

# Cultural Diversity and Ethnic Minority Psychology

## Cultural Background and Input Familiarity Influence Multisensory Emotion Perception

Peiyao Chen, Ashley Chung-Fat-Yim, Taomei Guo, and Viorica Marian

Online First Publication, January 23, 2023. <https://dx.doi.org/10.1037/cdp0000577>

### CITATION

Chen, P., Chung-Fat-Yim, A., Guo, T., & Marian, V. (2023, January 23). Cultural Background and Input Familiarity Influence Multisensory Emotion Perception. *Cultural Diversity and Ethnic Minority Psychology*. Advance online publication. <https://dx.doi.org/10.1037/cdp0000577>

# Cultural Background and Input Familiarity Influence Multisensory Emotion Perception

Peiyao Chen<sup>1</sup>, Ashley Chung-Fat-Yim<sup>2</sup>, Taomei Guo<sup>3, 4</sup>, and Viorica Marian<sup>2</sup>

<sup>1</sup> Department of Psychology, Swarthmore College

<sup>2</sup> Department of Communication Sciences and Disorders, Northwestern University

<sup>3</sup> State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University

<sup>4</sup> Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University



**Objectives:** During multisensory emotion perception, the attention devoted to the visual versus the auditory modality (i.e., modality dominance) varies depending on the cultural background of the perceiver. In the present study, we examined (a) how cultural familiarity influences multisensory emotion perception in Eastern and Western cultures and (b) the underlying processes accounting for the cultural difference in modality dominance. **Method:** Native Mandarin speakers from China and native English speakers from the United States were presented with audiovisual emotional stimuli from their own culture (i.e., familiar) and from a different culture (i.e., unfamiliar) and asked to evaluate the emotion from one of the two modalities. Across modalities, the emotions were either the same (i.e., congruent, happy face, and happy voice) or different (i.e., incongruent, happy face, and sad voice). **Results:** When the input was in a familiar cultural context, American participants were more influenced by the visual modality, while Chinese participants were more influenced by the auditory modality. While both groups integrated the incongruent emotion from the irrelevant modality, only the American group integrated the congruent emotion from the irrelevant modality. When the input was in a less familiar cultural context, both groups showed increased visual dominance, but only the Chinese group simultaneously showed decreased auditory dominance. **Conclusions:** We conclude that cultural background and input familiarity interact to influence modality dominance during multisensory emotion perception.


## Public Significance Statement

The present study reveals that American participants were more influenced by facial expressions than vocal expressions, while Chinese participants were more influenced by vocal expressions than facial expressions during multisensory emotion perception. Recognizing these differences could facilitate communication and interactions between individuals from East Asian and Western cultures.


**Keywords:** multisensory perception, emotions, cross-cultural differences, modality dominance, familiarity

**Supplemental materials:** <https://doi.org/10.1037/cdp0000577.supp>

Peiyao Chen  <https://orcid.org/0000-0003-2409-8703>

Ashley Chung-Fat-Yim  <https://orcid.org/0000-0002-6905-8302>


Taomei Guo  <https://orcid.org/0000-0002-4682-7818>

Viorica Marian  <https://orcid.org/0000-0002-8335-1433>

Research reported in this publication was supported in part by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number R01HD059858 to Viorica Marian. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Marc Pell for providing the vocal emotion stimuli. The authors also thank Pan Liu and the members of the *Bilingualism and Psycholinguistics Research Group* for their helpful comments and input.

Peiyao Chen played lead role in data curation, formal analysis,

investigation, methodology, software, validation and writing of original draft, supporting role in visualization and equal role in conceptualization, project administration and writing of review and editing. Ashley Chung-Fat-Yim played lead role in visualization, supporting role in writing of original draft and equal role in formal analysis and writing of review and editing. Taomei Guo played supporting role in resources and writing of review and editing. Viorica Marian played lead role in funding acquisition, resources and supervision, supporting role in data curation, formal analysis, investigation, methodology, software, validation, visualization and writing of original draft and equal role in conceptualization, project administration and writing of review and editing.

 The data are available at <https://osf.io/nbhpe/>.

Correspondence concerning this article should be addressed to Peiyao Chen, Department of Psychology, Swarthmore College, 500 College Avenue, Swarthmore, PA 19081, United States. Email: [pchen3@swarthmore.edu](mailto:pchen3@swarthmore.edu)

Emotions can be expressed nonverbally through several modalities, including the visual modality (i.e., facial expressions) and auditory modality (i.e., vocal expressions). Despite early evidence suggesting the universality of emotion recognition (Ekman & Friesen, 1971; Izard, 1969), recent studies have found that emotions are influenced by norms and values that vary across cultures (see Mesquita et al., 2016, for a review). People perceive facial expressions (e.g., Ekman et al., 1987; Fang et al., 2019; Jack et al., 2009; Masuda et al., 2008, Yuki et al., 2007), vocal expressions (e.g., Ishii et al., 2003; Kitayama & Ishii, 2002), and multisensory emotions (Liu et al., 2015a, 2015b; Tanaka et al., 2010) differently, depending on their cultural background. As our world becomes more diverse, we are routinely interacting with people from different cultures, leading to increased exposure to foreign languages and faces. Yet, it remains relatively unknown to what extent our cultural background influences how we evaluate the emotions of those from a different culture. The present study examines how cultural background and familiarity with the auditory and visual inputs interact to affect multisensory emotional processing.

### Cultural Differences in Multisensory Emotion Perception

Previous cross-cultural studies have found that individuals from Western cultures are more influenced by the information in the visual modality (i.e., visual dominance), while individuals from Eastern cultures are more influenced by the information in the auditory modality (i.e., auditory dominance). This cultural difference in visual and auditory dominance between Easterners and Westerners was first demonstrated by Tanaka et al. (2010). Japanese and Dutch participants were presented with audio-video recordings of actors displaying happy or sad facial expressions and speaking in either happy or sad voices. Participants were asked to judge the emotion in one modality as happy or sad and ignore the emotion in the other modality. Across modalities, the emotions could either be congruent (i.e., a happy face with a happy voice) or incongruent (i.e., a happy face with a sad voice). The influence of the auditory and visual modality was calculated by computing the difference in accuracy between the congruent and incongruent trials. Japanese participants were more influenced by the voice than Dutch participants when judging the facial expression. In contrast, Dutch participants were more influenced by the face than Japanese participants when judging the vocal expression. Consistent with these results, Mandarin-speaking Chinese participants experienced greater influence from irrelevant vocal cues than English-speaking North American participants (Liu et al., 2015a), whereas English-speaking North American participants experienced greater influence from irrelevant facial cues than Mandarin-speaking Chinese participants (Liu et al., 2015b). These findings demonstrate that multisensory emotion integration and perception are modulated by a person's cultural background.

