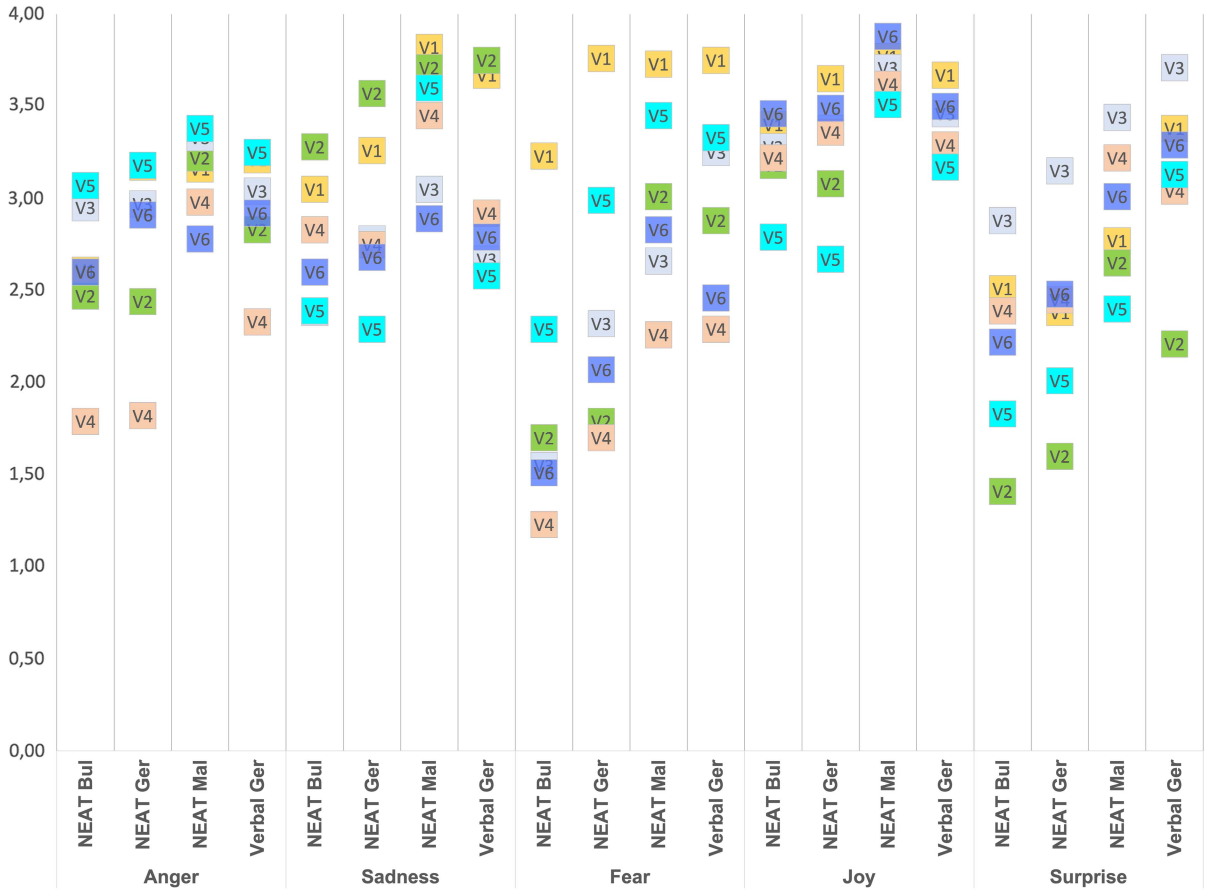
Exceptions are the slightly lower means of the Bulgarian sample on the fear scale and the slightly higher means of the verbal Likert-scale sample on the surprise scale. The vignette means of the Malaysian sample also deviated slightly upward from the rest of the samples on three of six emotions. There was a strong similarity in ranking patterns of the vignettes across the four samples, indicating that even though there were country-specific offsets in the mean intensity ratings, the ratings of the rank order of intensity across vignettes were very similar.

ICCs

Both the comparison between the verbal scales ratings from Study 2a and the NEAT ratings from Study 2 as well as the comparison of the NEAT ratings among the three countries can be used to determine the validity of the NEAT for the assessment of certain attributed emotions. With .42–.68 (with individual values for NEAT) and .56–.97 (with means for both measures) respectively, the ICCs as a measure of agreement between the verbal and NEAT scale as well as the comparison of NEAT scores across countries represented a fair to good effect (.42–.68) and a good to excellent effect (.56–.97; see Cicchetti, 1994 for the interpretation of the scores). When calculated separately by emotion scale, the range was between .23 and .83 (with individual values for NEAT) or .36–.99 (with means for both measures), whereby weak effects below .40 occurred for only two emotion scales (Cicchetti, 1994).

Figure 3

Comparison of the Mean Values of NEAT Scales Bulgarian, German, Malaysian, and Verbal Likert Scales per Emotion



Note. NEAT = Nonverbal Emotion Assessment Tool; V = vignette; number = vignette code; Bul = Bulgarian; Ger = German; Mal = Malaysian. See the online article for the color version of this figure.

The lowest values were consistently found for surprise and fear ratings. A high consistency across measures was found for anger, sadness, and joy ratings (where the minimum ICC reached .41). In two country comparisons, the fear scale scored below .40; the remaining effects can be classified as fair to good according to Cicchetti (1994). The above findings indicate that the NEAT (across the countries included) and the verbal Likert scales consistently captured the construct of attributed emotions and that large effects in terms of construct validity can be assumed for most emotion categories.

ICCs: Individual Values for NEAT

The Likert versus NEAT scale comparison included averaged scores for the verbal Likert scales and individual values for the NEAT. When comparing across countries, the individual value was only used for one country (see the Statistical Analyses section). The ICC value of .42 indicates a fair agreement between the raters from Bulgaria and Malaysia and thus a moderate interrater reliability. All total ICC scores can be classified as large effects. The total ICC scores of the cross-country comparison varied from .42 ($p < .01$) for the intercontinental NEAT comparison between Malaysia and Bulgaria to .68 ($p < .01$) for the European NEAT comparison between Bulgaria and Germany (see Table 6). The ICC value of .42 indicates fair agreement between the raters from Bulgaria and Malaysia and thus a moderate interrater reliability. The correlation between NEAT values in Bulgaria and Germany was .68, which can be classified as a good interrater agreement. The interrater reliability for verbal Likert and NEAT scales—calculated via the total score—was lowest for the comparison between Germany and Malaysia (.47, $p < .01$), medium–high for the comparison between Bulgaria and Germany (.55, $p < .01$) and highest for the comparison of both German samples with .66 ($p < .01$).

