sciendo

Psicológica (2020), 41, 21-38
doi: 10.2478/psicolj-2020-0002

Interpreting Foreign Smiles: Language Context and Type of Scale in the Assessment of Perceived Happiness and Sadness

Candice Frances[*1], Silvia Pueyo[2], Vanessa Anaya[2] & Jon Andoni Duñabeitia[3,4]

[1]*BCBL, Basque Center on Cognition, Brain and Language, Donostia, Spain*

[2]*Área de Lenguas y Educación, Universidad Europea del Atlántico, Santander, Spain*

[3]*Centro de Ciencia Cognitiva (C3), Universidad Nebrija, Madrid, Spain*

[4]*Department of Languages and Linguistics, The Arctic University of Norway, Tromsø, Norway*

The current study focuses on how different scales with varying demands can affect our subjective assessments. We carried out 2 experiments in which we asked participants to rate how happy or sad morphed images of faces looked. The two extremes were the original happy and original sad faces with 4 morphs in between. We manipulated language of the task—namely, half of the participants carried it out in their native language, Spanish, and the other half in their foreign language, English—and type of scale. Within type of scale, we compared verbal and brightness scales. We found that, while language did not have an effect on the assessment, type of scale did. The brightness scale led to overall higher ratings, i.e., assessing all faces as somewhat happier. This provides a limitation on the foreign language effect, as well as evidence for the influence of the cognitive demands of a scale on emotionality assessments.

[*] **Corresponding author:** Candice Frances. Basque Center on Cognition, Brain and Language (BCBL). Paseo Mikeletegi 69, 2nd floor, 20009 Donostia – Spain. E-mail: c.frances@bcbl.eu, candice.frances@ncf.edu. **Ackowledgements:** This research has been partially funded by grants PGC2018-097145-B-I00 (from the Agencia Estatal de Investigación) and SEV-2015-0490 from the Spanish Government. CF is supported by a MINECO predoctoral grant from Spanish government (BED-2016-077169).

The assessment of what we perceive may seem trivial but can have very important consequences, transcending seemingly simple evaluations and having strong implications for one's health. One example of this is the case of scales of pain perception (see Hjermstad et al., 2011 for a review). The literature on pain perception focuses on the importance of correctly assessing patients' subjective states in order to provide the best treatment. For example, understanding the intensity of their pain can help establish how much medication the patient needs, while staying within safe dosage limits. If pain is underestimated, patients suffer greatly for having to withstand high levels of pain, whereas if it is overestimated, they run the risk of getting excessive amounts of powerful and addictive medications.

Within the context of pain studies, verbal scales have been found to be helpful for putting our perceptions into words (Au et al., 1994). As mentioned before, these studies emphasize the fact that the scale we use can change our assessments (Brunelli et al., 2010) and consequently, treatment decisions. Therefore, it is essential to understand precisely how different scales affect our assessments in order to select the most effective ones and compensate for their biases when evaluating them. Several studies comparing various non-verbal scales—namely numerical (Brunelli et al., 2010) and visual analog scales (i.e., a continuous line with extreme labels at both ends)—with verbal ones, found that non-verbal scales are superior to verbal scales in providing more valid assessments of pain (Thong, Jensen, Miró, & Tan, 2018). This puts into question how helpful language-based scales are when it comes to assessing our subjective perceptions.

With the spread of globalization and migration, it is progressively more common to communicate in a foreign language in all aspects of one's life. This means that verbal scales are often used by non-native speakers of the language. This brings attention to the potential impact of the language of the scale when making decisions, as the imprecision of verbal scales may in fact be affecting people differently. Given the massive presence of English on the Internet, many people find themselves using this foreign language on a daily basis for a broad range of assessments. Some of these interactions occur in companies that operate across international and linguistic borders and use English as the lingua franca. It is relatively common in these scenarios to request feedback in that common language, and this feedback may have strong consequences, such as affecting workers' performance evaluations or job security. In addition, in cases of migration, foreign language use can affect assessments of health and need in minority groups, as responses may vary depending on whether surveys are provided in a native or foreign language (e.g., Moradi, Sidorchuk, & Hallqvist, 2010, but not Kinnunen et al., 2015).