Tanaka et al. (2010) and Liu et al. (2015a, 2015b) have attributed the cross-cultural differences in modality dominance to the display rules that each culture prescribes to. Display rules are a set of cultural norms learned from an early age that regulate how and when we should express our emotions in particular social situations (Ekman & Friesen, 1969). Cultural norms consequently influence how emotions are perceived and integrated (Marian, 2023). For example, individuals from Western individualistic cultures are encouraged to show their emotions through direct and explicit

means to influence others, such as making eye contact with others (Kitayama & Ishii, 2002). In contrast, individuals from Eastern collectivistic cultures are discouraged from expressing their feelings to maintain group harmony (Ekman, 1972; Markus & Kitayama, 1991; Matsumoto et al., 1998, 2008). During face-to-face interactions, individuals from Eastern cultures maintain less eye contact with others compared to individuals from Western cultures (e.g., Argyle & Cook, 1976; Hawrysh & Zaichkowsky, 1990; McCarthy et al., 2006, 2008). Similarly, when perceiving emotions, individuals from Eastern cultures may direct more of their attention to the auditory modality (Sanchez-Burks et al., 2003), demonstrating a different modality dominance compared to individuals from Western cultures.

As mentioned earlier, modality dominance is typically measured by calculating the accuracy difference between the congruent condition and the incongruent condition (Liu, Rigoulot & Pell 2015b; Tanaka et al., 2010). Because this calculation includes only congruent and incongruent conditions, it is unclear whether the cultural differences in modality dominance are due to facilitation from the congruent emotion in the unattended modality, interference from the incongruent emotion in the unattended modality, or a combination of the two. One possibility is that modality dominance could be due to *both* facilitation and interference effects. For example, the co-occurring voice may enhance the recognition of emotional facial expressions when the emotion is congruent (facilitation) but also lead to impairment when the emotion is incongruent (interference). Another possibility is that modality dominance could be due to either a larger facilitation effect *or* a larger interference effect from the unattended modality. For example, the co-occurring voice only interferes with, but does not facilitate, the recognition of emotional facial expressions. Dissociating interference and facilitation effects makes it possible to identify the source of cultural differences in modality dominance.

Takagi et al. (2015) examined facilitation, interference, and modality dominance effects in Japanese speakers. The authors found an interference effect where Japanese participants automatically integrated information from both modalities even when instructed to focus on only one modality. The facilitation effects were only evident when judging the voice, suggesting that the recognition of vocal emotions was enhanced by the congruent facial cues. Because the study by Takagi et al. consisted of only Japanese participants, it remains an open question whether Westerners would exhibit the same pattern of facilitation and interference effects in each task as Easterners do. The present study investigates the source of the modality dominance effect (i.e., facilitation and/or interference) in both Easterners and Westerners.

### Input Quality in Multisensory Emotion Perception

In addition to cultural background, modality dominance changes depending on the quality of the input. When the visual and auditory modalities are both optimal, individuals tend to prefer the visual modality (Collignon et al., 2008). However, as the quality of the visual input decreases, the reliance on the auditory modality increases. Along similar lines, ambiguity in one modality leads to increased reliance on the other modality (De Gelder & Vroomen, 2000; Massaro & Egan, 1996). Facial expressions that fall in the middle of the happy-sad continuum (i.e., more ambiguous and neutral) receive the greatest influence from the voice (De Gelder &

Vroomen, 2000). These results indicate that individuals can adjust their reliance on visual and auditory modalities depending on the characteristics of the input. But what happens when *both* the auditory and visual modalities are less familiar to the perceiver? Are multisensory emotions perceived and evaluated differently?

Quality and ambiguity of the input, as well as familiarity with the input, in one modality influence how much information one can extract from that modality. Moreover, neural evidence suggests that affective information, when it is ambiguous, degraded, or unfamiliar, would activate brain regions associated with cognitive control to signal that further exploration is needed to make a reliable judgment about the stimuli and environment (Watson et al., 2013; see Schreuder et al., 2016, for a review).

According to the modality precision theory (e.g., Choe et al., 1975; Fisher, 1968, also see Freides, 1974), modality appropriateness theory (O'Connor & Hermelin, 1972), and the estimated precision theory (Ernst & Bühlhoff, 2004), individuals favor the modality that provides the most precise information about the event. Because recognizing facial expressions is easier and more accurate than vocal expressions (Elfenbein & Ambady, 2002), greater sensitivity toward the visual modality is expected when processing unfamiliar audiovisual emotional information. To date, only one study has examined modality dominance in an unfamiliar context (Tanaka et al., 2010) and found that Japanese participants, but not Dutch participants, were influenced by auditory information when the language was unfamiliar. Furthermore, the findings from Liu et al. (2015a, 2015b) suggest that native Chinese speakers may be less distracted by auditory information, as there were no behavioral differences between the face and voice tasks. Therefore, we aimed to examine whether Chinese participants would show similar auditory dominance when perceiving stimuli from their own culture and from a less familiar culture.

## The Present Study

The present study examined two research questions: (a) How does familiarity with the visual and auditory inputs influence modality dominance in Eastern and Western cultures? and (b) Are the cultural differences in modality dominance in familiar and unfamiliar contexts due to facilitation, interference, or both? To address these questions, we used an emotion recognition task to compare multisensory emotion integration between native Mandarin speakers from China and native English speakers from the United States. Participants either saw a face, heard a meaningless pseudosentence, or were presented with both simultaneously. The task was to judge either the emotion of the face (i.e., face task) or the emotion of the voice (i.e., voice task) in their own culture and in a less familiar culture.

For the first research question, we predicted that, under the familiar context, American participants would be more influenced by the visual modality, whereas Chinese participants would be more influenced by the auditory modality or equally impacted by the two modalities, based on the cross-cultural literature in multisensory emotion perception (Liu et al., 2015a, 2015b; Tanaka et al., 2010). Specifically, we expected a larger modality dominance (i.e., a larger difference between congruent and incongruent trials) in the voice task than in the face task for American participants, but a larger modality dominance in the face task than in the voice task for Chinese participants. In the unfamiliar context, we predicted that

American and Chinese participants would show increased visual dominance, in line with the estimated precision hypothesis (Ernst & Bühlhoff, 2004), which means that the modality dominance in the voice task would increase for both groups of participants. However, another possibility is that Americans would show a different pattern of modality dominance between familiar and unfamiliar contexts, similar to the Dutch participants in Tanaka et al. (2010) study, while Chinese participants would maintain the same pattern of modality dominance across both contexts.

For the second research question, we examined the interference and facilitation effects in Chinese and American participants in both familiar and unfamiliar contexts. We predicted that American participants would have greater difficulty ignoring an incongruent facial emotion than an incongruent vocal emotion (i.e., larger interference effect) but also benefit more from a consistent facial emotion than a consistent vocal emotion (i.e., larger facilitation effect). For Chinese participants, we predicted that they would show greater difficulty ignoring incongruent voices, but also benefit more from a consistent vocal emotion than a consistent facial emotion (i.e., larger interference and facilitation effects).

## Method

### Participants

Thirty-four Mandarin-speaking Chinese participants living in China and 32 English-speaking American participants living in the United States took part in this study.<sup>1</sup> Chinese participants were recruited through an online platform at a local university in Beijing, while American participants were recruited through email listservs or flyers posted around a Midwestern university. The study was approved by the local institutional review board.