The same ranking was obtained for the scales calculated separately for each emotion. This ranking indicated that complete culture independence of the NEAT cannot be assumed. Even though ICCs do not provide direct proof of construct validity, a high level of agreement between raters (like ICC ≥ .75) points to a consistent interpretation of emotion expression across cultures. This supports face validity and consensus validity more than predictive validity. Since the level of agreement in the intercontinental comparison still indicated a large effect, the construct validity can nonetheless be classified as sufficient for this comparison. The range of .41–.83 of the ICCs for the individual emotion scales anger, joy and sadness indicated moderate to excellent interrater reliability. The interrater agreement was lower for the surprise and fear subscales. The ICCs for the joy scale was by far the highest (.55–.83, ps < .01), whereas it was lowest for the surprise scale (.23–.48, ps < .01). Thus, in this study, and unlike in Study 1, the NEAT icons for surprise seemed on face value to be less valid than the icons for the other emotions. However, the findings from Study 2a revealed that situations that elicited strong emotions and thus implied high emotional arousal were often also surprising. This confound could have resulted in the lower observed validity of the surprise icon. Therefore, the total score was considered again after eliminating the surprise scales.

ICCs: Means for Both Measures

When only mean scores were included in the calculation, the construct validity was generally higher than for individual values. The level of agreement ranged from .36 to .99, which consistently indicated a large effect. The European NEAT comparison led to estimates greater than or equal to .95 ($p < .01$) for both the total score and all separate scales. The intercontinental NEAT comparison showed lower validity of the total scores as well as the scales determined for the emotions individually compared with the European comparison (range between .36 and .81, ps < .01), with most ICC values ranging from moderate to excellent effects (Cicchetti, 1994). The total scores for the NEAT versus verbal Likert scale comparison—as with the individual values—were highest when only German samples were included and lowest when an intercontinental comparison was made. Compared with the individual ICC values, the total scores showed much higher reliability, with ICCs ranging from .63 to .88. The recognition rates

Table 6
Intraclass Correlations With Individual Values for NEAT

Variable	Total	Anger	Sadness	Fear	Joy	Surprise
Verbal Ger—NEAT Ger	.66** [.633, .692] (.73** [.721, .739])	.75** [.689, .788]	.69** [.671, .707]	.66** [.635, .676]	.83** [.816, .839]	.42** [.191, .581]
Verbal Ger—NEAT Bul	.55** [.523, .574] (.61** [.603, .623])	.62** [.583, .659]	.61** [.588, .627]	.49** [.456, .516]	.72** [.707, .736]	.33** [.143, .468]
Verbal Ger—NEAT Mal	.47** [.318, .574] (.49** [.296, .625])	.55** [.444, .635]	.41** [.120, .594]	.41** [.265, .516]	.60** [.360, .732]	.30** [.249, .357]
NEAT Bul—NEAT Ger	.68** [.669, .685] (.72** [.712, .728])	.78** [.762, .788]	.68** [.660, .695]	.58** [.552, .597]	.82** [.802, .837]	.48** [.459, .509]
NEAT Bul—NEAT Mal	.42** [.139, .598] (.46** [.177, .633])	.49** [.262, .637]	.41** [.080, .610]	.29** [.021, .493]	.59** [.368, .726]	.23** [.018, .396]
NEAT Mal—NEAT Ger	.43** [.168, .598] (.47** [.207, .633])	.50** [.304, .627]	.42** [.121, .609]	.36** [.131, .527]	.55** [.257, .715]	.23** [.036, .387]

Note. In the comparison of NEAT scales *Country* with verbal Likert scales in the first three rows, individual values for the variable NEAT scales *Country* are used. In the comparison NEAT scales *Country* with NEAT scales *Country* in the last three rows, individual values for either NEAT scales German or NEAT scales Malaysian are used. If Malaysia is included in the comparison, then this variable always contained the individual values. The other values are always mean values in both comparisons. In parentheses are the total ratings where the surprise ratings are excluded. The 95% confidence intervals are given in brackets. NEAT = Nonverbal Emotion Assessment Tool; Bul = Bulgarian; Ger = German; Mal = Malaysian.
$^{**}p < .01$.

for the different emotions also showed similar variability to the individual values. The surprise scale again had the lowest and joy the highest recognition rate (Table 7).

Rank Correlations for Intended Emotions at the Aggregate Level

Rank correlations determined for intended emotions also indicated high construct validity of the NEAT. Validity was highest for the NEAT European comparison, ρ = .92, *p* < .01, 95% CI [.911, .921], medium–high for the Bulgarian and Malaysian NEAT comparison, ρ = .84, *p* < .01, 95% CI [.825, .845], and lowest for the German and Malaysian NEAT comparison, ρ = .81, *p* < .01, 95% CI [.795, .817]. For the verbal Likert scale with NEAT association values ranged between ρ = .64, *p* < .01, 95% CI [.619, .657] for the association with Malaysia and ρ = .79, *p* < .01, 95% CI [.782, .806] for the association with Germany. The association for Bulgaria was .71, *p* < .01, 95% CI [.697, .728].

Exploratory Analysis: Boxplots for Target Versus Nontarget Emotions

Boxplots were created for a more detailed analysis of the finding that scores were consistently lowest for the Malaysian sample. To get a first impression, only the results for the fear vignettes are discussed below; boxplots were created separately for target and nontarget emotions and for each country (the plots can be found in the Figures A1–A6). For the boxplot "target emotion—German sample," the score of the German sample on each of the six fear vignettes was determined by including only the scale that captures the target emotion. The three boxplots for the target emotions indicated smaller differences between the samples than those for the nontarget emotions. Although the medians of the Malaysian sample appeared to be higher for the target emotions than those of the German and Bulgarian samples, this difference was not large. The median also fell within a narrower range in the Malaysian sample than in the Bulgarian sample, but the variance was similar to that in the German sample.

