Bilingualism: Language and Cognition

cambridge.org/bil

Introducing grip force as a nonverbal measure of bilingual feelings

Dieter Thoma, Julia Hüsam and Kimberley Wielscher

Department of English Linguistics, University of Mannheim, Mannheim, Germany

Research Article

Cite this article: Thoma D, Hüsam J, Wielscher K (2023). Introducing grip force as a nonverbal measure of bilingual feelings. *Bilingualism: Language and Cognition* **26**, 125–137. https://doi.org/10.1017/S1366728922000396

Received: 28 June 2021
Revised: 15 May 2022
Accepted: 15 May 2022
First published online: 15 June 2022

Keywords:
bilingualism; emotion; feelings; pupillometry; dynamometer

Address for correspondence:
Dieter Thoma
University of Mannheim
Schloss
68131 Mannheim
Germany
thoma@uni-mannheim.de

Abstract

Bilinguals' emotions can vary in intensity with the language of a stimulus. Yet, extant research has somewhat surprisingly accepted inconsistent results from implicit nonverbal and explicit verbal emotion measures. To date, it is unclear if this inconsistency recurs to conceptual or methodological differences. We therefore investigated if squeezing a handheld dynamometer is a valid nonverbal, "visceral" alternative to self-reported language-dependent feelings by comparing explicit ratings to neuro-physiological emotional reactions. We replicated two pupillometry experiments inducing language-dependent emotions through sentence reading (Study 1) and listening to narrative video commercials (Study 2) of low and high emotionality in the first or second language. Pupillometry confirmed that bilinguals are more sensitive to the low-high emotionality contrast in their first than second language. Grip force (but not duration) mirrored these findings, whereas verbal ratings did not. We thus recommend grip force as a new attentional, nonverbal measure for bilingualism research.

1. Introduction

Emotional experiences, or feelings, are intertwined with language (e.g., Lindquist, Satpute & Gendron, 2015), which may lead to differential emotional reactions in unbalanced bilinguals' dominant first language (L1) and weaker second language (L2) (e.g., Caldwell-Harris & Ayçiçeği-Dinn, 2009). While bilinguals' language-dependent emotions have been a topic of continued interest over the past decades, researchers have faced the latent problem that language is part of the research problem and the method when bilinguals verbally report on their feelings. In fact, verbalization shapes initially nonverbal dimensions of emotion (Kassam & Mendes, 2013), and the choice of the language used to ask bilinguals about their feelings influences the response (Marian & Neisser, 2000). In search for nonverbal measurement alternatives in monolingual contexts, Creswell, Sayette, Schooler, Wright and Pacilio (2018) recently showed that the voluntary muscle force applied when squeezing a handheld dynamometer predicts later eating behavior better than verbal ratings of hunger. Therefore, our research extended this novel methodology. Here, we explored if reporting the intensity of feelings nonverbally via grip force and duration offers a more valid assessment of language-dependent feelings than verbal ratings by comparing these explicit ratings to pupil reactions induced by emotional intensity.

1.1. Emotions, feelings, and language

Emotions can be understood as experiential, physiological, and behavioral reactions to personally relevant external and internal stimuli (Mauss & Robinson, 2009). Researchers have developed several emotion theories which conceptualize them, for example, as discrete states such as happiness, anger, or fear (e.g., Ekman, 1992) or as *ad hoc* combinations of a few affectual dimensions, most importantly of how pleasurable (positive vs. negative valence) and intense (low vs. high arousal) individuals experience an emotional reaction (e.g., Russell, 2003; Russell & Barrett, 1999). Although valence and arousal are theoretically distinct, empirical ratings of words (e.g., Bradley & Lang, 1999) and pictures (e.g., Kurdi, Lozano & Banaji, 2017) suggest a skewed U-shaped relationship between them, since strongly negative and strongly positive stimuli also tend to be more arousing and strongly negative ones most arousing.

According to the dimensional theory of constructed emotion (Barrett, 2006; Gendron, Lindquist, Barsalou & Barrett, 2012), language (re-)constructs or shapes emotion by providing category labels such as "happiness", "anger", and "fear" that allow us to bundle affectual sensations as self-conscious experiences (Lindquist et al., 2015). While a recent analysis of almost 2500 languages (Jackson et al., 2019) found that all languages differentiate emotions along the basic affect dimensions of valence and arousal, it also showed that their semantics of specific emotion terms such as "anger" and "shame" vary. Linguistic effects on emotion construction and perception may therefore depend on language-specific emotion vocabulary (Majid, 2012).

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

However, not all dimensions of emotion can be expressed verbally because emotions are not only represented or constructed in the brain but partially embodied (Niedenthal, 2007), i.e., perceptually, somatoviscerally, and motorically re-experienced. In a very broad sense, feelings are mental experiences of emotions that range from the perception of motivational, or VISCERAL, bodily sensations, i.e., "gut feelings" such as hunger and thirst, to cognitively complex states such as compassion and gratitude (Damasio & Carvalho, 2013).

Apart from its constructive function, the verbalization of emotions, or "affect labelling", i.e., translating feelings into verbal or other symbolic labels in self-report, typically reduces the intensity of explicit and implicit emotions (for review, see Torre & Lieberman, 2018). For example, Kassam and Mendes (2013) asked participants to solve mathematical tasks that induced a non-self-conscious emotion (anger), a self-conscious one (shame), or no emotion. Half of the participants verbally rated their feelings on numerical scales with verbal anchors, whereas the other half completed an unrelated control questionnaire. The authors measured a range of cardiovascular reactions indicating that participants responded to the emotion induction, e.g., with higher heart rates during anger and shame, but these reactions were weaker when participants rated their feelings verbally.

More generally, affect labelling has been interpreted as a case of VERBAL OVERSHADOWING (Schooler & Engstler-Schooler, 1990), which refers to the distortion of a range of initially nonverbal experiences such as taste preferences (Wilson & Schooler, 1991), body movements (Defrasne Ait-Said, Maquestiaux & Didierjean, 2014), and emotions (Kassam & Mendes, 2013; Lieberman et al., 2007). Verbal overshadowing has been alternatively attributed to insufficient verbal abilities that result in a distorted perception of non-verbal experiences (e.g., Schooler & Engstler-Schooler, 1990) or to a disadvantageous shift from a holistic and intuitive to an analytic and deliberate processing and decision mode in tasks where intuition would be more apt (e.g., Dijksterhuis, 2004).

In a recent approach to minimize verbal overshadowing, Creswell et al. (2018) validated reporting hunger as a visceral feeling by squeezing a handheld dynamometer, compared to rating it on a verbal self-report scale. They asked participants who had not eaten for at least four hours to report on their hunger before they were presented with and allowed to eat as much popcorn as they wished. In three conditions, participants indicated their hunger by rating it on a numerical self-report scale with verbal anchors ranging from 0 (= "not hungry at all") to 100 (= "the most intense hunger I have ever felt") and then, by squeezing a dynamometer (verbal first condition). Alternatively, they started with squeezing the dynamometer before the verbal rating (nonverbal first) or reported hunger only nonverbally. Results showed that the dynamometer responses in the nonverbal and nonverbal first condition, which were a compound measure of grip force and duration, predicted actual eating behavior well ($r > .50$) and better than verbal hunger ratings ($r < .17$). However, if the nonverbal rating followed the verbal one, it was an equally poor predictor of popcorn consumption. In a replication (Creswell et al., 2019), squeeze recordings of the urge to smoke also tended to be more reliable predictors than verbal ratings of latencies to smoke, i.e., how long smokers would wait to light a cigarette after at least 6 hours of abstinence. Theoretically, one could either argue that the attempt to verbalize the nonverbal visceral experience of hunger or urge to smoke is difficult, which yields unreliable measurements, or that the verbalization induces less intuitive and more deliberate decision processes during self-report. Creswell et al. (2018, 2019) concluded that applying a dynamometer as a measurement instrument for self-conscious experiences is a promising alternative to verbal self-reports whenever verbalization, i.e., language, modulates the feelings to be reported. Bilinguals' language-dependent emotions may be such a case.

1.2. Language-dependent emotions

Cross-experimentally, evidence for language-dependent emotions in unbalanced bilinguals – typically reduced ones in L2 – has been found in a variety of different measures such as electrodermal activity (Caldwell-Harris & Ayçiçeği-Dinn, 2009; Harris, Ayçiçeği-Dinn & Gleason, 2003), pupillometry (García-Palacios et al., 2018; Iacozza, Costa & Duñabeitia, 2017; Thoma, 2021; Thoma & Baum, 2019; Toivo & Scheepers, 2019), event-related potential (*ERP*) in brain activity (Jończyk, Boutonnet, Musiał, Hoemann & Thierry, 2016; Opitz & Degner, 2012; Sianipar, Middelburg & Dijkstra, 2015), word-based Stroop and Simon tasks (Sheikh & Titone, 2016; Sutton, Altarriba, Gianico & Basnight-Brown, 2007), hypothetical decision making (Costa et al., 2014; Keysar, Hayakawa & An, 2012) and explicit self-report (Caldwell-Harris & Ayçiçeği-Dinn, 2009; Dewaele, 2004, 2008; Imbault, Titone, Warriner & Kuperman, 2021; Vélez-Uribe & Rosselli, 2019). Explanations for bilinguals' language-dependent emotions are still debated (see Thoma, 2021; Williams, Srinivasan, Liu, Lee & Zhou, 2020 for review and empirical testing). In short, contextual accounts emphasize the role of different contexts of language learning and use, age of acquisition (AoA), and cultural expectations, whereas processing accounts assume structurally similar but less fluent emotion processing in L2 than L1.