Participants needed to meet specific inclusionary and exclusionary criteria. Chinese participants must have been born and raised in China, not be majoring in any Western languages or cultural studies, and have not previously lived in a Western country for more than 3 months. American participants must have been born and raised in the United States or Canada, not be majoring in any Asian languages or Asian studies, and have not previously lived in an East Asian country for more than 3 months. The study took place online, and informed consent was obtained in each participant's native language. At the end of the experiment, both groups were compensated with an electronic gift card for their time.

Six participants were excluded for not completing the task (2 Chinese, 1 American), not following instructions (1 Chinese), and performing below chance (1 Chinese, 1 American). In addition, one American participant produced mean response times that were 3 SDs above their group's mean and was thus removed from further analyses. The final sample consisted of 30 Chinese (10 males, 20 females) and 29 American participants (9 males, 20 females). Participants' linguistic and cultural background information (see Table 1) were obtained using the *Language Experience and Proficiency Questionnaire* (Marian et al., 2007). The American and

<sup>1</sup> Using G\*Power 3.1 (Faul et al., 2009), an a priori power analysis was performed. Based on the effect size of  $r = .35$  (Liu et al., 2015b),  $\alpha = .05$ , and power = .85, the minimum number of participants needed to obtain a similar effect was approximately  $N = 14$  in each group for the between-group comparison. To account for a potentially larger variance among remote participants, we more than doubled the projected sample size.

**Table 1**  
*Background Information by Cultural Group*

Measures	Chinese <i>M (SD)</i>	American <i>M (SD)</i>	<i>p</i> value
<i>N</i>	30	29	
Age in years	22.77 (3.13)	22.83 (4.23)	.95
Years of education	16.27 (2.21)	15.83 (2.74)	.50
Native language proficiency rating (/10)	9.03 (0.96)	9.79 (0.49)	<.001
Other language proficiency rating (/10)	5.57 (1.36)	3.38 (3.00)	<.001
Age of other language acquisition <sup>a</sup>	8.50 (3.31)	10.9 (4.48)	.034
Daily exposure to Western culture (%)	17.20 (15.07)	86.62 (15.73)	<.001
Daily exposure to Eastern culture (%)	78.00 (17.44)	5.96 (11.57)	<.001

*Note.* Proficiency was self-reported on a scale from 1 (*very low*) to 10 (*perfect*). Daily exposure was reported in terms of the percentage per day.

<sup>a</sup>Twenty American participants responded to this item.

Chinese groups were matched on age and years of education ( $t_s < 1$ ), with education level used as an index of socioeconomic status (Hollingshead, 1975). The race of all Chinese participants was Asian, and the race of all American participants was White. All participants had a normal or corrected-to-normal vision, no hearing impairments, and no previous history of neuropsychological disorders.

## Emotion Recognition Task

### Vocal Stimuli

Twenty Mandarin and 20 English pseudosentences (grammatically correct sentences with no semantic information; Mandarin: “他在地上拔冲.” English: “They nestered the flugs.”) were obtained from two validated vocal emotion databases (Liu & Pell, 2012; Pell et al., 2009). Pseudosentences were chosen because semantic content has been shown to influence emotional tone judgments (Ishii et al., 2003).<sup>2</sup> Pseudosentences in each language were delivered by four different native speakers of that language (2 females and 2 males) in five different emotions (happiness, sadness, disgust, fear, and anger). According to the normed data within each database (Liu & Pell, 2012; Pell et al., 2009), Chinese and English pseudosentences were matched on recognition rate (Mandarin = 86%, English = 88%), emotional intensity (Mandarin = 3.3 out of 5, English = 3.4 out of 5), and duration (Mandarin = 1.78 s, English = 1.79 s),  $t_s < 1$ .

### Face Stimuli

The Asian faces were obtained from the Taiwanese Facial Expression Image Database (Chen & Yen, 2007), and the Caucasian faces were obtained from the Karolinska Directed Emotional Faces database (Lundqvist et al., 1998). Four actors (2 females and 2 males) portraying all five emotions were selected from each database. According to the normed data within each database (Chen & Yen, 2007; Lundqvist et al., 1998), Asian and Caucasian faces did not differ in recognition rate (Asian = 83%, Caucasian = 84%) and emotional intensity ratings (Asian = 5.6 out of 9, Caucasian = 5.7 out of 9),  $t_s < 1$ . All faces were reprocessed to the same dimension (345 pixels wide  $\times$  430 pixels high) and resolution (300 dpi) and converted into grayscale using GNU Image Manipulation Program 2 (GNU Image Manipulation Program Development Team, 2018) to control for brightness and contrast.

## Bimodal Stimuli

For each culture, the voice and face stimuli were paired together to construct bimodal stimuli (Figure 1). Each unique voice was always paired with the same unique face of the same gender to maintain consistency between the face and voice identity. For each voice–face pairing, one facial expression (e.g., happy face) was paired once with a voice of the same emotion (e.g., happy voice) to construct a bimodal congruent trial and once with each of the remaining four emotions (e.g., sad, disgusted, fearful, and angry) to construct four bimodal incongruent trials, resulting in 20 bimodal congruent trials (face and voice exhibit the same emotion) and 80 bimodal incongruent trials (face and voice exhibit different emotions) in each culture.

Within a culture, four bimodal lists were created, each containing 20 congruent trials and 20 of the 80 incongruent trials. Thus, the same face (or voice) appeared once in the congruent condition and once in the incongruent condition. The emotions were equally distributed across congruent and incongruent conditions. In addition, two unimodal lists were created containing either 20 faces (unimodal face list) or 20 voices (unimodal voice list). Participants received one of the four bimodal lists and both unimodal lists from both Eastern and Western cultures. The same bimodal list was used for the face and voice tasks. Stimuli from the participant’s own culture were considered familiar, whereas stimuli from the other culture were considered unfamiliar.

<sup>2</sup> A potential concern related to using pseudosentences is that the vocal expressions may be easier to ignore due to the lack of semantic information, compared to the facial expressions. We considered using neutral sentences spoken in different emotional tones, however, as noted by Liu and Pell (2012) and Pell (2006), neutral sentences can produce different and unanticipated interpretations by listeners when combined with emotional prosody. Pseudosentences were used to avoid this confound and to compare our findings to previous cross-cultural work on multisensory emotion perception (e.g., Liu et al., 2015a, 2015b). To ensure that vocal expressions were not easier to ignore than facial expressions when processing pseudosentences, we conducted a  $2 \times 2$  ANOVA on the incongruent trials where the emotion of the vocal expression and the facial expression was different. If it is true that the vocal expressions in the present study were easier to ignore because they were meaningless, then we would expect the accuracy rate for the incongruent trials in the face task to be higher than the voice task. However, we found no significant main effect of task,  $F < 1$  (face task:  $M = 0.82$ ,  $SE = .015$ ; voice task:  $M = .81$ ,  $SE = .014$ ). This result suggests that pseudosentences and facial expressions produce similar levels of interference and that vocal expressions were not easier to ignore than facial expressions.