Much more meaningful were the findings revealed by the nontarget boxplots: The medians of the data fell into a much broader range or scattered more widely in the Malaysian sample than in the other two samples. Thus, disagreement in the assignment of scale scores for nontarget emotions appeared to be significantly greater in the Malaysian sample than in the other two samples. In addition, the medians in the Malaysian sample were significantly higher than those in the Bulgarian and German samples. Thus, the Malaysian sample assigned significantly higher scores to nontarget emotions across all vignettes than the other two samples. Since the created boxplots for the other five emotions also showed similar results (see Figures B1 through B24 in the additional online material, https://osf.io/3qpju/files/8hn4v?view_only=7b8ad19f4e9b45c48870d80b3fce71b8), it can be assumed that the findings of the analysis of the fear vignettes are generalizable: The NEAT's reduced validity when including the Malaysian sample can be attributed not so much to a lack of recognition of the target emotion but to a tendency to also perceive secondary, nontarget emotions. The finding that Malaysians perceive additional, nonintended emotions in vignettes echoes findings by Fang et al. (2019) for Chinese participants who perceive more such emotions in negative emotion expressions. This may point

Table 7
Intraclass Correlations With Means for Both Measures

Variable	Total	Anger	Sadness	Fear	Joy	Surprise
Verbal Ger—NEAT Ger	.88** [.809, .918] (.94** [.934, .951])	.91** [.766, .955]	.96** [.945, .975]	.92** [.896, .934]	.97** [.965, .978]	.63** [−.053, .851]
Verbal Ger—NEAT Bul	.83** [.760, .872] (.92** [.907, .922])	.90** [.772, .946]	.96** [.948, .968]	.81** [.740, .855]	.97** [.963, .970]	.52** [−.045, .770]
Verbal Ger—NEAT Mal	.63** [.102, .823] (.62** [.000, .833])	.74** [.356, .865]	.46** [−.094, .753]	.54** [−.042, .786]	.79** [−.038, .933]	.61** [.459, .710]
NEAT Bul—NEAT Ger	.97** [.973, .974] (.98** [.975, .976])	.99** [.985, .986]	.99** [.989, .990]	.95** [.924, .958]	.98** [.933, .987]	.96** [.946, .968]
NEAT Bul—NEAT Mal	.56** [−.089, .827] (.60** [−.082, .846])	.62** [−.005, .836]	.52** [−.090, .809]	.43** [−.075, .757]	.81** [−.041, .945]	.36** [−.064, .705]
NEAT Mal—NEAT Ger	.57** [−.087, .828] (.60** [−.077, .843])	.62** [.023, .826]	.54** [−.091, .814]	.50** [−.090, .797]	.73** [−.055, .910]	.38** [−.079, .714]

Note. Mean values for both variables are used in all six rows/comparisons. In parentheses are the total ratings where the surprise ratings are excluded. The 95% confidence intervals are given in brackets. NEAT = Nonverbal Emotion Assessment Tool; Bul = Bulgarian; Ger = German; Mal = Malaysian.
** $p < .01$.

to a more general tendency of East Asian participants toward the perception of mixed emotions.

Transparency and Openness

Within the methods sections of the included studies, we reported how we determined our sample size, all data exclusions (if any), all manipulations, and all measures used in the study, following Journal Article Reporting Standards (Kazak, 2018). All data, analysis code, and research materials for Studies 1, 2a, and 2 are available at https://osf.io/3qpju/?view_only=7b8ad19f4e9b45c48870d80b3fce71b8 (Pache et al., 2024). The data were analyzed using SPSS-Statistics Version 29.0.1.0. The design of the Studies and their analyses were not preregistered, because Study 1 was carried out in 2010 (when preregistrations were not yet common) and predictions were difficult given the objective of developing and validating an innovative research tool rather than testing hypotheses.

General Discussion

The present research explored a new, nonverbal tool based on pictographic emotion portrayals for assessing attributed emotions. In Study 1, children in primary school age matched 19 vignettes with and without additional emotional labels, as well as emotion labels alone with the corresponding NEAT emotion portrayals. In Study 2, participants from Germany, Bulgaria, and Malaysia rated 36 emotion vignettes on the NEAT scales. The results suggest that the NEAT is a valid tool for the assessment of emotions with children and in intercultural contexts. Despite the use of different vignettes across studies, the emotion identification rates were consistently relatively high, meaning that they are relatively independent of the vignettes. Across both studies and thus independently of the vignettes, the recognition rate was highest for joy and relatively low for fear and surprise.

These findings are comparable with data from other scales and procedures. Using verbal Likert scale ratings of facial expressions, Beaupré and Hess (2005) found across three cultural groups the highest accuracy for joy and the lowest for fear (surprise was not assessed). Similarly, better recognition rates for joy compared with negative facial expressions were found for photographs (Svärd et al., 2012) and schematic drawings (Leppänen & Hietanen, 2004, Study 2).

Lower recognition rates for fear in both Studies 1 and 2 replicate a frequent finding form facial expressions research that fear generally ranges among the least well-recognized emotions (e.g., Ekman & Friesen, 1971; Russell, 1994). However, whereas in studies involving facial expression on human faces fear was often confused with surprise (e.g., Ekman, 1994; Ekman & Friesen, 1976), in Study 1, fear was mainly confused with sadness, and vice versa. The misidentification with intended fear and decoded sadness was more pronounced than the opposite misidentification and hence may be due to the specifics of the vignettes used. As such, the findings show a general trend for some emotions to be better recognized than others and for some persistent confusions, which largely parallel findings from emotion expression research.

Although in Study 1 recognition performance differed when verbal labels were either included or omitted from the vignettes, this manipulation did not result in dramatic differences when only one emotion was portrayed. Each of the age groups included in the sample of Study 1 showed emotion recognition rates in at least a moderate range, so that generalizability across different age groups can also be assumed. Interestingly though, the findings revealed that the youngest children, unlike the older age groups, actually did not profit when labels were added to the vignettes. In fact, for some emotions the labels seemed to confuse them. This suggests that verbal labels were not fully acquired at that age. This is in line with a study by Kauschke et al. (2017), suggesting that the understanding of emotion terms increases significantly during primary school years. Whereas only about 55% of the 6-year-old pupils chose the terms given by the adult reference population in a task in which sentences were to be completed with emotion terms (productive sentence completion task), about 67% of the 9-year-old pupils did so. Similar results were obtained for the three other tasks relating to semantic processing. This also suggests that for this age group, the use of the NEAT is of definite advantage.