Further, there is little bilingual emotion research using multi-measure designs, and in these studies the findings derived from neuro-physiological and self-report measures do not always align. We limit the following review to skin conductance response (SCR) and pupillometry studies, but similar discrepancies have been observed between ERP and verbal responses (Jończyk et al., 2016; Wu & Thierry, 2012). For example, Harris et al. (2003) asked Turkish–English bilinguals to rate neutral and differentially valenced emotion words on a 1–7 scale for unpleasantness and simultaneously monitored SCR via fingertip electrodes. The SCR showed main effects of language and valence with stronger automatic reactions in L1 Turkish most clearly for taboo words and reprimands, while the verbal ratings were very similar across languages. In a replication (Caldwell-Harris & Ayçiçeği-Dinn, 2009, Exp. 1), the authors observed main yet no interaction effects of language and emotion word category in SCR, although the two factors yielded main and interaction effects in the verbal rating. The authors did not discuss these discrepancies. Using similar measures, Jankowiak and Korpal (2018) investigated if emotion-laden narratives in L1-Polish induced more intense SCR and verbal emotion ratings than in L2-English. Whereas SCR indicated physiologically stronger reactions in L1 than L2, the verbal ratings on semantic differentials with emotion adjectives as anchors were similar across languages. In their discussion, the authors questioned the reliability of their verbal measure as it may have been biased by social desirability or intimidation.

Compared to SCR, pupillometry is a relatively new and even less intrusive method for studying language-dependent emotions. Changes in pupil diameter result from the interplay of the sympathetic and parasympathetic autonomic nervous systems in

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

respectively prompting the dilator and sphincter pupillae muscles in the iris to widen or constrict the pupil (McDougal & Gamlin, 2015). Beyond its bi-directional sensitivity to differences in brightness (or luminance), the pupil reliably dilates as a function of the intensity of cognitive and affective processing (Beatty, 1982; Bradley, Miccoli, Escrig & Lang, 2008; Janisse, 1973). Although the pupil is sensitive to differences in valence and arousal, its dilation shows a skewed U-shaped response pattern similar to behavioral ratings (Bradley et al., 2008). Accordingly, dilation increases with arousal, is usually indicative of the intensity of valence but cannot distinguish its polarity (Collins, Ellsworth & Helmreich, 1967). This led pupil researchers to work with "emotional intensity" (Mathôt, Grainger & Strijkers, 2017) or "emotionality" (Citron, Gray, Critchley, Weekes & Ferstl, 2014) defined as a stimulus' absolute positive or negative deviation from neutral valence. As L2 processing is associated with increased cognitive load relative to the more fluent L1 (Oganian, Korn & Heekeren, 2016; J. Schmidtke, 2014), pupillometry studies on bilingual emotions should disentangle cognitive and affective causes of pupil dilation. In addition to a baseline correction for individual differences in pupil size and responsiveness, the multi-measure studies reviewed in the following have consequently implemented factorial designs crossing language and arousal/emotional intensity.

Iacozza et al. (2017) used a language (L1 Spanish vs. L2 English) by "emotionality" (neutral with low arousal vs. negative with high arousal) pupillometry design. Participants read sentences with emotion-manipulated adjective-noun phrases on an eye-tracked screen and instantly rated the "emotional impact" of each sentence (1 = "low, neutral impact"; 7 = "high, negative impact") in the language of the item. Pupil reactions were stronger to sentences with normatively higher emotionality. There was no main effect of language but a two-way interaction, such that the difference between neutral-low and negative-high sentences was significantly larger in L1 than in L2. However, verbal ratings showed a comparable emotionality effect in both languages. The authors only discussed the pupil-based language-emotionality interaction.

García-Palacios et al. (2018) varied the language (L1 Spanish vs. L2 English) of a fear conditioning paradigm cueing the expectation of a threat or not (neutral condition) during specific trials in a visual task. The threat, a mild electric shock, never occurred in fact. The authors measured SCR and pupil dilation relative to pre-trial baseline activity. Both measures showed a main effect of the fear condition with stronger reactions after a threat, a main effect of language with a higher level of reactivity in L2 and, importantly, a significant interaction such that the threat-neutral difference was more pronounced in L1 than L2. Interestingly, the two psychophysiological measures aligned, but the interaction effect was stronger and more reliable in the pupil data.

Toivo and Scheepers (2019) presented German–English and Finnish–English late bilinguals and an L1 English control group with an eye-tracked word recognition judgement task containing low vs. high arousal words statistically matched on various other characteristics including valence. Note that high-arousal words still were 1.3 to 1.9 times more negative than low-arousal on average. Participants started the task in L1 or L2. Notably and similar to Iacozza et al. (2017), participants' pupil dilation was more sensitive to the arousal manipulation in their L1 than L2. Accuracy in the word judgement task was >95% in all language conditions and practically identical for low and high arousal, so that it did not warrant conclusions about language-dependent emotions (speed was not measured).

Finally, Thoma and Baum (2019, Study 2) measured German–English bilinguals' pupil reactions during a sad high arousal advertising video which induced emotional reactions with a narrative in L1 or L2. Pupil dilation computed relative to a neutral low arousal street surveillance video was stronger in L1 than L2 with and without additional cognitive load indicating higher arousal than in L2. After watching the video, participants also reported their feelings on a pictorial valence scale with emoticons (Bradley & Lang, 1999). Pupil dilation and self-reports correlated moderately, but the authors did not conduct inferential analyses on self-reported feelings.

In sum, even though emotions should be studied with a multi-method approach, different measures of emotion cannot necessarily be used and interpreted interchangeably (Mauss & Robinson, 2009). It may well be that the effects of emotion-evocative linguistic stimuli differ systematically between emotion measures, but there is scarce evidence if and how these measurement issues affect bilingual emotion research.

1.3. Linguistic measurement equivalence

Research on language-dependent emotions usually uses stimuli that do not cause comprehension problems in any language condition. Explicit verbal measures of emotion may still be biased by the language of the question and response scales themselves (Dewaele, 2013; Weijters, Puntoni & Baumgartner, 2017). Some of the earliest research on bilinguals' emotions recognized that their narratives of autobiographical memories were more emotional, e.g., in the valence and arousal intensity of their words, if the language of encoding and retrieval were congruent (Marcos, 1976; Marian & Neisser, 2000). This illustrates that the measurement and conceptual problem are intertwined. Language choice could simply influence the accessibility of memory, or it could evoke culture-specific mindsets and expectations in bilinguals. The latter concept has been referred to as cultural frame switching (Benet-Martínez, Leu, Lee & Morris, 2002). In many studies, the contribution of the linguistic measurement as such to the observed effects is unclear. For example, in a study by Dylman and Bjärtå (2019), bilinguals reported higher levels of emotional distress after reading negative texts in L1 Swedish and answering the following questions in the congruent language. When they answered in L2 English, however, they reported lower levels of distress. The verbal ratings did not allow the researchers to attribute this asynchrony to differences in emotion or in L2 processing load. Further, Noriega and Blair (2008) found Hispanics in the United States to verbally report more emotional thoughts after reading same-content advertising slogans in Spanish than English, but they did not ask them to report on English in Spanish and vice versa. Finally, in a survey study (Ross, Xun & Wilson, 2002), Chinese–English bilinguals described themselves as more collective-minded and reported lower self-esteem on a verbal rating scale ("strongly disagree" – "strongly agree") in Chinese than English. Admittedly, cultural frame switching can also be induced with nonverbal stimuli (Benet-Martínez et al., 2002), and some studies suggest that linguistic measurement inequivalence and cultural frame switching are less of an issue in "fully balanced" bilingual populations (e.g., Schwartz et al., 2014). We still do not know if the exemplary findings reported above would have been similar if participants had responded nonverbally.

In a series of nine studies, Langhe, Puntoni, Fernandes, and van Osselaer (2011) tested the influence of the language of verbal

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

scales on bilingual emotion ratings. The authors argued that bilinguals interpret emotion-word scale anchors such as "happy" vs. "sad" less intensely in L2, so that they perceive them as more similar, mentally contract the scale between them and consequently, provide more extreme ratings in L2. They demonstrated this so-called anchor contraction effect (ACE) for ratings of nonverbal and language-specific verbal emotional stimuli, which participants consistently rated more emotionally intense on L2 scales. The ACE occurred independently of bilinguals' lower emotion ratings for L2 stimuli on L1 scales and on unipolar scales. However, when the authors added emoticons or dots in shades of red to the numerically labelled scale points between the two verbal anchors on bipolar scales, the ACE disappeared.

All in all, verbalization as such (Dhar & Gorlin, 2013; Dijksterhuis, 2004) and L2 use (Costa et al., 2014; Keysar et al., 2012) can induce more deliberate decision making which can alter self-reported emotions, i.e., feelings, relative to their nonverbal experience. Pictorial labels such as emoticons or colors on rating scale points can reduce disruptive effects of language (Langhe et al., 2011), but as they are also symbolic, they can still distort non-cognitive components of emotion (Kassam & Mendes, 2013). There is also no guarantee that bilingual participants will not subconsciously translate symbolic information, e.g., the facial expression of an emoticon, into divergent language-specific emotion vocabulary (Barbieri, Kruszewski, Ronzano & Saggion, 2016; Gendron et al., 2012). Consequently, it is important to validate verbal self-reports on emotion against nonverbal measurement alternatives.