In this area, namely foreign language research, most studies have focused on the effects of foreign language on emotionality. For example, performing tasks in a foreign language context affects how we experience and perceive emotions (Caldwell-Harris & Ayçiçeği-Dinn, 2009; Dewaele, 2004; Ivaz, Costa, & Duñabeitia, 2016). This is because foreign language use leads to reduced emotionality, modulating valence and intensity, as well as effecting how we interpret different cues (Costa, Foucart, Arnon, Aparici, & Apesteguia, 2014; Keysar, Hayakawa, & An, 2012). This relates to the anchor contraction effect (ACE, De Langhe, Puntoni, Fernandes, & van Osselaer, 2011), which suggests that participants have a tendency to report more extreme emotions using nonnative language end labels rather than native language ones. This phenomenon is explained by an underestimation of the intensity of the end labels that leads to assessments closer to the end-points of the scale. Therefore, participants equate the labels with lower emotionality and thus consider that their emotions match up better with more extreme labels in the foreign language.

These effects of language also relate to issues with verbal scales in general. For example, verbal scales tend to be vague and have a large amount of inter-individual variability (Budescu & Wallsten, 1995) even between experts in the subject matter (Rudram, 1996; Shor & Weisner, 1999). This issue is particularly salient when laypeople are asked to analyze expert assessments and translate them into percentage of support for a statement or alternative, in which case they tend to underestimate the true likelihood of the statement (Martire & Watkins, 2015). As mentioned before, number scales are often used as a way to avoid these problems, but another approach is to use images, although this has shown mixed results. Having emoticons accompany the verbal scale can reduce or even get rid of the ACE (De Langhe et al., 2011). This can also be effective for assessing emotions that are difficult to verbalize (Elder, 2018) and provide more consistent responses regardless of instruction quality (Toet et al., 2018). Furthermore, in the case of pain assessment, this type of scale provides responses that are less contaminated by other factors, such as unpleasantness (Thong et al., 2018). Nevertheless, several studies have suggested that these scales do not necessarily provide an improvement over verbal scales (DeCastellarnau, 2018), making the effects of verbal versus non-verbal scales quite unclear. A similar type of scale that has received less attention is color intensity. These types of scales can provide some of the same benefits as emoticon scales while being more general. In particular—and importantly for this study—, this type of scale also reduces the ACE (De Langhe et al., 2011). Given the importance of assessments for communication, it is relevant to ask whether and how they are affected by the type of label (verbal or non-verbal) as well

as by the proximity of the speaker to the language (foreign or native language).

The current study focuses on the impact of language on making decisions and providing judgements of emotional faces by comparing verbal and non-verbal scales within two language contexts—namely, native and foreign. The particular task we chose was an assessment of how happy or sad people in morphed images looked. The scale went from the original happy (100% happy) to the sad (100% sad) faces with 4 morphs in between, for a grand total of 6 levels. The reason for choosing this task was that it is simple and people are particularly good at it, especially when detecting joy (Martinez & Du, 2012). In addition, this assessment relies on subjective measures that can be contrasted against the objective reality of the stimuli. Furthermore, these 6 images had a one-to-one correspondence to the response scale, reducing the amount of variability between subjects and assessments. Additionally, we chose to test emotional faces because there is also evidence that the assessments of such stimuli can be affected by the context they are in (Rim Noh & Isaacowitz, 2013) and consequently, if there is emotional detachment in a foreign language, this language context is more likely to affect these assessments.

With this aim in mind, we carried out two experiments on facial emotion perception and assessment of sadness and happiness. In both experiments, we asked participants to label the emotion displayed on a scale from sad to happy (a valence assessment task). We compared ratings on a non-verbal scale (using a brightness scale) and a language-based verbal scale, either in participants' native or foreign language. This way, we expected to see how emotional assessments change as a function of the type of scale used and to establish the manner in which language affects scales differently depending on the nativeness of the language used. In particular, we chose a gray brightness scale because it is essentially visual and implies minimal language processing, as there are no specific names for each of the levels, making it difficult to translate into words. In addition, given that the semantic connotations of colors are not completely consistent between cultures—e.g., the placement of blue within the positive-negative spectrum is reported to be the opposite in Spanish (Soriano & Valenzuela, 2009) as it is reported in English as well as several other cultures (Adams & Osgood, 1973)—, we found that the gray brightness scale (see Figure 1) was the most appropriate, as brightness is consistently evaluated as positive and darkness as negative (Adams & Osgood, 1973; Hemphill, 1996; Soriano & Valenzuela, 2009; Wexner, 1954).