In Study 2, different statistical approaches confirmed large concordance in ratings across the three cultural groups as well when comparing verbal Likert scale ratings by Germans with each of the cultural groups. The vignette ratings were also ranked similarly on the intended scales by Germans, Bulgarians, and Malaysians. However, a number of cultural differences emerged in Study 2. In the German sample, agreement was highest between verbal rating and nonverbal rating; they were mid-level for the Bulgarian sample and lowest in the Malaysian sample. This phenomenon suggests greater cultural similarities between the two European samples and more differences with respect to the Southeast Asian sample. Similar differences were also found in the recognition of vocal expressions of emotions, in which emotion recognition rates were substantially higher in seven European countries and the United States than in Indonesia (Scherer et al., 2001). This is in line with the observation that in the case of Bulgarians and Germans, ICCs (including only mean scores) were .97 in total, and were never below .95 for any single emotion (see Table 7). Interrater reliabilities were notably lower (yet still high in an absolute sense) when comparing the responses of Bulgarians and Malaysians (.56) and the responses of the German and Malaysian participants (.57). Notably, the NEAT vignettes were created in the German cultural context and as such may lead to different attributions when used in a different context. This can be avoided by creating culturally appropriate vignettes for all studied cultural contexts and by asking participants from all cultures to rate all vignettes. This would allow to partial out the contribution of the cultural adequacy of the vignettes versus the NEAT.

Nonetheless, the overall high rates of concordance provide initial evidence that the NEAT can be used with participants from different countries in future studies. This is also supported by the fact that rank correlations of nonverbal comparison never fell below .81. Together, these findings suggest that the NEAT is a useful tool for use in research with children and in intercultural studies.

Limitations

Even though the present findings are convincing regarding the usefulness of the NEAT, there are some limitations. One issue regards the results for surprise vignettes. As noted above, for surprise, recognition rates were weakest in Study 2 and relatively low in Study 1. One explanation is that situations that evoke strong emotions are often surprising as well. While surprise may not be shown facially and hence does not confound research on facial emotion expressions, it is evident in the vignettes and as such may have led to confounds that reduced the validity of this scale. Another possibility is that participants may have interpreted the surprise scale as referring to emotional arousal in general. The observation that recognition rates for surprise were better in Study 1 suggests a potential influence of the emotion antecedent stories which differed between studies. Whereas the vignettes in Study 1 refer to third parties, the surprise vignettes in Study 2 mostly describe unexpected situations that affect the protagonists themselves or their immediate environment. Therefore, the more personal vignettes in Study 2 could also have led to a stronger identification with the described scenario and thus to a more differentiated evaluation of emotions, since past similar experiences and the emotions experienced in them are also taken into account.

A further caveat to be mentioned here is that our research does not show that emotion assessment with the NEAT is superior to the use of verbal Likert scales. Yet the results achieved using the NEAT scales by children and adults from three distinct cultures indicate that the NEAT scales can indeed be used to assess attributed emotions in different contexts.

A further limitation is the sample used for Study 2a: ideally, this study should have used a larger sample from each of the three cultures studied. The Bulgarian, German, and Malaysian participants in Study 2 varied with regard to age in that these samples also included some older participants, which, in the case of the Bulgarian and Malaysian samples, may also include participants who might be less familiar with the cultural background of their German counterparts. Despite these limitations and the ethnocentric bias, as it were, Study 2 has the advantage to provide a broader scope of information on the validity of the NEAT than using evidence from a single culture or country alone, as in Study 1.

Future Directions

The main innovation of the NEAT is that, without referring to emotion terms but based on drawn facial expressions instead, it assesses five qualitatively different emotions—anger, fear, joy, sadness, and surprise—both alone and in combination, as well as their intensity levels. This makes room for a wide application of the NEAT in different applied and research contexts. Further evaluation of the strengths and weaknesses of this tool is still needed. First, its applicability needs to be explored in developmental psychology with preschool children who are not yet able to read or not familiar with filling out questionnaires. In fact, given the findings from the youngest group in Study 1, the NEAT may prove an especially useful tool for younger ages. Second, given that Study 2 suggests that the NEAT might be relatively robust with respect to language barriers, its applicability in samples with language disorders remains to be explored. Third, cross-cultural comparisons including international and multiethnic target groups remain to be explored in more depth, including with rural populations who are not familiar with Western media. Fourth, the NEAT can also be a valuable component in multimethod research designs: Its nonverbal format allows for triangulation with verbal measures and offers the possibility to supplement or validate emotion ratings based on verbal labels. Conversely, the addition of verbal assessments can help refine or contextualize NEAT-based results. This mutual enrichment strengthens NEAT's potential as a flexible tool for various research environments. Finally, research into mixed emotions should also be further pursued with the NEAT; initial approaches to this have been outlined in Study 1 in this article. In order to further facilitate the use of the NEAT, in particular in its use with young children and in certain intercultural research contexts, the NEAT emotion portrayals may be printed on supports in the size of playing cards, with one facial expression per card.

In sum, the present results suggest that the NEAT has a high construct validity which is quite robust across different cultures, age groups, and vignette materials. This indicates that the NEAT and its applicability across distinct cultures, communities, and age groups are worthy of further investigation and that this multipurpose tool should be made available to the scientific community for further discussion and exploration.

References

Arifin, W. N. (2024). *Sample size calculator*. http://wnarifin.github.io

Baron-Cohen, S., Golan, O., & Ashwin, E. (2009). Can emotion recognition be taught to children with autism spectrum conditions? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3567–3574. https://doi.org/10.1098/rstb.2009.0191

Baron-Cohen, S., Golan, O., Wheelwright, S., Granader, Y., & Hill, J. (2010). Emotion word comprehension from 4 to 16 years old: A developmental survey. *Frontiers in Evolutionary Neuroscience*, *2*, Article 109. https://doi.org/10.3389/fnevo.2010.00109

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Beaupré, M. G., & Hess, U. (2005). Cross-cultural emotion recognition among Canadian ethnic groups. *Journal of Cross-Cultural Psychology*, *36*(1), 76–92. https://doi.org/10.1177/0022022104273656

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*(2), 179–190. https://doi.org/10.1037/0022-3514.92.2.179

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior*, *21*(1), 3–21. https://doi.org/10.1023/A:1024902500935

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., & Wigboldus, D. H. J. (2014). Stereotype associations and emotion recognition. *Personality and Social Psychology Bulletin*, *40*(5), 567–577. https://doi.org/10.1177/0146167213520458