2. The present study

The aim of the present research was to validate the force and duration of squeezing a dynamometer as nonverbal, visceral measures of self-reported language-dependent emotions experienced by bilinguals. Therefore, we replicated and extended two prior experiments (Iacozza et al., 2017; Thoma & Baum, 2019) using a multi-measure approach. This allowed us to cross-validate physiological (pupil dilation), visceral (grip force and duration), and verbal (self-report rating scale) measures of emotional reactions to stimuli presented either in the bilinguals' L1 or L2. In both experiments, we tested three hypotheses.

We developed two replication hypotheses following the original research (Iacozza et al., 2017; Thoma & Baum, 2019). First, we hypothesized that unbalanced bilinguals' emotions are stronger in L1 than in L2 (H1). Second, we assumed that stimulus emotionality moderates the language effect, such that negatively valenced high-arousal stimuli cause larger L1 vs. L2 emotion differences, compared to neutral-low arousing ones (H2). Finally, our main hypothesis followed from the observation that verbal reporting may alter (visceral) feelings (Creswell et al., 2018; Kassam & Mendes, 2013; Lieberman et al., 2007), which are mental experiences of emotions (Damasio & Carvalho, 2013). Emotions, in turn, are reflected in pupil dilation as an automatic – and thereby, unbiased – reaction of the sympathetic nervous system to emotional stimuli (Bradley et al., 2008). Accordingly, we hypothesized that visceral measures (grip force and duration) are better predictors of language-dependent implicit emotional reactions (pupil dilation) than verbal emotion ratings (H3).

To test the hypotheses, both experiments used a 2 (language: L1 German vs. L2 English) by 2 (emotionality: low vs. high) design with language as between and emotionality as within-subjects factor. We operationalized the emotionality conditions with stimuli of neutral valence with low arousal vs. negative valence with high arousal because prior research showed (a) the largest behavioral differences in valence- and arousal-induced intensity of emotional reactions between these conditions (Bradley & Lang, 1999; Kurdi et al., 2017) and (b) that pupil dilation is more sensitive to this contrast in L1 than L2 (García-Palacios et al., 2018; Iacozza et al., 2017; Thoma & Baum, 2019; Toivo & Scheepers, 2019). We estimated a sample size of about 30 per language condition and experiment *a priori*, given that Creswell et al. (2018) observed relatively strong correlations between their dynamometer measure of hunger and popcorn consumption in the nonverbal and nonverbal first groups ($r > .5$, $n = 35$). Due to the COVID-19 pandemic, we had to minimize traffic in the lab and the same participants took part in the same language conditions of both studies. They started with the first experiment which took about 35 minutes and had a 10-minute break afterwards, while the experimenter vented the room. Next, they completed the second experiment and the control and background tasks which took another 25 minutes before they were debriefed. Both experiments complied with the Declaration of Helsinki.

3. Study 1: Reading-induced emotions

Study 1 replicated the Spanish–English pupillometry experiment by Iacozza et al. (2017) in a German–English setting, where participants verbally rated the emotionality of individual sentences. We extended the task in that participants reported their perceived emotional intensity prior to each verbal rating by squeezing a dynamometer.

3.1. Method

Participants

The study included 66 participants with a mean age of 22.56 years ($SD = 7.87$) who took part for course credit or €10. All 45 female and 21 male participants were Germans, AoA: $M = .33$, $SD = .95$, who had learned English as a foreign language in school for at least nine years, AoA: $M = 7.68$, $SD = 1.76$. We tested their degree of bilingualism with a self-assessment and lexical decision task as well as with a language-learning-and-use questionnaire (all adopted from Thoma & Baum, 2019).

According to their mean self-assessment of speaking, reading, writing, listening, and grammar skills on 7-point scales, participants' proficiency in L1 German ($M = 6.47$, $SD = .65$) was significantly higher than in L2 English, $M = 5.76$, $SD = .69$; $t(65) = 7.24$, $p < .001$. Similarly, the lexical decision task, which contained 120 German and then 120 English items, indicated more automatic lexical access (coefficient of variation) in their L1 ($M = .24$, $SD = .05$) than in their L2, $M = .26$, $SD = .05$; $t(65) = -4.16$, $p < .001$. Finally, the language learning and use questionnaire with 17 items in each language confirmed that the participants were unbalanced bilinguals, dominant in German yet with advanced English proficiency, L1: $M = .71$, $SD = .07$, L2: $M = .36$, $SD = .07$; $t(65) = 31.94$, $p < .001$.

Materials

The experiment comprised 16 neutral-low and 16 negative-high valence-arousal sentences per language condition to manipulate low vs. high emotionality. Following Iacozza et al. (2017), we created 16 incomplete sentential frames in English into which we inserted either low or high emotionality words as exemplified in

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

(1). These sentences were then translated into German as displayed in (2) by highly proficient bilinguals.

(1) a. The ongoing consequences of global warming change thousands of lives.
b. The traumatic consequences of global warming destroy thousands of lives.

(2) a. Die anhaltenden Folgen der globalen Erwärmung verändern tausende Leben.
b. Die traumatischen Folgen der globalen Erwärmung zerstören tausende Leben.

English and German negative-high arousal words were selected from the Affective Norms for German Sentiment Terms (ANGST, D. S. Schmidtke, Schröder, Jacobs & Conrad, 2014) which provides norms for translation equivalents for the Affects Norms for English Words (ANEW, Bradley & Lang, 1999) and from (Iacozza et al., 2017). Neutral-low arousal words were on none of these lists.

We validated the emotional intensity of the sentences in a norming study with a different sample of 25 different L1 German and 26 L1 English university students. They read the neutral-low and negative-high version of the sentences in L1 and rated their valence and arousal on 5-point pictorial scales (SAM, Bradley & Lang, 1994). The results confirmed the intended difference in valence-induced intensity, $F(1, 48) = 118.17$, $p < .001$, as well as an arousal difference, $F(1, 48) = 155.90$, $p < .001$), whereas there was no significant language difference or interaction, all $F < .75$, $p > .38$. A list of the sentences, individual and summary norms is included in Appendix A. The sentences were also matched in the number of words across the four conditions with a range of $M = 11.25–11.50$, $SD = 2.57–3.05$, all $F < .05$, $p > .83$. They were framed as authentic advertising slogans either for a supermarket promoting their commitment to sustainability or an NGO promoting their commitment to nature conservation because we expected our student-aged population to show high interest in these topics. To further engage participants in the task, the slogans were presented in combination with advertising context-related, same-size pictures in the background. The pictures were identical in low and high intensity trials and across language conditions to counterbalance for parasympathetic pupil reactions in response to luminance differences (Bradley et al., 2008).

Procedure

Participants were invited in an email newsletter and tested individually in the eye-tracking lab. They were randomly assigned to the German ($n = 33$) or English ($n = 33$) condition. The experimenter informed them that the study focused on emotionality in advertising. After providing written informed-consent, participants were seated approximately 70 centimeters in front of a 24”-computer screen surveyed by a remote SMI RED 500 Hertz eye-tracking system. Room lightning was artificial and constant at 590 lux. We positioned a handheld dynamometer on the side of the participant’s dominant hand on a small tray at waist level and partly under the table supporting the eye-tracked screen, so that participants could comfortably grab the dynamometer handle yet were not tempted to look at the device. We used a microFET HandGRIP dynamometer from Hoggan Scientific that measures maximum grip force in Newton and grip duration in seconds and transmits data via a blue tooth interface (accuracy ± 0.4 N, operational range 0–880 N).

Grip size of the dynamometer handle was adjusted for each individual, and hand dominance was initially determined in terms of strongest grip force. In a picture-based practice task, participants were further familiarized with the hardware use and procedure. The hardware instructions were in German, while all subsequent tasks were in the language of the experimental condition. In the practice task, participants saw two high and two low arousal pictures sampled from the Open Affective Standardized Image Set (OASIS, Kurdi et al., 2017). Each picture was displayed for 5 s, and the experimenter asked “how emotional” it was. Participants learned to squeeze the dynamometer harder and longer to indicate stronger emotional reactions (Creswell et al., 2018) and to proceed to the next screen by fixating the lower right corner of it for 3 s. After each visceral emotion rating, they reported their feelings on a 7-point rating scale with the verbal anchors “unemotional” and “emotional” which are identical cognates in German and English.

The main sentence rating task started with a 9-point eye-tracking calibration that was repeated until position accuracy was ≤ .5 degrees of visual angle. Figure 1 illustrates the procedure of a trial repeated for all 32 sentences. It started with a 1-s fixation inter-trial screen followed by the display of the target sentence in its advertising frame. Participants proceeded to the next screen by fixating its lower right corner for 3 s. It showed the question “How emotional was this slogan?” and the instruction to respond by squeezing the dynamometer. The last screen displayed the same question in combination with a verbal rating scale described above. The participant said the number and the experimenter recorded it. We started with the visceral measures because Creswell et al. (2018) observed stronger correlations between the nonverbal and verbal measures of hunger in this order and to avoid the verbal overshadowing effect they reported.