**Figure 1**  
*Example of Bimodal Stimuli From the Eastern (Left) and Western (Right) Cultures*



*Note.* In the left panel, the Asian face is paired with a Mandarin pseudo-sentence and in the right panel, the Caucasian face is paired with an English pseudo-sentence. The image of the Asian face, F21, is from *Taiwanese Facial Expression Image Database*, by L. F. Chen, and Y. S. Yen, National Yang-Ming University, 2007 (<http://bml.ym.edu.tw/~download/html/>). Copyright 2007 by Brain Mapping Laboratory Institute of Brain Science, National Yang Ming Chiao Tung University. Reprinted with permission. The image of the Caucasian face is from *The Karolinska directed emotional faces—KDEF (CD ROM)*, by D. Lundqvist, A. Flykt and A. Öhman, 1998, Karolinska Institute. Copyright 1998 by Karolinska Institute, Psychology section. Reprinted with permission.

### Fillers

To discourage participants from closing their eyes or muting the sound, 12 bimodal filler trials with a new set of faces and voices were added to each task. On half of the filler trials of the face task, a 500-ms duration beep was inserted in the speech stream. On half of the filler trials of the voice task, a small but detectable red dot was added on the cheek of the face.

### Procedure

The task was programmed using a combination of HTML, JavaScript, and CSS, and conducted via the internet. Experimental lists with paths for each voice clip and picture were stored in MySQL databases. The voice clips and pictures were preloaded on each webpage to ensure they were presented simultaneously at the beginning of each trial. Participants were instructed to complete the task in a quiet room using Google Chrome. Before starting the experiment, they were instructed to adjust the sound to their level of comfort.

Each trial began with a prompt indicating the current task (i.e., “judge the voice emotion” or “judge the face emotion”). For the voice task, participants were asked to judge the emotion of the voice while ignoring the face. For the face task, participants were asked to judge the emotion on the face while ignoring the voice. After clicking on the prompt, for the bimodal trials, the face and voice appeared simultaneously. The face remained on the screen for the duration of the speech. For the unimodal trials of the face task, the face appeared anywhere between 1,500 and 2,000 ms with a 100-ms interval (i.e., 1,500, 1,600, 1,700, 1,800, 1,900, and 2,000), which is

consistent with the duration that 95% of the voice stimuli fell in. For the unimodal trials of the voice task, a fixation cross appeared and remained on the screen for the duration of the speech. After the stimulus presentation, a display with five emotions in English (happiness, sadness, disgust, fear, and anger) appeared, and participants were instructed to select the emotion they perceived as quickly as possible. Once an emotion was selected, participants rated the intensity of the emotion on a scale from 0 (*not intense at all*) to 6 (*extremely intense*). For the filler trials, the beep and the red dot occurred for 500 ms within the last 700 ms of a trial. Instead of making a judgment on the emotion, participants were instructed to report whether they saw a flashing red dot on the face or heard a beep in the speech stream by clicking “yes” or “no.”

Unimodal, bimodal congruent, bimodal incongruent trials, and filler trials were intermixed and randomly presented. The familiar and unfamiliar contexts were also randomly presented within each task. The order of receiving the face or the voice task first was counter-balanced across participants. At the beginning of each task, participants were given eight practice trials, including two filler trials. There were three breaks embedded within each task. The experiment was self-paced and took approximately 60–90 min to complete.

### Statistical Analysis

The mean accuracy rates and response times (RTs) for each trial type (unimodal, bimodal congruent, bimodal incongruent) by the level of familiarity (familiar, unfamiliar), cultural group (Chinese, American), and task (face, voice) are presented in Table 2. Only correct trials were included in the RT analyses (80.7% of the data). RTs were measured from the onset of a trial until a response was made and those that fell below 500 ms or above 5,000 ms were excluded from the data (2.5% of the data). For accuracy rates, modality dominance was computed by subtracting the mean of bimodal incongruent trials from that of bimodal congruent trials. For RTs, modality dominance was computed by subtracting the mean of bimodal congruent trials from that of bimodal incongruent trials. In IBM SPSS Version 27.0 (IBM Corp), separate two-way repeated-measures analyses of variance (ANOVAs) on modality dominance were conducted on accuracy rates and RTs with the task (voice, face) as a within-subjects factor and cultural group (Chinese, American) as a between-subjects factor. All within-subjects repeated-measures results were reported with Greenhouse–Geisser corrected *p* values and all pairwise comparisons were adjusted using Bonferroni corrections.

To examine what accounts for the cultural differences in modality dominance, we examined facilitation (i.e., congruent trials vs. unimodal trials) and interference (i.e., incongruent vs. unimodal trials) effects separately. Separate three-way repeated-measures ANOVAs with the task (voice, face) and type (facilitation, interference) as the within-subject factors and cultural group (American, Chinese) as the between-subject factor were performed on accuracy rates and RTs. Bonferroni corrections were applied for all post hoc comparisons.

## Results

### Modality Dominance

#### *Familiar Context*

In the accuracy rate analyses of modality dominance, a significant interaction between group and task emerged,  $F(1, 57) = 11.37$ ,

**Table 2***Mean Accuracy Rates (ACC) and Response Times (RT) by Level of Familiarity, Task, and Cultural Groups for Each Trial Type and Effect*

Measures	Familiarity	Task	Group	Bimodal congruent	Bimodal incongruent	Unimodal	Modality dominance	Facilitation effects	Interference effects	
ACC	Familiar	Face	Chinese	.88 (.10)	.76 (.14)	.85 (.11)	.12 (.14)	.035 (.088)	.085 (.098)	
			American	.91 (.079)	.86 (.087)	.89 (.085)	.05 (.10)	.022 (.099)	.028 (.080)	
		Voice	Chinese	.90 (.083)	.84 (.096)	.89 (.062)	.06 (.087)	.008 (.064)	.052 (.083)	
			American	.90 (.072)	.78 (.12)	.84 (.10)	.13 (.096)	.067 (.14)	.059 (.16)	
	Unfamiliar	Face	Chinese	.78 (.11)	.77 (.11)	.77 (.096)	.017 (.11)	.012 (.090)	.005 (.087)	
			American	.89 (.082)	.84 (.12)	.89 (.073)	.053 (.10)	.003 (.080)	.050 (.088)	
		Voice	Chinese	.74 (.14)	.60 (.13)	.67 (.12)	.14 (.14)	.072 (.096)	.068 (.12)	
			American	.81 (.10)	.60 (.18)	.72 (.12)	.21 (.17)	.090 (.086)	.12 (.15)	
	RT	Familiar	Face	Chinese	1,488 (385)	1,619 (424)	1,526 (356)	131 (273)	38.19 (183.50)	93.12 (255.79)
				American	1,376 (309)	1,396 (308)	1,441 (257)	20 (232)	64.15 (206.05)	-44.36 (199.79)
			Voice	Chinese	1,562 (402)	1,641 (376)	1,636 (393)	79 (300)	73.52 (294.99)	5.42 (265.35)
				American	1,479 (280)	1,644 (314)	1,534 (332)	165 (230)	54.55 (190.47)	110.32 (296.13)
Unfamiliar		Face	Chinese	1,429 (303)	1,538 (278)	1,572 (372)	110 (223)	143.40 (342.80)	-33.78 (284.63)	
			American	1,410 (273)	1,350 (268)	1,401 (291)	-59 (256)	-8.75 (245.82)	-50.48 (268.53)	
		Voice	Chinese	1,618 (414)	1,762 (382)	1,756 (373)	144 (384)	138.41 (299.55)	6.06 (298.79)	
			American	1,596 (298)	1,714 (355)	1,620 (349)	118 (245)	24.63 (279.86)	93.57 (334.93)	

Note. Standard deviations are in parentheses.