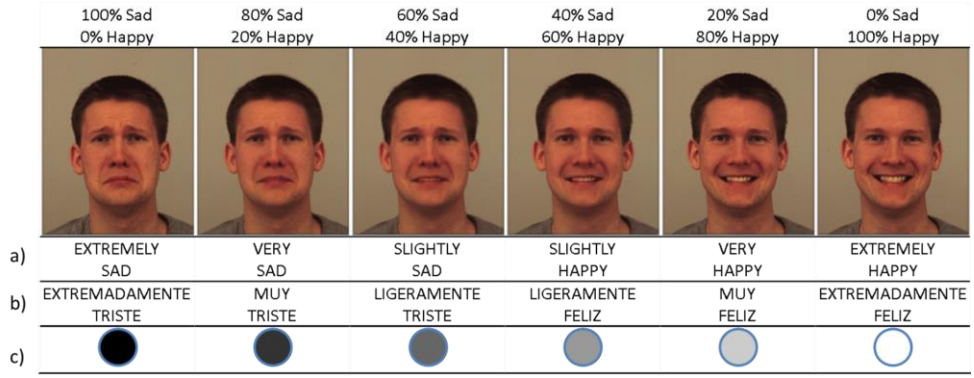


Fig. 1: Continuum of morphs for one of the stimulus faces. (a) Verbal labels used in the FL version (English). b) Verbal labels used in the NL version (Spanish). (c) Gray labels used in the brightness scale of the NL and FL versions.

EXPERIMENT 1: Brightness and Verbal Scale in a Mixed Design

METHOD

Participants. Participants were 84 native Spanish-speaking students (61 females, M_{age} = 36.38 years, *SD* = 9.36, see Appendix) from the Universidad Europea del Atlántico (UNEATLANTICO). Half of the participants were randomly assigned to the native language and the rest to the foreign language context, with conditions matched for age, gender, English knowledge, percentage of daily English use, age of acquisition of English, Spanish language skills (Spanish Lextale; Izura, Cuetos, & Brysbaert, 2014), and English language skills (English Lextale, Lemhöfer & Broersma, 2012). See Appendix for means and standard deviations by group and experiment. All participants gave informed consent and the experimental protocol was approved by the Ethics Committee of UNEATLANTICO.

Stimuli. The images of 3 male and 3 female faces displaying happy and sad expressions were taken from the Karolinska Directed Emotional Faces (KDEF) (Lundwist, Flyict, & Ohman, 1998) and morphed using FreeMorphing software to create 6 levels of emotion (see Figure 1), resulting in a total of 36 images. Images were jpg format and 400 by 300 pixels, with hair and part of a gray t-shirt visible in every image (see Blair, Murray, & Mitchell, 2001, and Niedenthal, Halberstadt, Margolin, & Innes-Ker, 2000, for similar approaches).

Procedure. Participants did an online survey, first answering demographic and linguistic background questions, and then the experimental

tasks. Critically, language context varied across groups—all text displayed either in their native language (Spanish) or foreign language (English). The language condition was kept strictly between subjects for several reasons. First, it was important to avoid the effects of changing or mixing languages (Gollan & Ferreira, 2009), as well as contamination between conditions. In addition, the task itself was already quite long, and diminishing the number of stimuli (currently 6 per level) would have rendered the power too low for the experiment. Therefore, we opted for a between-subjects design with strict matching between groups.

For the experimental task, participants were presented with an image and a scale and were asked to evaluate how happy or sad the person in each image looked. The scale was either a verbal or a brightness one—only one type of scale per page (see Figure 1 for labels)—and each page contained 9 images to evaluate. The pages with each type of scale were randomized such that the scales were intermixed (e.g., first evaluate 9 images using the verbal scale, then 9 using the brightness scale, then 9 more with the brightness scale, followed by 9 in the verbal, and so on for 8 pages). The order of the values within the scale was presented randomized by page (i.e., they were randomized for each page, but consistent throughout the page), so that participants had to read each label in order to correctly complete the task (see Figure 2 for an example of what the page looked like). They saw each of the 36 images twice in sets of 9 so that there were 8 pages, 4 using a brightness scale and 4 a verbal scale. By the end of the task, participants had rated each image twice, once using each scale, but on different pages.