Analyses

We programed a Python script to clean and aggregate the pupillometry data according to conventional procedures (Kinner et al., 2017; Lemercier et al., 2014; Thoma & Baum, 2019). The script deleted recordings during blinks and computed a mean diameter from the left and right pupil size for each measurement. To calculate pupil dilation, each participant’s mean pupil size within the last 500 ms during the directly preceding fixation screen was subtracted from each pupil diameter value recorded during the display of the target stimulus. Finally, the script downsampled the dilation values from 500 Hz to 1-s epochs. To account for dwell time outliers, we cut off pupil measures recorded after 14 s per slogan screen excluding 2.10% of the data.

Creswell et al. (2018, 2019) analyzed the Area Under the Curve (AUC) as a conjoint dynamometer-generated visceral measure multiplicatively integrating grip force and duration. However, we report separate analyses for maximum grip force and total duration because pretests indicated that participants immediately understood the instruction to squeeze more forcefully yet asked questions about squeezing longer to indicate emotional intensity. Grip force and duration measures were log-transformed to normalize them. To allow for direct comparisons with Creswell et al.’s original studies using the dynamometer, we included AUC-based analyses in Appendix C (they replicated the grip force findings).

To test our replication hypotheses (H1 and H2), we fitted linear mixed-effects regression models (LMMs) using the *lmer* function from the *lme4* package (Bates et al., 2021) in R Studio (RStudio Team, 2021) for all ratio measures of emotion (pupil

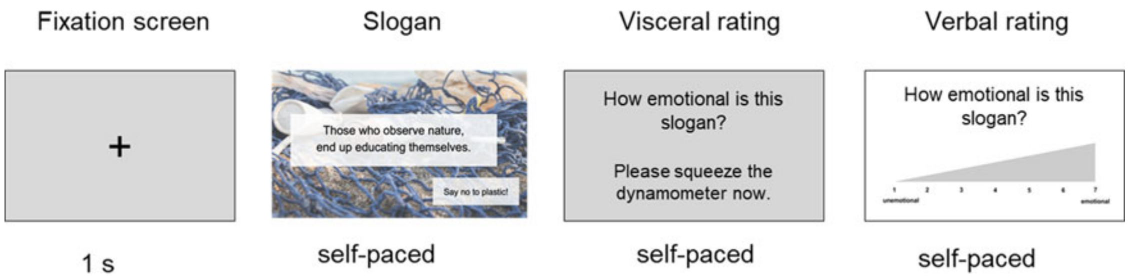
Fixation screen
Slogan
Visceral rating
Verbal rating
+
Those who observe nature, end up educating themselves.
Say no to plastic!
How emotional is this slogan?
Please squeeze the dynamometer now.
How emotional is this slogan?
1
2
3
4
5
6
7
unemotional
emotional
1 s
self-paced
self-paced
self-paced

Figure 1. Procedure of a target trial in the sentence emotion rating task in the English condition.

dilation, grip force, grip duration). For the ordinal verbal ratings, we used a *glmer* function from the Poisson family. All models included by-participant and by-item random intercepts and slopes for emotionality (as a within-subjects variable) and random intercepts for items (Barr, Levy, Scheepers & Tily, 2013). Categorical fixed factors were deviation-coded (-.5 vs. .5), and continuous predictors were centered to achieve comparable effect sizes. In addition to language, emotionality, and their interaction, all models included sex as a control variable because males tend to have stronger grip force (Corrêa et al., 2020). The pupil dilation models further entailed epoch as a predictor since pupil reactions evolve over time (Lemercier et al., 2014). Further interactions were included if they improved model fit (which they did not). Estimates with $|t| \geq 1.96$ fall within the 95% confidence interval and are considered significant, which was confirmed by *p*-values estimated with *lmerTest* (Kuznetsova, Brockhoff & Christensen, 2017). To test our research hypothesis (H3), we used likelihood ratio tests (Pinheiro & Bates, 2009) analyzing if grip force, grip duration, and verbal ratings improved the model fit relative to a null model predicting pupil dilation from epochs and sex alone in three separate models with the same random-effect structure as described above.

3.2. Results and discussion

Pupil dilation

The LMM predicting pupil dilation (Table 1) confirmed a significant main effect of language, in that pupil dilation in millimeter was larger in L1 ($M = .46$, $SD = .24$) than in L2 ($M = .38$, $SD = .23$) and a significant main effect of emotionality with stronger dilation for high-emotionality slogans ($M = .51$, $SD = .23$), compared to low ones ($M = .33$, $SD = .21$). Importantly, as depicted in Figure 2, first panel, a significant interaction between language and emotionality qualified the main effects, such that the language difference was significant for high, $b = -.124$, $SE = .024$, $t = -5.16$, but not for low-emotionality slogans, $b = -.007$, $SE = .028$, $t = -.24$. Consistent with the physiology of pupil reactions, dilation increased over time, and women and men reacted similarly.

Grip force

Grip force responses across the 32 slogans were highly consistent with Cronbach's $\alpha = .99$ in both languages. The LMM (Table 1) found a main effect of language with stronger maximum force in L1 ($M = 88.44$, $SD = 51.45$) than in L2, $M = 78.69$, $SD = 44.72$, and a main effect of emotionality with substantially stronger force for high ($M = 94.60$, $SD = 50.49$) than low ($M = 72.53$, $SD = 43.61$) intensity slogans. Notably, there was a significant language-emotionality interaction. As the second panel in Figure 2 suggests – in parallel to the pupil reactions – the gap between L1 and L2 was larger for high-emotionality slogans, $b = -.052$, $SE = .051$, $t = -1.02$, compared to low ones, $b = .015$, $SE = .058$, $t = .26$. Finally, men's grip force ($M = 100.34$, $SD = 54.51$) was significantly stronger compared to women's, $M = 75.74$, $SD = 43.17$.

Grip duration

The measure of grip duration was internally consistent across slogans with Cronbach's $\alpha = .96$ in L1 and .98 in L2. In the LMM (Table 1), language did not account for a significant share of variance, L1: $M = 1.55$, $SD = 1.11$, L2: $M = 1.91$, $SD = 1.65$, and it did not interact with emotionality (see Figure 2, third panel). There was a significant main effect of emotionality indicating that participants squeezed the dynamometer longer after high- ($M = 1.95$, $SD = 1.58$) relative to low-emotionality ($M = 1.51$, $SD = 1.19$) slogans. The effect of sex approached significance.

Verbal rating

The verbal ratings of how (un-)emotional participants felt after reading each slogan were very consistent with Cronbach's $\alpha = .96$ in L1 and .93 in L2. The Poisson model (Table 1) yielded no significant main or interaction effect of language (see Figure 2, fourth panel) since the ratings were similar in L1 ($M = 3.71$, $SD = 1.55$) and L2 ($M = 3.95$, $SD = 1.48$). High-emotionality slogans received significantly higher ratings ($M = 4.55$, $SD = 1.36$) than low ones, $M = 3.12$, $SD = 1.32$, while none of the other predictors was significant.

Comparison of measures

Grip force explained pupil dilation beyond the null model, $\chi^2(1) = 26.94$, $p < .001$, such that larger pupil dilation was associated with stronger grip force, $b = .018$, $SE = .003$, $t = 5.20$. Grip duration also improved model fit, $\chi^2(1) = 17.07$, $p < .001$. However, longer grip duration was associated with weaker pupil dilation, $b = -.012$, $SE = .003$, $t = -4.14$. Finally, the inclusion of verbal ratings did not improve model fit, $\chi^2(1) = .49$, $p = .784$; $b = -.002$, $SE = .003$, $t = -.67$.

As the negative relationship between pupil dilation and grip duration was unexpected, and participants repeatedly asked the experimenters about the meaning of "squeezing longer", we conducted additional explorative analyses. Grip duration correlated significantly with the mean time participants dwelled on the advertising slogans, $r(66) = .340$, $p = .005$, while none of the other three emotion measures showed such a correlation, all $r < |.15|$, $p > .25$. Participants also tended to dwell slightly longer on L2 slogans, $M = 4.29$, $SD = .59$ vs. $M = 3.99$, $SD = 0.63$; $t(65) = 2.00$, $p = .05$. Accordingly, an extended LMM found that dwell time significantly predicted grip duration ($b = .067$, $SE = .025$,

Table 1. Mixed-effects models predicting reactions in the sentence-reading task.

Dependent variable	Pupil diameter change [mm]			Grip force (log) [Newton]			Grip duration (log) [s]			Verbal emotion rating [1-7]		
Fixed effects	*b*	*SE*	*t*	*b*	*SE*	*t*	*b*	*SE*	*t*	*b*	*SE*	*z*
(Intercept)	0.368	0.018	20.93	1.872	0.029	64.46	0.156	0.030	5.24	1.276	0.037	34.92
Language (Lg.)	−0.067	0.025	**-2.71**	−0.014	0.053	−0.27	0.059	0.052	1.12	0.043	0.051	0.83
Emotionality	0.193	0.020	**9.48**	0.132	0.011	**11.60**	0.113	0.013	**9.01**	0.383	0.033	**11.44**
Epoch	0.033	0.002	**18.90**	-	-	-	-	-	-	-	-	-
Sex	0.012	0.026	0.46	0.139	0.055	**2.53**	0.092	0.056	1.64	−0.153	0.055	**−2.80**
Lg. × emotion.	−0.115	0.015	**−7.63**	−0.073	0.023	**−3.23**	0.016	0.025	0.62	−0.047	0.067	−0.71

Note. Absolute *t*-values > 1.96 are significant (in bold). Continuous predictors were standardized. (-) = variable not relevant for the model.