Bijlstra, G., Holland, R. W., Dotsch, R., & Wigboldus, D. H. J. (2019). Stereotypes and prejudice affect the recognition of emotional body postures. *Emotion*, *19*(2), 189–199. https://doi.org/10.1037/emo0000438

Binetti, N., Roubtsova, N., Carlisi, C., Cosker, D., Viding, E., & Mareschal, I. (2022). Genetic algorithms reveal profound individual differences in emotion recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(45), Article e2201380119. https://doi.org/10.1073/pnas.2201380119

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brechet, C. (2017). Children's recognition of emotional facial expressions through photographs and drawings. *The Journal of Genetic Psychology*, *178*(2), 139–146. https://doi.org/10.1080/00221325.2017.1286630

Broekens, J., & Brinkman, W.-P. (2013). AffectButton: A method for reliable and valid affective self-report. *International Journal of Human–Computer Studies*, *71*(6), 641–667. https://doi.org/10.1016/j.ijhcs.2013.02.003

Bruner, J. S., & Tagiuri, R. (1954). The perception of people. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 2, pp. 634–655). Addison-Wesley.

Camras, L. A., & Allison, K. (1985). Children's understanding of emotional facial expressions and verbal labels. *Journal of Nonverbal Behavior*, *9*(2), 84–94. https://doi.org/10.1007/BF00987140

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Covic, A., von Steinbüchel, N., & Kiese-Himmel, C. (2020). Emotion recognition in kindergarten children. *Folia Phoniatrica et Logopaedica*, *72*(4), 273–281. https://doi.org/10.1159/000500589

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., & Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, *589*(7841), 251–257. https://doi.org/10.1038/s41586-020-3037-7

Cüceloglu, D. (1970). Perception of facial expressions in three different cultures. *Ergonomics*, *13*(1), 93–100. https://doi.org/10.1080/00140137008931123

Darwin, C. (1998). *The expression of the emotions in man and animals*. Oxford University Press. (Original work published 1872). https://doi.org/10.1093/oso/9780195112719.001.0001

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation, 1971* (Vol. 19, pp. 207–283). University of Nebraska Press.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, *115*(2), 268–287. https://doi.org/10.1037/0033-2909.115.2.268

Ekman, P., & Ekman, E. (2018). *Atlas of emotion*. Retrieved August 2024, from https://atlasofemotions.org/#introduction/

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129. https://doi.org/10.1037/h0030377

Ekman, P., & Friesen, W. V. (1976). *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system: The manual*. Research Nexus.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). Universals and cultural differences in the judgements of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717. https://doi.org/10.1037/0022-3514.53.4.712

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203

Elfenbein, H. A., Beaupré, M. G., Levesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion*, *7*(1), 131–146. https://doi.org/10.1037/1528-3542.7.1.131

Fang, X., Sauter, D. A., & Van Kleef, G. A. (2019). Seeing mixed emotions: The specificity of emotion perception from static and dynamic facial expressions across cultures. *Emotion*, *19*(5), 856–870. https://doi.org/10.1037/emo0000459

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. Academic Press.

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science*, *27*(4), 211–219. https://doi.org/10.1177/0963721417746794

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, *5*(1), 60–65. https://doi.org/10.1177/1754073912451331

Hess, U. (2017). Emotion categorization. In C. Lefebvre & H. Cohen (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 107–126). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00005-1

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 B: Biological Sciences*, *364*(1535), 3497–3504. https://doi.org/10.1098/rstb.2009.0165

Hess, U., & Kafetsios, K. (2022). Infusing context into emotion perception impacts emotion decoding accuracy. *Experimental Psychology*, *68*(6), 285–294. https://doi.org/10.1027/1618-3169/a000531

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology*, *55*(9), 1830–1849. https://doi.org/10.1037/dev0000686

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*(6), 640–643. https://doi.org/10.1046/j.0956-7976.2003.psci_1478.x

Hugenberg, K., & Sacco, D. F. (2008). Social categorization and stereotyping: How social categorization biases person perception and face memory. *Social and Personality Psychology Compass*, *2*(2), 1051–1072. https://doi.org/10.1111/j.1751-9004.2008.00090.x

Hupka, R. B., Lenton, A. P., & Hutchison, K. A. (1999). Universal development of emotion categories in natural language. *Journal of Personality and Social Psychology*, *77*(2), 247–278. https://doi.org/10.1037/0022-3514.77.2.247

Izard, C. E. (1971). *The face of emotion*. Appleton-Century-Crofts.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, *115*(2), 288–299. https://doi.org/10.1037/0033-2909.115.2.288

Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of Emotion. *Journal of Experimental Psychology: General*, *145*(6), 708–730. https://doi.org/10.1037/xge0000162

Kafetsios, K., & Hess, U. (2023). Reconceptualizing emotion recognition ability. *Journal of Intelligence*, *11*(6), Article 123. https://doi.org/10.3390/jintelligence11060123

Kauschke, C., Bahn, D., Vesker, M., & Schwarzer, G. (2017). Semantische Repräsentation von Emotionsbegriffen bei Kindern im Grundschulalter [The semantic representation of emotion terms in primary-school–age children]. *Kindheit und Entwicklung*, *26*(4), 251–260. https://doi.org/10.1026/0942-5403/a000238

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kilbride, J. E., & Yarczower, M. (1976). Recognition of happy and sad facial expressions among Baganda and U.S. children. *Journal of Cross-Cultural Psychology*, *7*(2), 181–194. https://doi.org/10.1177/002202217672006

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*(2), 363–374. https://doi.org/10.2307/2529786

Leiner, D. J. (2019). *SoSci Survey* (Version 3.1.06) [Computer software]. https://www.soscisurvey.de

Leppänen, J. M., & Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research*, *69*(1), 22–29. https://doi.org/10.1007/s00426-003-0157-2

Lindquist, K. A., Jackson, J. C., Leshin, J., Satpute, A. B., & Gendron, M. (2022). The cultural evolution of emotion. *Nature Reviews Psychology*, *1*(11), 669–681. https://doi.org/10.1038/s44159-022-00105-4

Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The hundred-year emotion war: Are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychological Bulletin*, *139*(1), 255–263. https://doi.org/10.1037/a0029038

Matsumoto, D. (2005). Scalar ratings of contempt expressions. *Journal of Nonverbal Behavior*, *29*, 91–104. https://doi.org/10.1007/s10919-005-2742-0