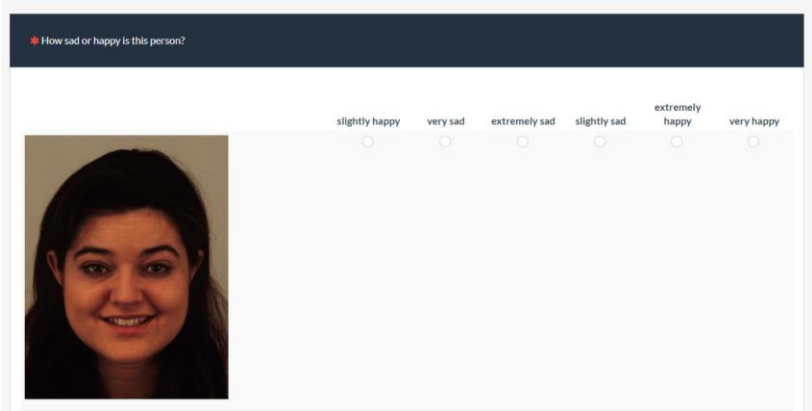


Fig. 2: Example page from the English verbal section.

RESULTS

Analyses were conducted with linear mixed-effect models (lme) using the lme4 (Bates, Maechler, Bolker, & Walker, 2015) package in R (R Core Team, 2018). Significance p-values and Type III F-statistics for main effects, interactions, and planned comparisons were calculated using Satterthwaite approximations to denominator degrees of freedom as implemented in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017). With Rating as the dependent variable, the fixed structure of the models was composed of the factors Language Context (native vs. foreign), Scale (verbal vs. brightness), and the ordinal variable Level (range: 1 to 6), as well as by their interactions. The model with the maximal within-unit random effects structure (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013) did not converge. Therefore, the model included by participant random slopes and intercepts for Scale and Level as well as their interaction and by items random slopes for the interaction of Language, Scale, and Level. The predictors Language, Scale, and Level were centered prior to analysis so that the reference point (the intercept) corresponded to the average between Languages and Scales for the midpoint of Level (i.e., the average over all the morphs).

This analysis showed an expected main effect of Level, with "happy" faces rated as happier and "sad" faces as sadder, $F(1, 8.10) = 403.31$, $p < .001$, estimate [Lower – Upper 95% CI] = -.714 [-.785 – -.643]. There was also a main effect of Scale, with items rated as more positive or "happy" using the brightness scale than the verbal scale, $F(1, 81.89) = 42.61$, $p < .001$, estimate [Lower – Upper 95% CI] = -.298 [-.389 – -.207]. Finally, there was no main effect of Language, $F(1, 81.91) = .17$, $p = .68$, estimate [Lower – Upper 95% CI] = .021 [-.083 – .126]. There were also no interactions between Language and Scale [$F(1, 81.89) = 1.10$, $p = .30$, estimate [Lower – Upper 95% CI] = -.096 [-.278 – .087]], Language and Level [$F(1, 46.32) = .93$, $p = .34$, estimate [Lower – Upper 95% CI] = -.037 [-.114 – .040]], Scale and Level [$F(1, 75.60) = 1.84$, $p = .17$, estimate [Lower – Upper 95% CI] = -.039 [-.096 – .018]], nor the triple interaction [$F(1, 62.02) = .75$, $p = .39$, estimate [Lower – Upper 95% CI] = .051 [-.067 – .168]].

Scores using the brightness scale were more positive (higher) than those coming from verbal scales. On the other hand, there was no effect of language. This suggests that language context (native or foreign) does not have a strong influence in the way people assess emotional faces, but that other factors such as type of scale—in this case brightness versus verbal—lead to different assessments.