$t = 2.66$) and interacted with language, $b = .104$, $SE = .050$, $t = 2.09$. In LMMs predicting pupil dilation, grip force, and verbal ratings, dwell time was irrelevant. We therefore assume that participants reported a subjective mix of the intensity of their emotional reaction and the perceived time they spent reading a slogan in the duration of squeezing the dynamometer.

Interim discussion

In sum, the sentence rating experiment successfully replicated the pupillometry results from Iacozza et al. (2017) with weaker L2 emotions for high emotionality stimuli, yet no language differences in verbal ratings. In addition to the language-emotionality interaction, our sample and materials also showed a main effect of language on pupil dilation, so that our first and second hypothesis were confirmed. As the visual comparison of the four interaction plots in Figure 2 and the result pattern in the LMMs in Table 1 suggest, grip force mirrored the main and interaction effects observed for pupil dilation. Model comparisons showed that grip force explained pupil reactions well. However, the verbal ratings did not correspond with pupil dilation. These findings supported our third hypothesis. The result pattern for grip duration was more similar to verbal ratings than to pupil dilation, and participants may also have confused reading time and emotionality.

4. Study 2: Narrative video-induced emotions

Study 2 replicated parts of the German–English pupillometry experiment by Thoma and Baum (2019) inducing emotions in bilinguals with authentic narrative video commercials. We extended the task by showing participants several videos of low and high emotionality (instead of a neutral baseline and one high emotionality video) and by a combination of visceral and verbal self-report ratings of emotion as in Study 1.

4.1. Method

Participants and materials

Participants were the same 66 as in Study 1. The present experiment included two authentic low- and two high-emotionality video commercials for which a German and English language version was available or created. We compiled a larger pool of videos which convey their message primarily through audio narration and do not feature talking actors. Their potential emotionality was discussed in a focus group consisting of two German and two English L1 speakers. We selected two low-high emotionality pairs that were equivocally classified and of comparable length. If necessary, we edited the videos in length and removed sequences showing written language other than the brand name. We resampled the audio tracks of the high-emotionality videos, such that the same balanced bilingual speaker, who had a sonorant mature male or a female child voice, narrated both language versions. In the low-emotionality videos, the narrator voices were comparable. The visual tracks of the videos were identical across language conditions. We programed a Python software to estimate the mean perceived relative luminance per video second according to the luminous efficiency function (Poynton, 2012, Eq. 24.1) from red, green and blue tristimulus values of the individual images within this second.

In detail, we sampled the high-emotionality 58-s video *Hochzeitstag* ("Wedding Day", Minghella, 2006) used and resampled by Thoma and Baum (2019). It features an elderly

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

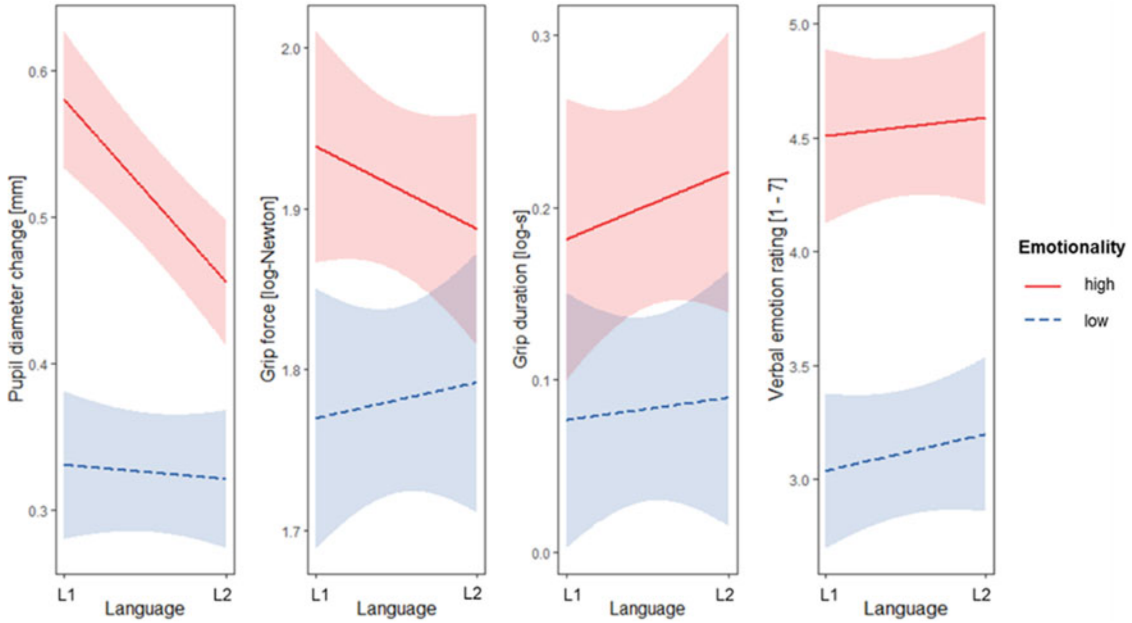

Figure 2. Covariate adjusted interaction plots for language × emotionality with means and 95%-CIs for the four measures of emotions induced by the sentence-reading task.

lady on a cemetery who receives flowers from a delivery service her late husband had pre-ordered. Second, we edited a 42-s low emotionality video ("Letterbox Flowers", Bloom & Wild, 2019), showing a variety of flowers and recommending their delivery. Third, we used a 170-s high-emotionality video ("My father is a liar", MetLife, 2015) telling the story of an unemployed single father doing everything for his child daughter. Fourth, we selected a 197-s low-emotionality corporate video (Endress+Hauser, 2019) describing the spirit of a Swiss engineering company.

The videos were further normed for valence and arousal in two unrelated studies, where they served as fillers (see Appendix B for values). Thirty-four L1 English and 30 L1 German speakers rated the two low or the two high emotionality videos. Their ratings confirmed that high emotionality stimuli were more negative ($F(1, 63) = 54.75$, $p < .001$) and more arousing ($F(1, 63) = 313.44$, $p < .001$), while there was no significant language difference or interaction, all $F < 1.85$, $p > .17$.

Procedure and analyses

The procedure was as in Study 1, with the advertising slogans in the sentence rating task being replaced with the four video stimuli presented in random order in both language conditions. After a video had played, the task proceeded automatically to the screen presenting the visceral rating instruction (compare Figure 1).

Data preparation and analyses also followed the rationales of Study 1. Since the video stimuli differed in emotionality and luminance, the LMMs predicting pupil dilation additionally included relative luminance per epoch to control for the latter.

4.2. Results and discussion

Pupil dilation

As summarized in Table 2, the LMM revealed no main effect of language with only slightly stronger pupil dilation in L1, $M = .32$, $SD = .48$, compared to L2, $M = .27$, $SD = .48$. The model confirmed a very strong main effect of emotionality with hardly any reaction to low emotionality videos, $M = .04$, $SD = .41$, yet considerable pupil dilation during high-emotionality videos, $M = .57$, $SD = .39$. This main effect was qualified by a significant interaction with language, such that the difference between low vs. high emotionality was significantly larger in L1 ($b = .446$, $SE = .097$, $t = 4.59$) than in L2, $b = .257$, $SE = .102$, $t = 2.53$ (see Figure 3, first panel). A different way to interpret the two-way interaction is to say that, regarding language-dependent emotions, pupil dilation was weaker in L2 during listening to high-emotionality video narratives, $b = -.159$, $SE = .069$, $t = -2.31$, while language made no significant difference during low-emotionality videos, $b = .032$, $SE = .059$, $t = .54$. As expected, the model revealed a significant effect of epoch with increasing pupil dilation over time and a massive constriction of pupil diameter in response to higher luminance. Finally, women reacted slightly more intensely ($M = .31$, $SD = 0.50$) than men, $M = .26$, $SD = .44$.

Grip force

Even though the task included only four video stimuli, the reliability of grip force was good with Cronbach's $\alpha = .90$ in L1 and .87 in L2. Table 2 reports no significant main effect of language effect despite descriptively stronger force in L1 ($M = 110.07$, $SD = 67.95$) than in L2, $M = 96.21$, $SD = 54.99$. The LMM confirmed a main effect of emotionality with clearly stronger grip force in response to high ($M = 130.42$, $SD = 61.18$) than low ($M = 75.86$, $SD = 49.96$) emotionality videos and also, a significant language-emotionality interaction. Inspection of this interaction (see Figure 3, second panel) revealed that the difference between low and high emotionality was more pronounced in L1 ($b = .338$, $SE = .033$, $t = 10.35$) than in L2, $b = .205$, $SE = .036$, $t = 5.76$. As physiognomically expected, men squeezed the dynamometer significantly harder ($M = 132.46$, $SD = 75.69$) than women, $M = 89.46$, $SD = 49.14$.

Grip duration

Participants squeezed the dynamometer consistently long across the video stimuli with Cronbach's $\alpha = .70$ in L1 and .86 in L2. Grip duration did not differ significantly between the language

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

Table 2. Mixed-effects models predicting reactions in the video task.