$p = .001$ ,  $\eta_p^2 = .17$  (Figure 2a). Post hoc analyses revealed that the American group had a significantly larger modality dominance in the voice task than face task,  $F(1, 28) = 8.60$ ,  $p = .007$ ,  $\eta_p^2 = .24$ , 95% CI [-0.13, -0.023]. In contrast, the Chinese group had a marginally larger modality dominance in the face task than the voice task,  $F(1, 29) = 3.81$ ,  $p = .061$ ,  $\eta_p^2 = .24$ , 95% CI [-0.003, .12]. These results suggest that under familiar cultural contexts, American participants are more attuned to facial cues, while Chinese participants are slightly more attuned to vocal cues. The main effects of task and group were not significant,  $F_s < 1$ .

For the RTs analyses of modality dominance, there was a significant group-by-task interaction,  $F(1, 57) = 4.06$ ,  $p = .049$ ,  $\eta_p^2 = .067$  (Figure 2b). In the American group, the voice task ( $M = 164.87$ ,  $SE = 42.72$ ) produced a larger modality dominance than the face task ( $M = 19.78$ ,  $SE = 42.99$ ),  $F(1, 28) = 6.51$ ,  $p = .016$ ,  $\eta_p^2 = .19$ , 95% CI [-261.54, -28.62], suggesting that American participants are more influenced by facial expressions. However, in the Chinese group, there were no differences between the face and voice tasks,  $p = .51$ , 95% CI [79.09, -109.39]. All other effects were not significant,  $F_s < 1$ .

### Unfamiliar Context

When the culture was unfamiliar to both groups, the accuracy rate analyses of modality dominance produced a main effect of task,  $F(1, 57) = 35.32$ ,  $p < .001$ ,  $\eta_p^2 = .38$ , and group,  $F(1, 57) = 4.16$ ,  $p = .046$ ,  $\eta_p^2 = .068$ , 95% CI [-0.10, -0.001], but no group-by-task interaction,  $F < 1$ . The face task ( $M = .035$ ,  $SE = .014$ ) produced a smaller modality dominance than the voice task ( $M = .17$ ,  $SE = .020$ ), 95% CI [-0.19, -0.092], indicating that both groups are more influenced by facial expressions in unfamiliar contexts. Thus, when the cultural context was less familiar, both groups showed increased reliance on the visual modality (Figure 3). The Chinese group ( $M = .078$ ,  $SE = .018$ ) had an overall smaller modality dominance than the American group ( $M = .13$ ,  $SE = .018$ ).

The analyses of modality dominance on RTs yielded a main effect of task,  $F(1, 57) = 4.54$ ,  $p = .037$ ,  $\eta_p^2 = .074$ , in which participants

had a larger modality dominance in the voice task ( $M = 131.56$ ,  $SE = 41.71$ ) than face task ( $M = 26.6$ ,  $SE = 32.83$ ). All other effects were not significant,  $p_s > .080$ .

## Interference and Facilitation Effects

### Familiar Context

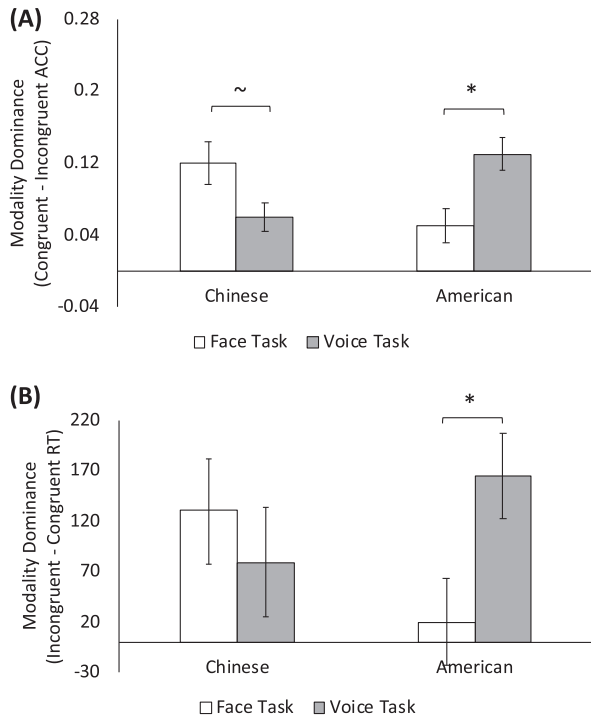
For the analyses on accuracy rates, there was a significant interaction between type and group,  $F(1, 57) = 11.37$ ,  $p = .001$ ,  $\eta_p^2 = .17$  (Figure 4). In the Chinese group, the interference effect was significantly larger than the facilitation effect,  $F(1, 29) = 4.20$ ,  $p = .045$ ,  $\eta_p^2 = .069$ , 95% CI [-0.092, -0.001]. In contrast, the facilitation and interference effects were equivalent in the American group,  $F < 1$ , 95% CI [-0.048, .045]. In other words, the American group had greater difficulty ignoring the incongruent irrelevant modality but also benefited more from the congruent irrelevant modality, while the Chinese group experienced mainly interference from the incongruent modality. No other effects and interactions were significant,  $p_s > .14$ .

For the analyses on RTs, only significant interaction between task and group emerged,  $F(1, 57) = 4.06$ ,  $p = .049$ ,  $\eta_p^2 = .067$ . In the Chinese group, there was no difference between the voice task ( $M = 9.89$ ,  $SE = 27.39$ ) and the face task ( $M = 65.65$ ,  $SE = 24.92$ ),  $F < 1$ , 95% CI [-94.95, 42.58]. However, in the American group, the voice task ( $M = 82.43$ ,  $SE = 21.36$ ) produced a larger difference score in RT than the face task ( $M = 39.47$ ,  $SE = 21.50$ ),  $F(1, 28) = 6.51$ ,  $p = .016$ ,  $\eta_p^2 = .19$ , 95% CI [2.60, 142.48], indicating greater influence from facial expressions. All other main effects and interactions were not significant,  $p_s > .088$ .