EXPERIMENT 2: Brightness and Verbal Scale in a Blocked Design

In order to verify the results of Experiment 1, we explored our initial observations in further detail. In Experiment 2, we increased the number of stimuli and participants and showed the images one-by-one in a blocked design. By blocking presentation by scale, we maximized the chances of uncovering any potential difference between conditions and, if no differential effects arise, then one could safely conclude that the processing of emotional faces is not affected by the language in which the emotions are being rated.

METHOD

Participants. Participants were 130 native Spanish-speaking students (86 females, M_{age} = 34.52 years, *SD* = 8.53—see Appendix) from the same subject pool and distribution as in Experiment 1.

Stimuli. Stimuli were built as in Experiment 1 (see Figure 1), but with additional images for a total of 10 faces (5 male, 5 female).

Procedure. Participants followed the same procedure as in Experiment 1, except that images were presented one-by-one—one image per page—and blocked by scale—all of the faces were rated using the brightness scale first and then using the verbal scale. The rationale behind this was to avoid any possible interference from the verbal scale on the brightness scale. Whereas the brightness scale is unlikely to influence the verbal scale, doing the assessment using the verbal scale first might lead to "converting" the brightness into a proxy for the verbal scale. This way, this type of contamination was avoided.

RESULTS AND DISCUSSION

Analyses were conducted with linear mixed-effect models (lme) in the same way as in Experiment 1, with the same variables (same response variable and random and fixed effects).

This analysis showed an expected main effect of Level, with "happy" faces rated as happier and "sad" faces as sadder, $F(1, 31.49) = 1599.90$, $p < .001$, estimate [Lower – Upper 95% CI] = -.725 [-.761 – -.688]. There was also a main effect of Scale, with items rated as more positive or "happy" using the brightness scale than the verbal scale, $F(1, 126.93) = 25.33$, $p < .001$, estimate [Lower – Upper 95% CI] = -.437 [-.611 – -.263]. Finally, there was

no main effect of Language, $F(1, 127.01) = .62$, $p = .43$, estimate [Lower – Upper 95% CI] = -.082 [-.290 – .126]. There were also no interactions between Language and Scale [$F(1, 126.94) = .39$, $p = .53$, estimate [Lower – Upper 95% CI] = -.108 [-.456 – .239]], Language and Level [$F(1, 128.55) = .81$, $p = .37$, estimate [Lower – Upper 95% CI] = .026 [-.032 – .084]], Scale and Level [$F(1, 127.07) = .0003$, $p = .99$, estimate [Lower – Upper 95% CI] = -4.06×10^{-4} [-.046 – .045]], nor the triple interaction [$F(1, 127.58) = .69$, $p = .41$, estimate [Lower – Upper 95% CI] = .038 [-.053 – .130]].

Although in this experiment we cannot fully exclude the possibility that the participants' responses to the verbal scale were influenced by their memory of the color section, the results are fully in line with those of Experiment 1. In this second experiment, the effect of scale showed more positive scores for the brightness scale than for the verbal scale, but we again failed to find an effect of language. It seems clear that foreign language does not affect the assessment of emotionality in static faces, while the use of non-verbal labels clearly changes our decisional criteria.