Michalson, L., & Lewis, M. (1985). What do children know about emotions and when do they know it? In M. Lewis & C. Saarni (Eds.), *The socialization of emotions* (pp. 117–139). Springer. https://doi.org/10.1007/978-1-4613-2421-8_6

Naumann, S., Bayer, M., Kirst, S., van der Meer, E., & Dziobek, I. (2023). A randomized controlled trial on the digital socio-emotional competence training Zirkus Empathico for preschoolers. *npj Science of Learning*, *8*(1), Article 20. https://doi.org/10.1038/s41539-023-00169-8

Nunnally, J. C. (1978). An overview of psychological measurement. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders: A handbook* (pp. 97–146). Academic Press. https://doi.org/10.1007/978-1-4684-2490-4_4

Ogarkova, A. (2013). Folk emotion concepts: Lexicalization of emotional experiences across languages and cultures. In J. J. R. Fontaine, K. R. Scherer, & C. Soriano (Eds.), *Components of emotional meanings: A sourcebook* (pp. 46–62). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199592746.003.0004

Ogarkova, A. (2016). Translatability of emotions. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 575–599). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100508-8.00023-0

Ogarkova, A. (2021). Cross-lingual translatability of emotion terms: A review. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 909–935). Elsevier. https://doi.org/10.1016/B978-0-08-100508-8.00023-0

Pache, M., Miketta, L., Banse, R., & Hess, U. (2024, November 12). *NEAT_Nonverbal_Emotion_Assessment_Tool*. Open Science Framework. https://osf.io/3qpju/?view_only=7b8ad19f4e9b45c48870d80b3fce71b8

Ridgeway, D., Waters, E., & Kuczaj, S. A. (1985). Acquisition of emotion descriptive language: Receptive and productive vocabulary norms for ages 18 months to 6 years. *Developmental Psychology*, *21*(5), 901–908. https://doi.org/10.1037/0012-1649.21.5.901

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin*, *110*(3), 426–450. https://doi.org/10.1037/0033-2909.110.3.426

Russell, J. A. (1994). Is there universal recognition from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102–141. https://doi.org/10.1037/0033-2909.115.1.102

Russell, J. A. (1995). Facial expressions of emotion: What lies beyond minimal universality? *Psychological Bulletin*, *118*(3), 379–391. https://doi.org/10.1037/0033-2909.118.3.379

Russell, J. A., & Widen, S. C. (2002). A label superiority effect in children's categorization of facial expressions. *Social Development*, *11*(1), 30–52. https://doi.org/10.1111/1467-9507.00185

Saarni, C. (1999). *The development of emotional competence*. Guilford Press.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychological Science*, *26*(3), 354–356. https://doi.org/10.1177/0956797614560771

Scarantino, A. (2019). Affective pragmatics extended: From natural to overt expressions of emotions. In U. Hess & S. Hareli (Eds.), *The social nature of emotion expression* (pp. 49–81). Springer. https://doi.org/10.1007/978-3-030-32968-6_4

Scarantino, A., Hareli, S., & Hess, U. (2022). Emotional expressions as appeals to recipients. *Emotion*, *22*(8), 1856–1868. https://doi.org/10.1037/emo0001023

Scherer, K. R., Banse, R., & Wallbott, H. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92. https://doi.org/10.1177/0022022101032001009

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Svärd, J., Wiens, S., & Fischer, H. (2012). Superior recognition performance for happy masked and unmasked faces in both younger and older adults. *Frontiers in Psychology*, *3*, Article 520. https://doi.org/10.3389/fpsyg.2012.00520

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, *17*(1), 101–110. https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press. https://doi.org/10.1017/CBO9780511521256

Wingenbach, T. S., Morello, L. Y., Hack, A. L., & Boggio, P. S. (2019). Development and validation of verbal emotion vignettes in Portuguese, English, and German. *Frontiers in Psychology*, *10*, Article 1135. https://doi.org/10.3389/fpsyg.2019.01135

Yang, Y., & Wang, Q. (2019). Culture in emotional development. In N. A. Fox, R. J. Davidson, & C. A. Kalin (Eds.), *Handbook of emotional development* (pp. 569–593). Springer. https://doi.org/10.1007/978-3-030-17332-6_22

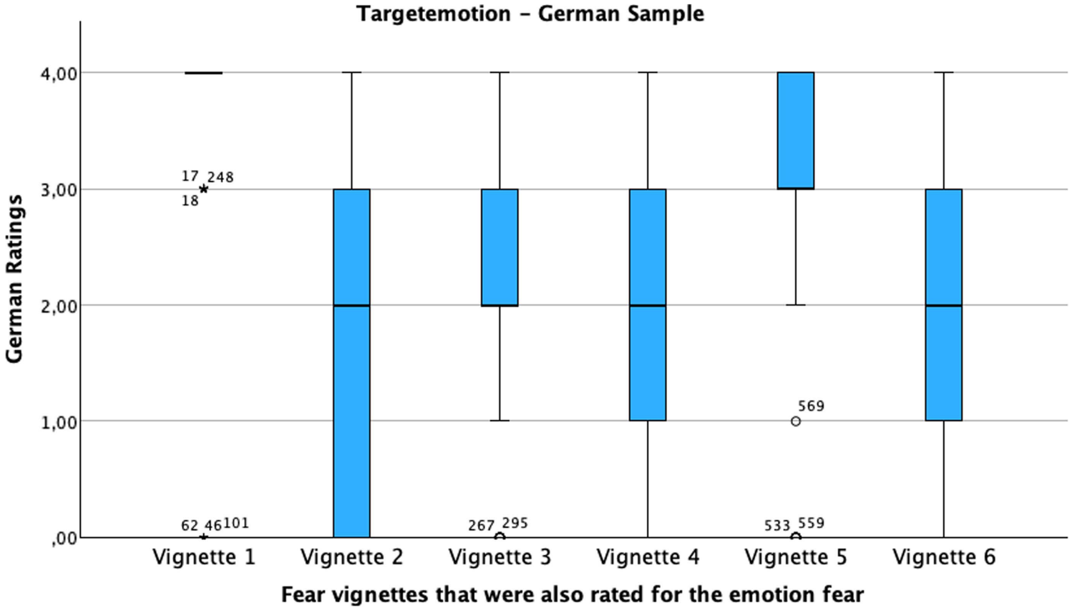
(*Appendix follows*)