Dependent variable	Pupil diameter change [mm]			Grip force (log) [Newton]			Grip duration (log) [s]			Verbal emotion rating [1-7]		
Fixed effects	*b*	*SE*	*t*	*b*	*SE*	*t*	*b*	*SE*	*t*	*b*	*SE*	*z*
(Intercept)	0.327	0.054	6.05	1.956	0.026	74.55	0.261	0.026	10.02	1.452	0.033	43.91
Language (Lg.)	−0.063	0.051	−1.23	−0.006	0.050	−0.13	0.065	0.050	1.32	0.036	0.063	0.58
Emotionality	0.353	0.103	**3.42**	0.272	0.020	**13.57**	0.330	0.022	**15.02**	0.734	0.061	**11.99**
Epoch	0.012	0.002	**6.66**	-	-	-	-	-	-	-	-	-
Luminance	−0.142	0.002	**−89.46**	-	-	-	-	-	-	-	-	-
Sex	−0.013	0.005	**−2.41**	0.158	0.051	**3.06**	0.083	0.053	1.56	−0.068	0.064	−1.07
Lg. × emotion.	−0.191	0.079	**-2.41**	−0.133	0.040	**-3.33**	−0.041	0.043	−0.95	−0.081	0.122	−0.66

Note. Absolute *t*-values > 1.96 are significant (in bold). Continuous predictors were standardized. (-) = variable not relevant for

conditions (Table 2). Descriptively, L1 participants squeezed shorter ($M = 2.01$, $SD = 1.30$) than the L2 group, $M = 2.48$, $SD = 2.05$. The effect of emotionality was significant, such that participants squeezed the dynamometer longer after high-emotionality ($M = 2.95$, $SD = 1.74$) than low-emotionality ($M = 1.54$, $SD = 1.42$) videos. Language and emotionality did not interact (see Figure 3, third panel). The effect of sex was negligible.

Verbal rating

The reliability of the verbal emotion ratings was acceptable with Cronbach's $\alpha = .77$ in L1 and .71 in L2. The Poisson model (Table 2) found no significant main or interaction effect of language (see Figure 3, fourth panel) as the ratings were similar in L1 ($M = 4.54$, $SD = 2.01$) and L2, $M = 4.70$, $SD = 1.89$. High-emotionality videos were rated significantly more emotional ($M = 4.72$, $SD = 1.94$) than low-emotionality ones, $M = 4.39$, $SD = 1.95$. Moreover, men reported weaker feelings ($M = 4.39$, $SD = 1.95$) than women, $M = 4.72$, $SD = 1.94$.

Comparison of measures

In parallel to Study 1, we first computed a null model including epoch, relative luminance, and sex as fixed predictors of pupil dilation and, then, used likelihood ratio tests to analyze the contribution of visceral and verbal ratings. Grip force improved model fit significantly beyond the null model, $\chi^2(1) = 24.43$, $p < .001$, based on a positive association between pupil dilation and force, $b = .024$, $SE = .005$, $t = 5.08$. In a separate model, grip duration also improved model fit, $\chi^2(1) = 6.61$, $p < .001$, yet longer grip duration was associated negatively with pupil dilation, $b = -.012$, $SE = .004$, $t = -2.82$. Finally, a model with verbal ratings was not significantly better than the null model, $\chi^2(1) = 1.03$, $p = .311$; $b = .006$, $SE = .006$, $t = 1.01$.

Interim discussion

Study 2 replicated the reduced pupil dilation effect in L2 relative to L1 in response to listening to authentic narratives in video commercials for high-emotionality videos originally observed by Thoma and Baum (2019). Consistent with prior research (García-Palacios et al., 2018; Iacozza et al., 2017; Toivo & Scheepers, 2019), pupil dilation clearly responded to the emotional intensity of the stimuli, whereas language had no significant direct influence but qualified the emotionality effect, such that it was more pronounced in L1 than L2. Therefore, Study 2 confirmed the replication hypotheses H1 and H2. With respect to our research hypothesis, H3, visceral ratings via grip force indeed mirrored the result pattern observed for pupil dilation more closely and predicted pupil dilation in an LMM more accurately than verbal ratings. As in Study 1, there was an unexpected negative relationship between pupil dilation and grip duration.

5. Discussion

The present study aimed at validating visceral measures obtained from squeezing a handheld dynamometer for the assessment of bilingual feelings. Therefore, we replicated and extended two pupillometry experiments (Iacozza et al., 2017; Thoma & Baum, 2019) investigating bilinguals' language-dependent reactions induced by reading sentences (Study 1) and listening to narrative video commercials (Study 2) of low and high emotionality in L1 or L2. In addition to pupil dilation as an index of automatic emotional reactions of the sympathetic nervous system, we obtained explicit visceral and verbal emotion ratings for each stimulus.

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

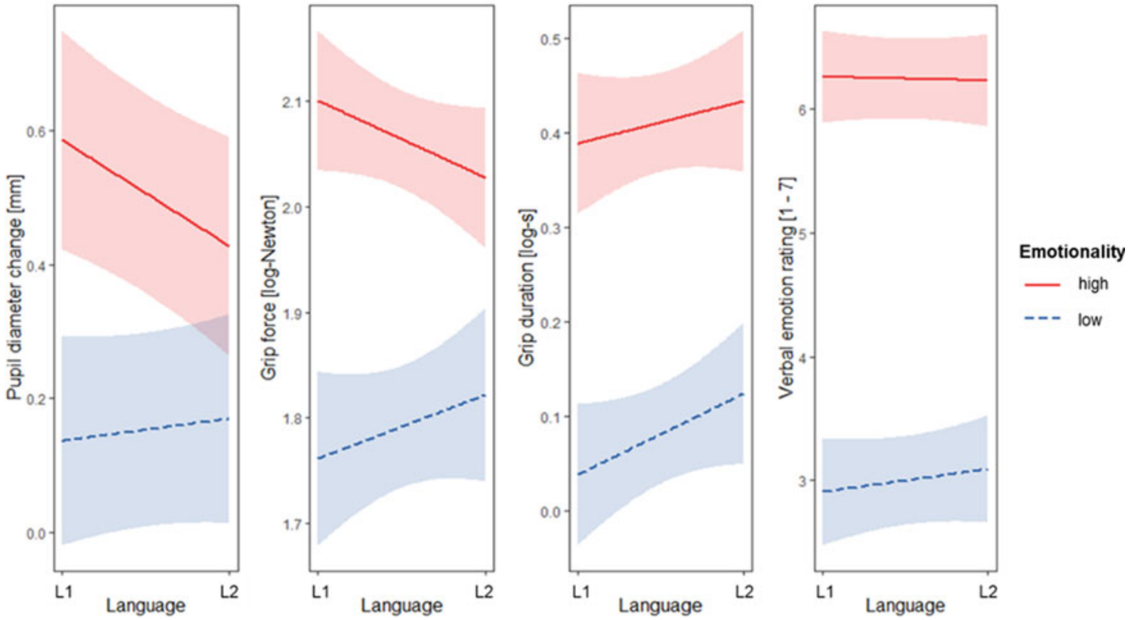

Figure 3. Covariate-adjusted interaction plots for language × emotionality with means and 95%-CIs for the four measures of emotion in the video task.

The most important replication result in both studies was that – based on a language-emotionality interaction in pupil dilation – bilinguals were more sensitive to stimulus emotionality in their L1 than L2. In other words, the difference in emotional reaction between neutral low-arousal vs. negative high-arousal sentences and narratives was larger in L1. This interaction has been shown with pupillometry for conjoint valence and arousal-induced emotional intensity (García-Palacios et al., 2018; Iacozza et al., 2017; Thoma & Baum, 2019) and arousal (Toivo & Scheepers, 2019), as well as for valence in studies measuring SCR (Caldwell-Harris & Ayçiçeği-Dinn, 2009) and ERP (Jończyk et al., 2016; Wu & Thierry, 2012). Also consistent with some prior research (Harris et al., 2003; Iacozza et al., 2017; Jończyk et al., 2016), verbal ratings differed with stimulus emotionality in the sense of a manipulation check, but they did not show a language or interaction effect and were uncorrelated with pupil dilation. Maximum grip force, however, predicted pupil dilation well and, most importantly, reproduced the language-emotionality interaction observed in the pupillometry data.

The present research has important methodological and theoretical implications but also limitations that may be addressed in future research. At a practical level, we suggest that for emotion research, dynamometer responses provide an economical, easy-to-use nonverbal measurement alternative to pupillometry and electrodermal activity in terms of equipment and data complexity. The methodology further allows researchers to measure attentional – rather than automatic and involuntary – responses without having to choose a language for the reporting. Visceral ratings could possibly be used in domains of bilingualism research beyond emotion, where the dependent variable is at risk of being unknowingly influenced by the choice of the response language such as attitudes, preferences, and even the acceptability of grammatical structures. Notwithstanding, our evidence is limited because we did not include other than verbal symbolic rating scales, e.g., with iconic anchors or color points instead of numbers (Langhe et al., 2011) in the validation. To maximize the power to confirm our replication hypotheses (H1 and H2) that were the prerequisite to test our research hypothesis (H3), we limited stimulus emotionality to neutral-low vs. negative-high arousal items. In line with extant research (e.g., Harris et al., 2003; Wu & Thierry, 2012), we expect similar but weaker emotionality and language effects for positively valenced sentences and narratives. However, how these will be reflected in pupillary, dynamometer, and verbal responses remains to be tested.