DISCUSSION

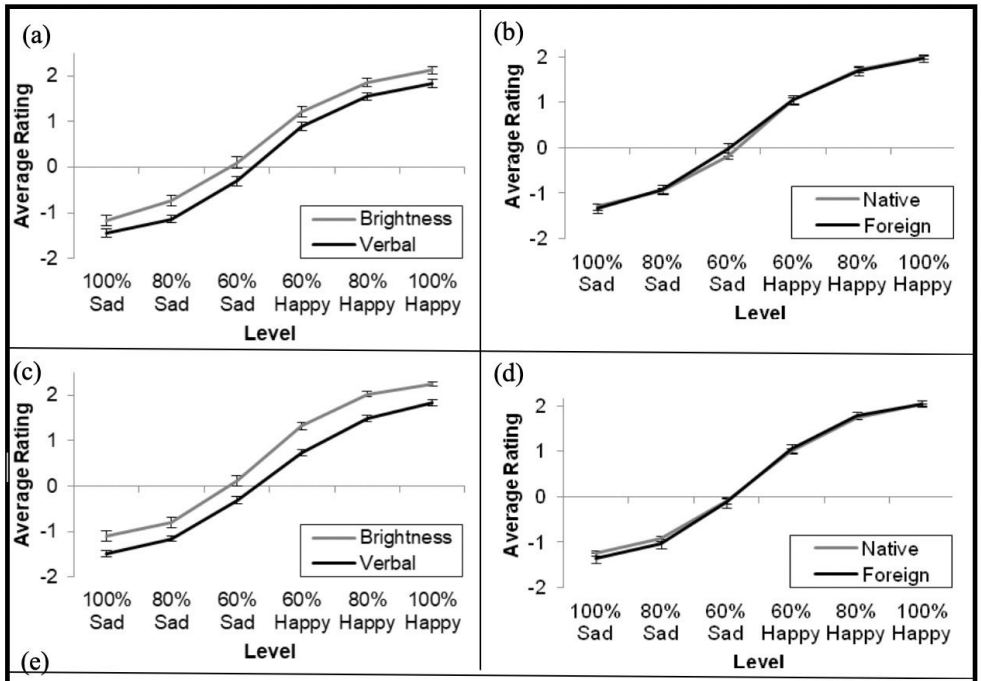
In the current exploratory, methodologically-oriented study, we explored how verbal and non-verbal scales affect our assessments. To this end, we asked participants to rate their perceived emotionality of a series of morphed faces using verbal and non-verbal scales. In addition, we assessed whether language context has an effect on these assessments, exploring potential differences between native and foreign language-mediated judgements. To this end, we had participants evaluate the sadness and happiness of emotional faces using verbal and non-verbal (brightness) labels in a native or a foreign language.

In both experiments, we observed a positive shift in values when using the brightness scale. This suggests that using a brightness scale can lead to more positive assessments. Put differently, this implies that when using a verbal scale, responses tend to be less emotionally charged overall as compared with a brightness scale. As we suggested before, brightness scales include no linguistic information and, by removing language, the task becomes fully visual with the comparison occurring in the same modality. Here, we observed a general positive shift towards brighter, "happier" tones. Although it may seem somewhat counterintuitive at first, one possible explanation for this is that cognitive load could be increased. This is because both the assessment and the object are in the same modality and, thus, more items need to be assessed at the same time through the same network (Lavie,

2005; Lavie & Cox, 1997). Furthermore, research on cognitive load has found similar results when load is increased in other ways (Sweller, Ayres, & Kalyuga, 2011). In fact, preceding studies have found that increasing cognitive load diverted attention away from negatively-valenced stimuli (Maranges, Schmeichel, & Baumeister, 2017) or consequences (Drolet & Frances Luce, 2004). For this reason, it is suggested that the measurement instrument should aim to minimize perceptual load in order to avoid interfering with the assessment (Wissmath, Weibel, & Mast, 2010). The current results cannot disambiguate precisely whether the differential effects are the result of increased cognitive load or of an attentional shift. Although we recognize that the original goal of this study was not to delve into cognitive load, our results do seem to be explained in this context and fit well within this literature.

Another possible explanation has to do with the vagueness of verbal labels. Perhaps the color labels are easier to adapt to the precise level of arousal elicited by an image, whereas matching this arousal to a verbal label might be more difficult. In other words, the amount of brightness in the scale marks a very clear level within the scale, whereas linguistic modifiers (e.g., extremely, very, and slightly) might hold a more arbitrary, less clear relationship with arousal.

One advantage that our task had—e.g., over pain assessment—is that an actual correct value within the range of stimuli could be calculated. By equating each of the levels of the morphs with the levels of the scale, we can easily calculate the expected average assessment for each of the morphs and then contrast this with the actual ratings. In practical terms, if the full 100% happy morph is equated with “extremely happy” or a value of 2.5 and, conversely, the full 100% sad morph is equated with “extremely sad” or a value of -2.5, we could also extrapolate that the morph that is 60% sad and 40% happy should get approximately a -.5 value (or “slightly sad”), on average. Following this logic, it is worth noting that the averaged reported values within the verbal scale were actually closer to the expected values than those within the brightness scale. For example, the aforementioned morph that should get approximately a -.5 value on average (see Figure 3c), gets a value much closer to this “optimal” value with the verbal scale (Experiment 1: -.306; Experiment 2: -.318) than with the color scale (Experiment 1: .095; Experiment 2: .109). This would suggest that, in this case, the verbal scale not only leads to less positive assessments, but that it allows for more accurate ratings than the non-verbal scale, at least within this stimulus set.