### Unfamiliar Context

For the facilitation and interference effects on accuracy rates in the unfamiliar context, there was a main effect of task,  $F(1, 57) = 35.32$ ,  $p < .001$ ,  $\eta_p^2 = .38$ , in which the voice task ( $M = .087$ ,  $SE = .010$ ) produced a larger difference score in accuracy ratings than the

**Figure 2**  
Modality Dominance Effects for Accuracy Rates (A) and Response Times (B) by Cultural Group in the Familiar Context



*Note.* Modality dominance for accuracy rates was calculated by subtracting the accuracy of the bimodal incongruent from the accuracy of the bimodal congruent condition. Modality dominance for response times was calculated by subtracting the response times of the bimodal congruent from the response times of the bimodal incongruent condition. A larger modality in the face task reflects a larger influence of the voice, whereas a larger bias in the voice task reflects a larger influence of the face. In Panel A, Chinese participants (left) had a larger modality dominance in the face task than in the voice task, whereas American participants (right) had a larger modality dominance in the voice task than in the face task. In Panel B, only the American participants had a larger modality dominance in the voice task than in the face task. Error bars represent one standard error of the mean.  
\*  $p < .05$ , ~  $p = .061$ .

face task ( $M = .018$ ,  $SE = .007$ ), 95% CI [.046, .093], suggesting a larger influence from facial expressions. There was also a main effect of the cultural group,  $F(1, 57) = 4.16$ ,  $p = .046$ ,  $\eta_p^2 = .068$ . The Chinese group ( $M = .039$ ,  $SE = .009$ ) had a smaller difference score in accuracy ratings than the American group ( $M = .065$ ,  $SE = .009$ ), 95% CI [-.051, .00]. No other main effects or interactions were significant,  $ps > .15$ .

For the facilitation and interference effects in RTs, only a main effect of task emerged,  $F(1, 57) = 4.54$ ,  $p = .037$ ,  $\eta_p^2 = .074$ , in which the voice task ( $M = 65.67$ ,  $SE = 21.02$ ) produced a larger difference score in RT than the face task ( $M = 12.60$ ,  $SE = 15.59$ ), 95% CI [3.21, 102.93]. Again, this suggests greater influence from the visual modality. All other effects and interactions did not reach significance,  $ps > .080$ . The lack of a group-by-type interaction in the analyses of accuracy rates and RTs suggests that cultural background does not affect the source of the modality dominance when the emotional information is unfamiliar.

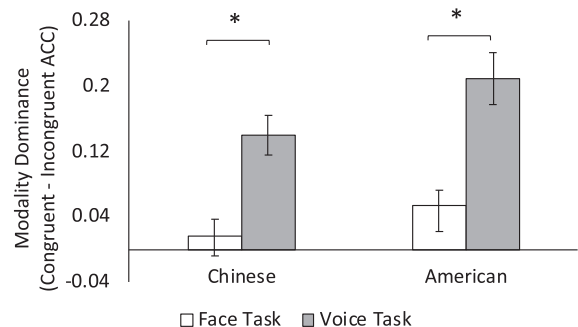
## Discussion

The present study compared modality dominance in American and Chinese participants when integrating audiovisual emotional information from their own culture and from a less familiar culture. When the audiovisual emotional information was from their own culture, we found that American participants were more sensitive to facial expressions than to vocal expressions of emotions, while Chinese participants tended to be more sensitive to vocal expressions of emotions than to facial expressions. Crucially, we observed that American participants' modality dominance was due to both facilitation and interference effects from the unattended modality. In contrast, Chinese participants' modality dominance was primarily due to interference from the unattended modality. When the audiovisual emotional information was from a less familiar culture, both groups increased their reliance on the visual modality, lending support to the modality precision theory (e.g., Choe et al., 1975; Fisher, 1968), modality appropriateness theory (O'Connor & Hermelin, 1972), and estimated precision hypothesis (Ernst & Bühlhoff, 2004). The present study suggests that modality dominance in emotion perception may be driven by different underlying processes in Eastern and Western cultures and may change depending on the familiarity of the context.

## Cultural Context and Modality Dominance

Consistent with previous cross-cultural studies on modality dominance during multisensory emotion integration (Liu et al., 2015a; Tanaka et al., 2010), the accuracy data revealed that English speakers from the United States were influenced to a greater extent by facial expressions, whereas Mandarin speakers from China were marginally more influenced by the emotion of the voice. However, in the study by Liu et al. (2015b), the Chinese participants did not show a clear preference for the auditory modality. Discrepant findings between studies may be attributed to differences in the participants' level of exposure to Western cultures. In Liu et al. (2015b) study, the Chinese participants were living in a Western

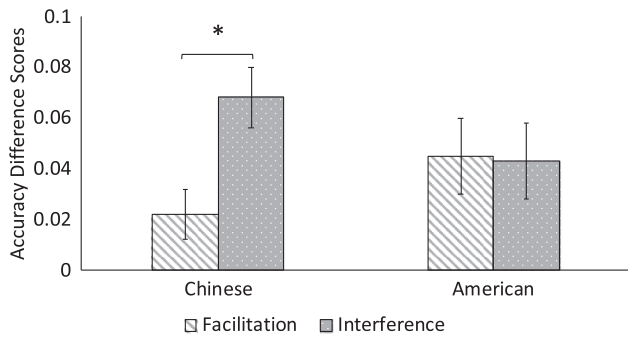
**Figure 3**  
Modality Dominance Effects for Accuracy Rates by Cultural Group in the Unfamiliar Context



*Note.* Modality dominance for accuracy rates was calculated by subtracting the accuracy of the bimodal incongruent from the accuracy of the bimodal congruent condition. Both groups of participants had a larger modality dominance on the voice task than the face task. Error bars represent one standard error of the mean.  
\*  $p < .05$ .



**Figure 4**  
Facilitation and Interference Effects for Accuracy by Cultural Group in the Familiar Context



*Note.* The face task measures the influence of the voice, and the voice task measures the influence of the face. The interference effects (raw accuracy differences between the unimodal and the bimodal incongruent condition) are larger than the facilitation effects (raw accuracy differences between the bimodal congruent and the unimodal condition) in both the face and voice tasks. Chinese participants had a larger interference than facilitation effect, whereas American participants had similar levels of facilitation and interference effects. Error bars represent one standard error of the mean.

\*  $p < .05$ .

country for an average of 10 months. In contrast, the Chinese participants in our study did not spend more than 1 month in a Western country (except for one participant who lived in a Western country for longer than 1 month but less than 3 months). Therefore, it is possible that in Liu et al.'s study (2015b), the participants' cultural immersion experience influenced their modality dominance at the time of testing. Indeed, a recent study confirms this explanation. Chinese–English bilinguals who immigrated from China to North America did not show a clear preference for the auditory modality (Chen et al., 2022).

Consistent with the estimated precision hypothesis (Ernst & Bühlhoff, 2004), both American and Chinese participants were more distracted by the visual modality when the cultural context was less familiar. This is likely because facial expressions are more easily recognized and easier to process than vocal expressions of emotions across cultures (Elfenbein & Ambady, 2002). Therefore, both American and Chinese groups showed increased visual dominance when both visual and auditory inputs become less familiar. As Collignon et al. (2008) suggested, this increased visual dominance in the less familiar context indicates that emotion perception, similar to other types of perception, might also follow a perceptual framework in which the degree of uncertainty determines the relative dependence on each sensory modality.