In terms of theory development, it is central to further explore in how far the present findings represent measurement or conceptual differences. On the one hand, we could take up Creswell et al.'s (2018, 2019) verbal-overshadowing argument that visceral states or "gut feelings" such as hunger and urge to smoke are distorted by verbal ratings and can, thus, be assessed more validly with visceral dynamometer responses. If researchers are interested in pre-verbal or even pre-cognitive affectual states, then measuring them via verbal responses may indeed bias what they intend to measure, so that visceral responses would allow for more valid conclusions about basic affectual reactions. On the other hand, we could argue that the discrepancies between pupil responses and verbal ratings occurred not so much because translating stimulus-evoked emotionality into a numerical scale with verbal anchors (in our case same-orthography German and English cognate words) was a difficult task. Instead, the verbal rating may have triggered additional thinking and, in turn, more deliberate decision-making (Dijksterhuis, 2004) and stronger emotion regulation (Torre & Lieberman, 2018) before participants reported the intensity of their current feelings. Deliberate decisions can but need not correspond with intuition, in particular, if the intuition is weak (Dhar & Gorlin, 2013). The intuitive pupillary reaction to linguistic emotional stimuli could therefore reflect a different stage of emotion processing, compared to the deliberate verbal rating behavior. This interpretation aligns with bilingual emotion research claiming that bilinguals' weaker emotions in L2 are highly automatic reactions that depend on the time course of emotion processing and may vanish in later, more intense and deliberate stages of emotion processing (Jończyk et al., 2016;

Opitz & Degner, 2012; Thoma, 2021). On this note, we could tentatively order the four emotion measures used in Studies 1 and 2 on a continuum from intuitive to deliberate starting with pupil dilation followed by grip force, grip duration and verbal ratings (see Figures 2 and 3). Whereas maximum grip force represented an impulsive reflex close to the automatic pupil responses, grip duration responses were naturally slower, more controlled and paralleled those of deliberate verbal responses. Based on Creswell et al. (2018, 2019), we did not expect deviations between grip force and duration, but there was also evidence that participants associated grip duration with stimulus duration, which cannot happen when reporting on hungriness or an urge. We could not fully resolve the issue with grip duration at the instructional level. However, to test the proposed continuum, it would be interesting to compare all four measures when bilinguals experience valence and arousal (core affect) in contrast to specific emotions – in particular, as languages may be actively and differentially involved in constructing specific emotions (Barrett, 2006; Gendron et al., 2012).

Taken together, the present findings warrant two different conclusions that await further testing. First, we could argue that we compared explicit visceral and verbal ratings of language-dependent emotional experiences by bilinguals and validated them against implicit prior emotional reactions reflected in pupil dilation. Then, our findings provide initial evidence for the superior validity of grip force as an explicit yet nonverbal, visceral self-report measure of bilingual feelings. Second, and maybe even more interestingly, the findings could suggest that grip force captures bilinguals' feelings at an early, relatively intuitive stage of emotion processing and reporting. This stage still closely resembles the language-dependent differences of automatic physiological emotional reactions that may change gradually with increasing deliberation, i.e., verbalization, and the choice of the language of reporting.

Data availability statement. Non-copyrighted experimental stimuli, most data files and the key R core are available at: https://osf.io/zhv4n/

Supplementary Material. For supplementary material accompanying this paper, visit https://doi.org/10.1017/S1366728922000396

References

Barbieri, F, Kruszewski, G, Ronzano, F and Saggion, H (2016) How cosmopolitan are emojis? In A Hanjalic, C Snoek, M Worring, D Bulterman, B Huet, A Kelliher, Y Kompatsiaris and J Li (Eds), *Mm'16: Proceedings of the 2016 ACM Multimedia Conference: October 15-19, 2016, Amsterdam, The Netherlands*, ACM. https://doi.org/10.1145/2964284.2967278 pp. 531–535.

Barr, DJ, Levy, R, Scheepers, C and Tily, HJ (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrett, LF (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review* **10**, 20–46. https://doi.org/10.1207/s15327957pspr1001_2

Bates, D, Maechler, M, Bolker, B, Walker, S, Bojesen Christensen, RH, Singmann, H, Dai, B, Scheipl, F, Grothendieck, G, Green, P and Fox, J (2021). *Package Lme4: Linear mixed-effects models using Eigen and S4* (Version 1.1-27.1) [Computer software]. https://cran.r-project.org/package=lme4

Beatty, J (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* **91**, 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Benet-Martínez, V, Leu, J, Lee, F and Morris, MW (2002). Negotiating biculturalism. *Journal of Cross-Cultural Psychology* **33**, 492–516. https://doi.org/10.1177/0022022102033005005

Bloom & Wild (Producer) (2019). *Letterbox flowers*. Bloom, & Wild Limited, available at: https://www.youtube.com/watch?v=−ZnCu99Nij8.

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

Bradley, MM and Lang, PJ (1999). *Affective Norms for English Words (ANEW): Instruction manual and affective ratings*. University of Florida.

Bradley, MM, Miccoli, L, Escrig, MA and Lang, PJ (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* **45**, 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Caldwell-Harris, CL and Ayçiçeği-Dinn, A (2009). Emotion and lying in a non-native language. *International Journal of Psychophysiology* **71**, 193–204. https://doi.org/10.1016/j.ijpsycho.2008.09.006

Citron, FMM, Gray, MA, Critchley, HD, Weekes, BS and Ferstl, EC (2014). Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia* **56**, 79–89. https://doi.org/10.1016/j.neuropsychologia.2014.01.002

Collins, BE, Ellsworth, PC and Helmreich, RL (1967). Correlations between pupil size and the semantic differential: An experimental paradigm and pilot study. *Psychonomic Science* **9**, 627–628. https://doi.org/10.3758/BF03327922

Corrêa, TGC, Donato, SVS, Lima, KCA, Pereira, RV, Uygur, M and Freitas, PBde (2020). Age- and sex-related differences in the maximum muscle performance and rate of force development scaling factor of precision grip muscles. *Motor Control* **24**, 1–17. https://doi.org/10.1123/mc.2019-0021

Costa, A, Foucart, A, Hayakawa, S, Aparici, M, Apesteguia, J, Heafner, J and Keysar, B (2014). Your moral depends on language. *PLoS ONE* **9**, 1–7. https://doi.org/10.1371/journal.pone.0094842

Creswell, KG, Sayette, MA, Schooler, JW, Wright, AGC and Pacilio, LE (2018). Visceral states call for visceral measures: Verbal overshadowing of hunger ratings across assessment modalities. *Assessment* **25**, 173–182. https://doi.org/10.1177/1073191116645910

Creswell, KG, Sayette, MA, Skrzynski, CJ, Wright, AGC, Schooler, JW and Sehic, E (2019). Assessing cigarette craving with a squeeze. *Clinical Psychological Science* **7**, 597–611. https://doi.org/10.1177/2167702618815464

Damasio, A and Carvalho, GB (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience* **14**, 143–152. https://doi.org/10.1038/nrn3403

Defrasne Ait-Said, E, Maquestiaux, F and Didierjean, A (2014). Verbal overshadowing of memories for fencing movements is mediated by expertise. *PLoS ONE* **9**, 1–5. https://doi.org/10.1371/journal.pone.0089276

Dewaele, JM (2004). The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of Multilingual and Multicultural Development* **25**, 204–222. https://doi.org/10.1080/01434630408666529

Dewaele, JM (2008). The emotional weight of I love you in multilinguals' languages. *Journal of Pragmatics* **40**, 1753–1780. https://doi.org/10.1016/j.pragma.2008.03.002

Dewaele, JM (2013). *Emotions in multiple languages* (Paperback ed.). Basingstoke: Palgrave Macmillan.

Dhar, R and Gorlin, M (2013). A dual-system framework to understand preference construction processes in choice. *Journal of Consumer Psychology* **23**, 528–542. https://doi.org/10.1016/j.jcps.2013.02.002

Dijksterhuis, A (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology* **87**, 586–598. https://doi.org/10.1037/0022-3514.87.5.586

Dylman, AS and Bjärtå, A (2019). When your heart is in your mouth: The effect of second language use on negative emotions. *Cognition and Emotion* **33**, 1284–1290. https://doi.org/10.1080/02699931.2018.1540403

Ekman, P (1992). Are there basic emotions? *Psychological Review* **99**, 550–553. https://doi.org/10.1037/0033-295x.99.3.550

Endress+Hauser (Producer). (2019). *Corporate video*. Endress+Hauser AG, available at https://www.youtube.com/watch?v=yJa7D91vu6A.