Appendix

Boxplots for Target Versus Nontarget Emotions

Figure A1

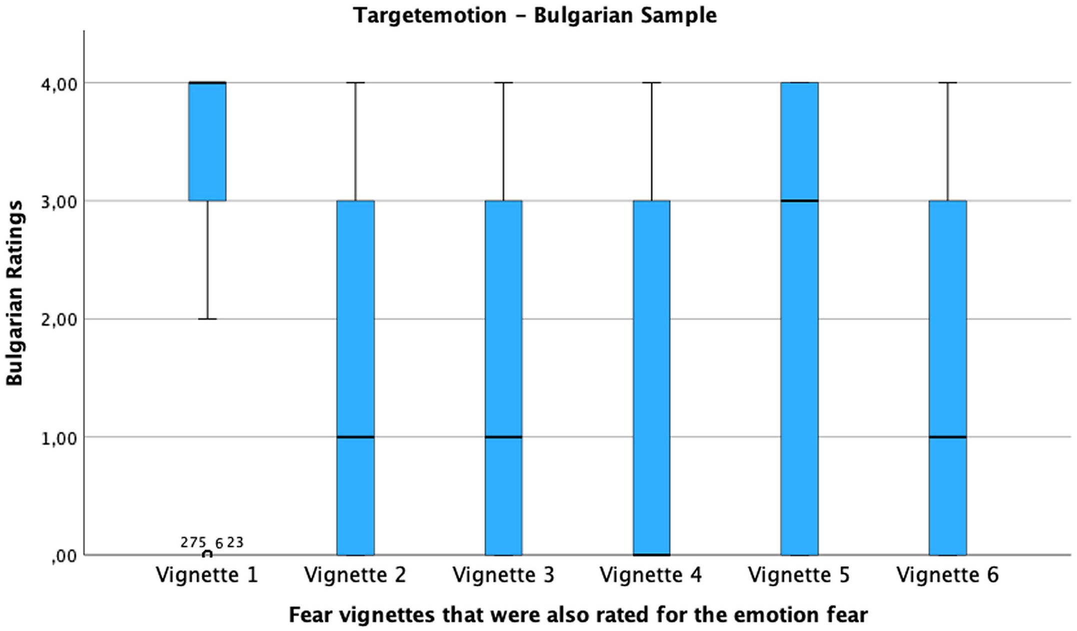
Boxplot Target Emotion Fear: German



Note. The small circle indicates mild outliers and the asterisk denotes extreme outliers. See the online article for the color version of this figure.

Figure A2

Boxplot Target Emotion Fear: Bulgarian

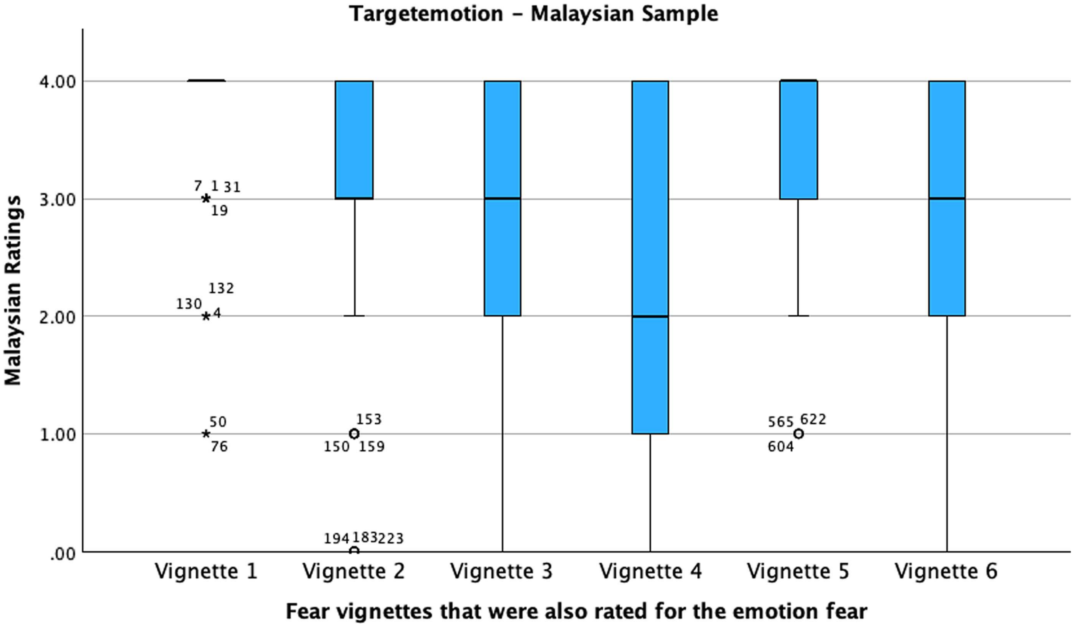


Note. The small circle indicates mild outliers and the asterisk denotes extreme outliers. See the online article for the color version of this figure.

(*Appendix continues*)