A closer look at the data reveals that the Chinese group decreased their auditory dominance when the input changed from a familiar to a less familiar culture (Figures 2a and 3). To examine this observation further, a two-way ANOVA of task (voice, face) and familiarity (familiar, unfamiliar) was conducted for each cultural group. The task by familiarity interaction was significant for both groups [Americans:  $F(1, 28) = 4.79, p = .037, \eta_p^2 = .15$ ; Chinese:  $F(1, 29) = 23.18, p < .001, \eta_p^2 = .44$ ]. Pairwise comparisons revealed that for the American group, modality dominance increased when the context went from familiar to unfamiliar in the voice task,  $t(28) = 2.95, p = .006, 95\%$

CI  $[-.14, -.025]$ , but not in the face task,  $t(28) < 1, p > .89, 95\%$  CI  $[-.057, .050]$ . For the Chinese participants, not only did the modality dominance increase when the context went from familiar to unfamiliar in the voice task,  $t(29) = 3.26, p = .003, 95\%$  CI  $[-.13, -.030]$ , but there was also a significant decrease in modality dominance when the context went from familiar to unfamiliar in the face task,  $t(29) = 3.83, p < .001, 95\%$  CI  $[.048, .16]$ . In other words, as the inputs became unfamiliar, both groups increased their reliance on the visual modality, but only the Chinese participants simultaneously decreased their reliance on the auditory modality.

Previous studies have shown that when affective information is presented in a nonnative language, it tends to be perceived as more emotionally distant than when presented in the native language (Ayçiçeği & Harris, 2004; Chen et al., 2015; for a review, see Pavlenko, 2005). Therefore, one possible reason for the reduced dominance of the auditory modality in the Chinese group could be because the English language was perceived to be less emotional, making it easier to ignore. Unlike Chinese participants, American participants' auditory dominance remained low regardless of their familiarity with the language, whereas their visual dominance increased when the cultural context became less familiar. This behavioral pattern is consistent with the "visual dominance" and "visual capture" effects typically observed among Western participants (Chong et al., 2015; Collignon et al., 2008; Liu et al., 2015b).

### Facilitation and Interference Effects in Modality Dominance

One limitation among previous studies is that the source of the modality dominance in each culture was unknown because the comparisons were between bimodal congruent and incongruent conditions (Liu et al., 2015b; Tanaka et al., 2010). By including two unimodal conditions in the design, the present study was able to separate the source of the modality dominance effect within each culture. When the cultural context was familiar, modality dominance in the American group could be explained in terms of facilitation and interference, while modality dominance in the Chinese group could be explained in terms of interference only. American participants experienced similar levels of facilitation and interference from the irrelevant modality, likely due to their overall reliance on the visual modality. In contrast, the Chinese group experienced more interference than facilitation from the irrelevant modality, suggesting that modality dominance in the Chinese participants is likely due to the intrusion of the voice. The interference effect could be related to the tonal characteristic of Mandarin speech, where semantic meaning is affected by the change in pitch contour. Hence, it may be more difficult for Mandarin speakers to ignore the change in pitch in the voice task (see Liu et al., 2015b, for a similar discussion). Interference and facilitation effects in an unfamiliar context did not differ between cultural groups. Therefore, the facilitation and interference effects in modality dominance vary by cultural background when presented with emotional information from one's own culture, but not when presented with emotional information from a less familiar culture.

### Limitations

While the present study reveals cross-cultural differences in modality dominance during multisensory emotion perception, it

is limited by the fact that the Chinese group had some exposure to Western cultures and some proficiency in English, due to the prevalence of Western culture and the mandatory second-language education policy in China. In contrast, the American group had very little, if any, exposure to East Asian cultures and no knowledge of Mandarin. Another limitation of the present study is that we were unable to examine the impact of emotion type due to the small number of observations per emotion and trial type. Considering previous studies suggest that the type of emotion plays a role in multisensory emotion perception (Kawahara et al., 2017; Liu et al., 2021; Takagi et al., 2015), future research should consider the effect of emotion type on multisensory emotion perception when comparing across cultures.

## Conclusion and Implication

Our findings demonstrate that multisensory emotion perception is regulated by both physical properties of the stimulus and cognitive appraisals based on existing knowledge. Cultural background and input familiarity are both examples of existing knowledge that influence the interpretation and response to emotional stimuli. These findings have implications for communicating with individuals from linguistically and culturally diverse backgrounds online and suggest that video conferencing may be more beneficial than audio-only teleconferencing. Having input available from both visual and auditory modalities may boost the accuracy of emotion perception. Our study also indicates that when meeting someone from a different and unfamiliar culture, using facial expressions to communicate emotions may be more effective given the greater reliance on the visual modality under such circumstances. We conclude that multisensory perception of emotion is a dynamic process that relies on the interaction between input familiarity and the perceivers' cultural background.

## References

- Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press.
- Ayçiçeği, A., & Harris, C. L. (2004). Bilinguals' recall and recognition of emotion words. *Cognition and Emotion*, *18*(7), 977–987. <https://doi.org/10.1080/02699930341000301>
- Chen, L.-F., & Yen, Y.-S. (2007). *Taiwanese facial expression image database*. Brain Mapping Laboratory, Institute of Brain Science, National Yang Ming Chiao Tung University. <https://bmlab.web.nycu.edu.tw/>
- Chen, P., Chung-Fat-Yim, A., & Marian, V. (2022). Cultural experience influences multisensory emotion perception in bilinguals. *Languages*, *7*(1), 12. <https://doi.org/10.3390/languages7010012>
- Chen, P., Lin, J., Chen, B., Lu, C., & Guo, T. (2015). Processing emotional words in two languages with one brain: ERP and fMRI evidence from Chinese–English bilinguals. *Cortex*, *71*, 34–48. <https://doi.org/10.1016/j.cortex.2015.06.002>
- Choe, C. S., Welch, R. B., Gilford, R. M., & Juola, J. F. (1975). The “ventriloquist effect”: Visual dominance or response bias? *Perception & Psychophysics*, *18*(1), 55–60. <https://doi.org/10.3758/BF03199367>
- Chong, C. S., Kim, J., & Davis, C. (2015). *Visual vs. auditory emotion information: How language and culture affect our bias towards the different modalities* [Conference session]. Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (pp. 46–51).
- Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–135. <https://doi.org/10.1016/j.brainres.2008.04.023>
- De Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*(3), 289–311. <https://doi.org/10.1080/026999300378824>
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotions. In J. Cole (Ed.), *Nebraska symposium on motivation* (pp. 207–282). University of Nebraska Press.
- Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*(1), 49–98. <https://doi.org/10.1515/semi.1969.1.1.49>
- 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., 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 judgments 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>
- Ernst, M. O., & Bühlhoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169. <https://doi.org/10.1016/j.tics.2004.02.002>
- Fang, X., van Kleef, G. A., & Sauter, D. A. (2019). Revisiting cultural differences in emotion perception between easterners and westerners: Chinese perceivers are accurate, but see additional non-intended emotions in negative facial expressions. *Journal of Experimental Social Psychology*, *82*, 152–159. <https://doi.org/10.1016/j.jesp.2019.02.003>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fisher, G. H. (1968). Agreement between the spatial senses. *Perceptual and Motor Skills*, *26*(3), 849–850. <https://doi.org/10.2466/pms.1968.26.3.849>
- Freides, D. (1974). Human information processing and sensory modality: Cross-modal functions, information complexity, memory, and deficit. *Psychological Bulletin*, *81*(5), 284–310. <https://doi.org/10.1037/h0036331>
- GNU Image Manipulation Program Development Team. (2018). *GIMP*. <https://www.gimp.org>
- Hawrysh, B. M., & Zaichkowsky, L. J. (1990). Cultural approaches to negotiations: Understanding the Japanese. *International Marketing Review*, *7*(2), 28–42. <https://doi.org/10.1108/EUM0000000001530>
- Hollingshead, A. B. (1975). *Four factor index of social status* [Unpublished manuscript]. Yale University. [https://sociology.yale.edu/sites/default/files/files/yjs\\_fall\\_2011.pdf#page=21](https://sociology.yale.edu/sites/default/files/files/yjs_fall_2011.pdf#page=21)
- Ishii, K., Reyes, J. A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science*, *14*(1), 39–46. <https://doi.org/10.1111/1467-9280.01416>
- Izard, C. E. (1969). The emotions and emotion constructs in personality and culture research. In R. B. Cattell (Ed.), *Handbook of modern personality theory* (pp. 496–510). Aldine Press.
- Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology*, *19*(18), 1543–1548. <https://doi.org/10.1016/j.cub.2009.07.051>
- Kawahara, M., Sauter, D., & Tanaka, A. (2017). *Impact of culture on the development of multisensory emotion perception* [Conference session]. Proceedings of the 14th International Conference on Auditory-Visual Speech Processing (pp. 109–114). <https://doi.org/10.21437/AVSP.2017-21>

- Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition and Emotion*, *16*(1), 29–59. <https://doi.org/10.1080/0269993943000121>
- Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods*, *44*(4), 1042–1051. <https://doi.org/10.3758/s13428-012-0203-3>
- Liu, P., Rigoulot, S., Jiang, X., Zhang, S., & Pell, M. D. (2021). Unattended emotional prosody affects visual processing of facial expressions in Mandarin-speaking Chinese: A Comparison with English-speaking Canadians. *Journal of Cross-Cultural Psychology*, *52*(3), 275–294. <https://doi.org/10.1177/0022022121990897>
- Liu, P., Rigoulot, S., & Pell, M. D. (2015a). Cultural differences in on-line sensitivity to emotional voices: Comparing East and West. *Frontiers in Human Neuroscience*, *9*, Article 311. <https://doi.org/10.3389/fnhum.2015.00311>
- Liu, P., Rigoulot, S., & Pell, M. D. (2015b). Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. *Neuropsychologia*, *67*, 1–13. <https://doi.org/10.1016/j.neuropsychologia.2014.11.034>
- Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska directed emotional faces—KDEF (CD ROM)*. Karolinska Institute, Department of Clinical Neuroscience, Psychology Section.
- Marian, V. (2023). *The power of language*. Penguin Random House.
- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967. [https://doi.org/10.1044/1092-4388\(2007\)067](https://doi.org/10.1044/1092-4388(2007)067)
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253. <https://doi.org/10.1037/0033-295X.98.2.224>
- Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*(2), 215–221. <https://doi.org/10.3758/BF03212421>
- Masuda, T., Ellsworth, P. C., Mesquita, B., Leu, J., Tanida, S., & Van de Veerdonk, E. (2008). Placing the face in context: Cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology*, *94*(3), 365–381. <https://doi.org/10.1037/0022-3514.94.3.365>
- Matsumoto, D., Takeuchi, S., Andayani, S., Kouznetsova, N., & Krupp, D. (1998). The contribution of individualism vs. collectivism to cross-national differences in display rules. *Asian Journal of Social Psychology*, *1*(2), 147–165. <https://doi.org/10.1111/1467-839X.00010>
- Matsumoto, D., Yoo, S. H., Fontaine, J., Anguas-Wong, A. M., Arriola, M., Ataca, B., Bond, M. H., Boratav, H. B., Breugelmans, S. M., Cabecinhas, R., Chae, J., Chin, W. H., Comunian, A. L., Degere, D. N., Djunaidi, A., Fok, H. K., Friedlmeier, W., Ghosh, A., Glamcevski, M., ... Grossi, E. (2008). Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of Cross-Cultural Psychology*, *39*(1), 55–74. <https://doi.org/10.1177/0022022107311854>
- McCarthy, A., Lee, K., Itakura, S., & Muir, D. W. (2006). Cultural display rules drive eye gaze during thinking. *Journal of Cross-Cultural Psychology*, *37*(6), 717–722. <https://doi.org/10.1177/0022022106292079>
- McCarthy, A., Lee, K., Itakura, S., & Muir, D. W. (2008). Gaze display when thinking depends on culture and context. *Journal of Cross-Cultural Psychology*, *39*(6), 716–729. <https://doi.org/10.1177/0022022108323807>
- Mesquita, B., Boiger, M., & De Leersnyder, J. (2016). The cultural construction of emotions. *Current Opinion in Psychology*, *8*, 31–36. <https://doi.org/10.1016/j.copsyc.2015.09.015>
- O'Connor, N., & Hermelin, B. (1972). Seeing and hearing and space and space and time. *Perception & Psychophysics*, *11*(1), 46–48. <https://doi.org/10.3758/BF03212682>
- Pavlenko, A. (2005). *Emotions and multilingualism*. Cambridge University Press.
- Pell, M. D. (2006). Cerebral mechanisms for understanding emotional prosody in speech. *Brain and Language*, *96*(2), 221–234. <https://doi.org/10.1016/j.bandl.2005.04.007>
- Pell, M. D., Paulmann, S., Dara, C., Allasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*(4), 417–435. <https://doi.org/10.1016/j.wocn.2009.07.005>
- Sanchez-Burks, J., Lee, F., Choi, I., Nisbett, R., Zhao, S., & Koo, J. (2003). Conversing across cultures: East-West communication styles in work and nonwork contexts. *Journal of personality and social psychology*, *85*(2), 363–372. <https://doi.org/10.1037/0022-3514.85.2.363>
- Schreuder, E., van Erp, J., Toet, A., & Kallen, V. L. (2016). Emotional responses to multisensory environmental stimuli: A conceptual framework and literature review. *SAGE Open*, *6*(1), 1–19. <https://doi.org/10.1177/2158244016630591>
- Takagi, S., Hiramatsu, S., Tabei, K., & Tanaka, A. (2015). Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality. *Frontiers in Integrative Neuroscience*, *9*, Article 1. <https://doi.org/10.3389/fnint.2015.00001>
- Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., & de Gelder, B. (2010). I feel your voice. Cultural differences in the multisensory perception of emotion. *Psychological Science*, *21*(9), 1259–1262. <https://doi.org/10.1177/0956797610380698>
- Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2013). Dissociating task difficulty from incongruence in face-voice emotion integration. *Frontiers in Human Neuroscience*, *7*, Article 744. <https://doi.org/10.3389/fnhum.2013.00744>
- Yuki, M., Maddux, W. W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, *43*(2), 303–311. <https://doi.org/10.1016/j.jesp.2006.02.004>

Received April 4, 2022

Revision received November 11, 2022

Accepted November 18, 2022 ■