García-Palacios, A, Costa, A, Castilla, D, Del Río, E, Casaponsa, A and Duñabeitia, JA (2018). The effect of foreign language in fear acquisition. *Scientific Reports* **8**, 1157. https://doi.org/10.1038/s41598-018-19352-8

Gendron, M, Lindquist, KA, Barsalou, L and Barrett, LF (2012). Emotion words shape emotion percepts. *Emotion* **12**, 314–325. https://doi.org/10.1037/a0026007

Harris, C, Ayçiçeği-Dinn, A and Gleason, JB (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics* **24**, 561–579. https://doi.org/10.1017/S0142716403000286

Iacozza, S, Costa, A and Duñabeitia, JA (2017). What do your eyes reveal about your foreign language? Reading emotional sentences in a native and foreign language. *PLoS ONE* **12**, e0186027. https://doi.org/10.1371/journal.pone.0186027

Imbault, C, Titone, D, Warriner, AB and Kuperman, V (2021). How are words felt in a second language: Norms for 2,628 English words for valence and arousal by L2 speakers. *Bilingualism: Language and Cognition* **24**, 281–292. https://doi.org/10.1017/S1366728920000474

Jackson, JC, Watts, J, Henry, TR, List, JM, Forkel, R, Mucha, PJ, Greenhill, SJ, Gray, RD and Lindquist, KA (2019). Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522. https://doi.org/10.1126/science.aaw8160

Janisse, MP (1973). Pupil size and affect: A critical review of the literature since 1960. *Canadian Psychologist/Psychologie Canadienne* **14**, 311–329. https://doi.org/10.1037/h0082230

Jankowiak, K and Korpal, P (2018). On modality effects in bilingual emotional language processing: Evidence from galvanic skin response. *Journal of Psycholinguistic Research* **47**, 663–677. https://doi.org/10.1007/s10936-017-9552-5

Jończyk, R, Boutonnet, B, Musiał, K, Hoemann, K and Thierry, G (2016). The bilingual brain turns a blind eye to negative statements in the second language. *Cognitive, Affective, & Behavioral Neuroscience* **16**, 527–540. https://doi.org/10.3758/s13415-016-0411-x

Kassam, KS and Mendes, WB (2013). The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. *PLoS ONE* **8**, e64959. https://doi.org/10.1371/journal.pone.0064959

Keysar, B, Hayakawa, SL and An, SG (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science* **23**, 661–668. https://doi.org/10.1177/0956797611432178

Kinner, VL, Kuchinke, L, Dierolf, AM, Merz, CJ, Otto, T and Wolf, OT (2017). What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes. *Psychophysiology* **54**, 508–518. https://doi.org/10.1111/psyp.12816

Kurdi, B, Lozano, S and Banaji, MR (2017). Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods* **49**, 457–470. https://doi.org/10.3758/s13428-016-0715-3

Kuznetsova, A, Brockhoff, PB and Christensen, RHB (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* **82**. https://doi.org/10.18637/jss.v082.i13

Langhe, Bde, Puntoni, S, Fernandes, D and van Osselaer, SMJ (2011). The anchor contraction effect in international marketing research. *Journal of Marketing Research* (*JMR*) **48**, 366–380. https://doi.org/10.1509/jmkr.48.2.366

Lemercier, A, Guillot, G, Courcoux, P, Garrel, C, Baccino, T and Schlich, P (2014). Pupillometry of taste: Methodological guide – from acquisition to data processing - and toolbox for MATLAB. *Tutorials in Quantitative Methods for Psychology* **10**, 179–195. https://www.tqmp.org/RegularArticles/vol10-2/p179/p179.pdf

Lieberman, MD, Eisenberger, NI, Crockett, MJ, Tom, SM, Pfeifer, JH and Way, BM (2007). Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science* **18**, 421–428. https://doi.org/10.1111/j.1467-9280.2007.01916.x

Lindquist, KA, Satpute, AB and Gendron, M (2015). Does language do more than communicate emotion? *Current Directions in Psychological Science* **24**, 99–108. https://doi.org/10.1177/0963721414553440

Majid, A (2012). Current emotion research in the language sciences. *Emotion Review* **4**, 432–443. https://doi.org/10.1177/1754073912445827

Marcos, LR (1976). Bilinguals in psychotherapy: Language as an emotional barrier. *American Journal of Psychotherapy* **30**, 552–560. https://doi.org/10.1176/appi.psychotherapy.1976.30.4.552

Marian, V and Neisser, U (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General* **129**, 361–368. https://doi.org/10.1037/0096-3445.129.3.361

Mathôt, S, Grainger, J and Strijkers, K (2017). Pupillary responses to words that convey a sense of brightness or darkness. *Psychological Science* **28**, 1116–1124. https://doi.org/10.1177/0956797617702699

Mauss, IB and Robinson, MD (2009). Measures of emotion: A review. *Cognition and Emotion* **23**, 209–237. https://doi.org/10.1080/02699930802204677

McDougal, DH and Gamlin, PD (2015). Autonomic control of the eye. *Comprehensive Physiology* **5**, 439–473. https://doi.org/10.1002/cphy.c140014.

MetLife (Producer). (2015). *My father is a liar*. Metropolitan Life Insurance Company, available at https://www.youtube.com/watch?v=ura3lJgei5g.

Minghella, A (Director). (2006). *Hochzeitstag*. Fleurop-Interflora AG (Schweiz), available at https://www.fleurop.ch/de/s/werbung.

Niedenthal, PM (2007). Embodying emotion. *Science* (*New York, N.Y.*) **316**, 1002–1005. https://doi.org/10.1126/science.1136930

Noriega, J and Blair, E (2008). Advertising to bilinguals: Does the language of advertising influence the nature of thoughts? *Journal of Marketing* **75**, 69–83. https://doi.org/10.1509/jmkg.72.5.069

Oganian, Y, Korn, CW and Heekeren, HR (2016). Language switching-but not foreign language use per se-reduces the framing effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition* **42**, 140–148. https://doi.org/10.1037/xlm0000161

Opitz, B and Degner, J (2012). Emotionality in a second language: It's a matter of time. *Neuropsychologia* **50**, 1961–1967. https://doi.org/10.1016/j.neuropsychologia.2012.04.021

Pinheiro, JC and Bates, DM (2009). *Mixed-effects models in S and S-PLUS* (reprinted paperback ed. of the 2000 ed.). *Statistics and computing*. New York: Springer.

Poynton, C (2012). *Digital video and HD algorithms and interfaces* (2nd ed.). *The Morgan Kaufmann Series in Computer Graphics*. Burlington: Elsevier Science.

Ross, M, Xun, WQE and Wilson, AE (2002). Language and the bicultural self. *Personality and Social Psychology Bulletin* **28**, 1040–1050. https://doi.org/10.1177/01461672022811003

RStudio Team. (2021). *RStudio: Integrated Development Environment for R* (Version 2021.9.1.372) [Computer software]. RStudio, PBC. Boston, Mass. http://www.rstudio.com/

Russell, JA (2003). Core affect and the psychological construction of emotion. *Psychological Review* **110**, 145–172. https://doi.org/10.1037/0033-295x.110.1.145

Russell, JA and Barrett, LF (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology* **76**, 805–819. https://doi.org/10.1037/0022-3514.76.5.805

Schmidtke, DS, Schröder, T, Jacobs, AM and Conrad, M (2014). Angst: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods* **46**, 1108–1118. https://doi.org/10.3758/s13428-013-0426-y

Schmidtke, J (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology* **5**, 1–16. https://doi.org/10.3389/fpsyg.2014.00137

Schooler, JW and Engstler-Schooler, TY (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology* **22**, 36–71. https://doi.org/10.1016/0010-0285(90)90003-M

Schwartz, SJ, Benet-Martínez, V, Knight, GP, Unger, JB, Zamboanga, BL, Des Rosiers, SE, Stephens, DP, Huang, S and Szapocznik, J (2014). Effects of language of assessment on the measurement of acculturation: Measurement equivalence and cultural frame switching. *Psychological Assessment* **26**, 100–114. https://doi.org/10.1037/a0034717

Sheikh, NA and Titone, D (2016). The embodiment of emotional words in a second language: An eye-movement study. *Cognition and Emotion* **30**, 488–500. https://doi.org/10.1080/02699931.2015.1018144

Sianipar, A, Middelburg, R and Dijkstra, T (2015). When feelings arise with meanings: How emotion and meaning of a native language affect second language processing in adult learners. *PLoS ONE* **10**, 1–33. https://doi.org/10.1371/journal.pone.0144576

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press

Sutton, TM, Altarriba, J, Gianico, JL and Basnight-Brown, DM (2007). The automatic access of emotion: Emotional Stroop effects in Spanish–English bilingual speakers. *Cognition and Emotion* **21**, 1077–1090. https://doi.org/10.1080/02699930601054133

Thoma, D (2021). Emotion regulation by attentional deployment moderates bilinguals' language-dependent emotion differences. *Cognition and Emotion* **35**, 1121–1135. https://doi.org/10.1080/02699931.2021.1929853

Thoma, D and Baum, A (2019). Reduced language processing automaticity induces weaker emotions in bilinguals regardless of learning context. *Emotion* **19**, 1023–1034. https://doi.org/10.1037/emo0000502

Toivo, W and Scheepers, C (2019). Pupillary responses to affective words in bilinguals' first versus second language. *PLoS ONE* **14**, e0210450. https://doi.org/10.1371/journal.pone.0210450

Torre, JB and Lieberman, MD (2018). Putting feelings Into words: Affect labeling as implicit emotion regulation. *Emotion Review* **10**, 116–124. https://doi.org/10.1177/1754073917742706

Vélez-Uribe, I and Rosselli, M (2019). The auditory and visual appraisal of emotion-related words in Spanish–English bilinguals. *Bilingualism: Language and Cognition* **22**, 30–46. https://doi.org/10.1017/S1366728917000517

Weijters, B, Puntoni, S and Baumgartner, H (2017). Methodological issues in cross-linguistic and multilingual advertising research. *Journal of Advertising* **46**, 115–128. https://doi.org/10.1080/00913367.2016.1180656

Williams, A, Srinivasan, M, Liu, C, Lee, P and Zhou, Q (2020). Why do bilinguals code-switch when emotional? Insights from immigrant parent-child interactions. *Emotion* **20**, 830–841. https://doi.org/10.1037/emo0000568

Wilson, TD and Schooler, JW (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality, & Social Psychology* **60**, 181–192. https://doi.org/10.1037//0022-3514.60.2.181

Wu, YJ and Thierry, G (2012). How reading in a second language protects your heart. *The Journal of Neuroscience* **32**, 6485–6489. https://doi.org/10.1523/JNEUROSCI.6119-11.2012

https://doi.org/10.1017/S1366728922000396 Published online by Cambridge University Press