Figure A3

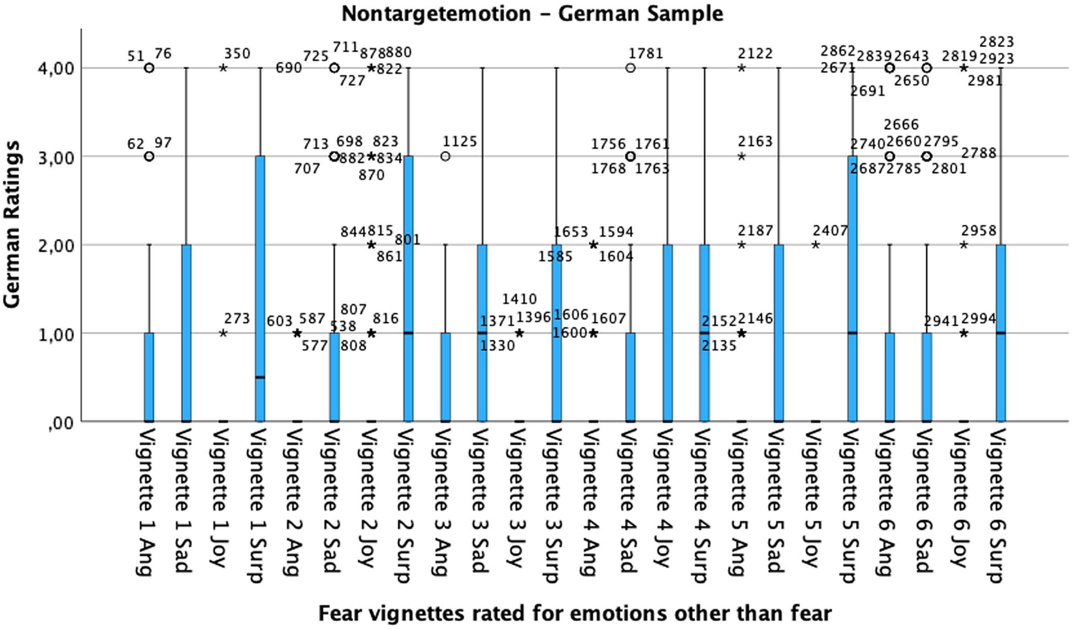
Boxplot Target Emotion Fear: Malaysian



Note. The small circle indicates mild outliers and the asterisk denotes extreme outliers. See the online article for the color version of this figure.

Figure A4

Boxplot Nontarget Emotion Fear: German

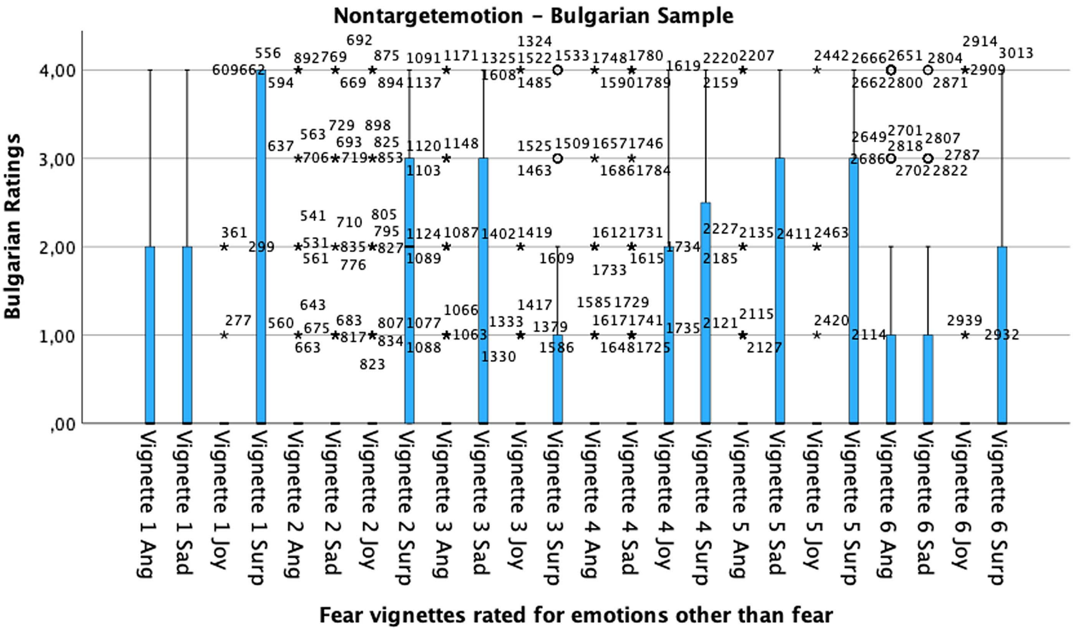


Note. The small circle indicates mild outliers and the asterisk denotes extreme outliers. Ang = anger; Sad = sadness; Surp = surprise. See the online article for the color version of this figure.

(*Appendix continues*)

Figure A5

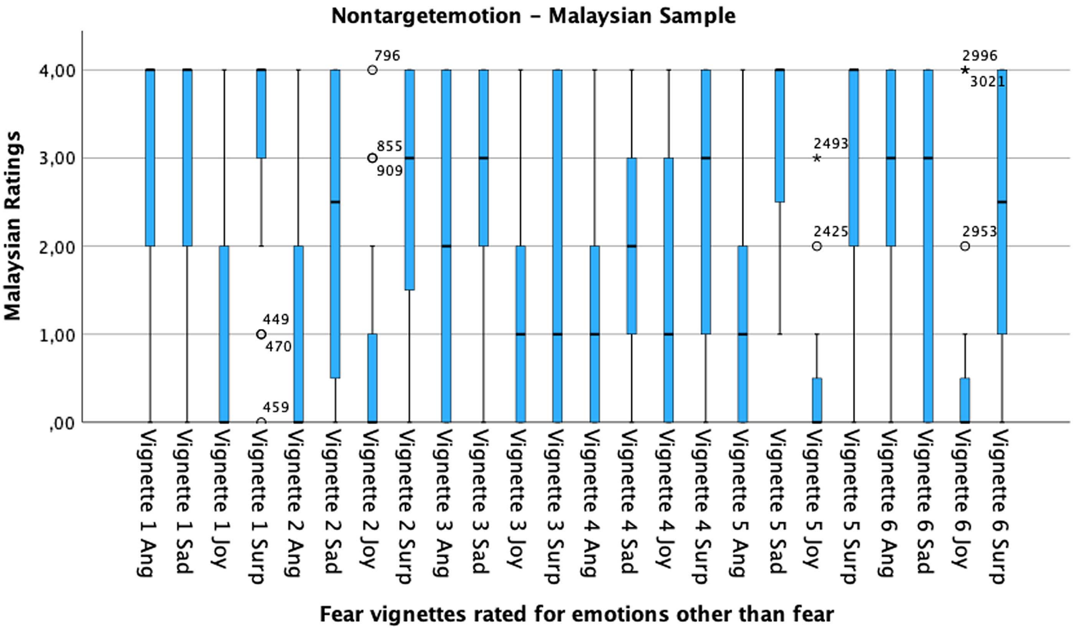
Boxplot Nontarget Emotion Fear: Bulgarian



Note. The small circle indicates mild outliers and the asterisk denotes extreme outliers. Ang = anger; Sad = sadness; Surp = surprise. See the online article for the color version of this figure.

Figure A6

Boxplot Nontarget Emotion Fear: Malaysian



Note. The small circle indicates mild outliers and the asterisk denotes extreme outliers. Ang = anger; Sad = sadness; Surp = surprise. See the online article for the color version of this figure.

Received February 1, 2024
Revision received August 28, 2025
Accepted December 10, 2025 ■