(e)

		Experiment 1		Experiment 2	
	Theoretical	Brightness	Verbal	Brightness	Verbal
Extremely Happy	2.5	2.12 (0.36)	1.83 (0.39)	2.25 (0.29)	1.83 (0.36)
Very Happy	1.5	1.85 (0.44)	1.55 (0.40)	2.02 (0.36)	1.49 (0.36)
Slightly Happy	0.5	1.21 (0.50)	0.88 (0.42)	1.32 (0.51)	0.74 (0.39)
Slightly Sad	-0.5	0.09 (0.6)	-0.30 (0.48)	0.10 (0.67)	-0.31 (0.45)
Very Sad	-1.5	-0.72 (0.53)	-1.13 (0.37)	-0.80 (0.68)	-1.15 (0.38)
Extremely Sad	-2.5	-1.17 (0.58)	-1.44 (0.39)	-1.10 (0.66)	-1.48 (0.39)

Fig. 3(a) Average rating by scale for each level of emotion in Experiment 1. (b) Average rating by language for each level of emotion in Experiment 1. (c) Average rating by scale for each level of emotion in Experiment 2. (d) Average rating by language for each level of emotion in Experiment 2. In all cases, error bars show the 95% confidence intervals. (e) Means and standard deviations for each of the levels and scales by experiment.

The second relevant finding of this study corresponds to the (lack of) impact of the language of the verbal scales (foreign vs. native) on emotion assessments. Given that using a foreign language increases cognitive load due to the differences in the knowledge and use of foreign and native languages, we expected similar results for the foreign language as we observed with the brightness scale. In fact, preceding studies have demonstrated a partial emotional detachment of bilingual participants when presented with certain scenarios in their foreign language (see García-Palacios et al., 2018; Iacozza, Costa, & Duñabeitia, 2017). But, contrary to our initial intuition, evaluations were not affected by a foreign language effect, nor was the effect of scale

modulated by language. We did not observe any effects of language, even in the second study, with an increased number of stimuli and, more importantly, a blocked design.

The literature on the foreign language effect has found many far reaching consequences of using a foreign language (Caldwell-Harris, 2009; Corey et al., 2017; Costa et al., 2014; Dewaele, 2004, 2010, 2011; Harris et al., 2006; Keysar et al., 2012; Pell et al., 2009; Schrauf, 2000). However, limits of the foreign language effect have been also reported in the literature on emotion recognition. The current study expands on the findings of Lorette and Dewaele (2015) which suggest that differences in recognition ability can be explained by linguistic ability and culture. Hence, the lack of differences between the foreign and native language conditions in the current study could be linked to the fact that participants in both groups were of the same culture and background and were sufficiently proficient to carry out the task without difficulty.

Here, we showed that the use of scales that vary in their cognitive demands can affect our assessments of emotions in faces. In particular, brightness scales led to more positive assessments. Importantly, we did not find a foreign language effect. This suggests that the foreign language effect is contingent on the difficulty of the language used, and it does not simply reflect an overall reduction in emotionality due to experience with the language. Our results suggest further effects of the type of scale and measures used to assess perceived emotions in faces. These need to be taken into consideration in order to fully understand how emotions are processed and evaluated. It seems that emotions in faces are assessed differently depending on the elements we use to provide our judgements. On the bright side, be it at home or on holidays in a foreign country, we can always detect a friendly smile.

RESUMEN

El estudio actual se centra en cómo escalas diferentes con demandas cognitivas variadas pueden afectar nuestras evaluaciones subjetivas. Se realizaron dos experimentos en los que se les pidió a los participantes que evaluaran cuán felices o tristes les resultaban las expresiones de algunas caras. Los dos extremos eran las caras tristes y felices originales, con cuatro variaciones en el medio. Manipulamos el idioma de la tarea, de tal manera que la mitad de los participantes realizaron el estudio en su idioma nativo (español) y la otra mitad en su idioma extranjero (inglés), y también variamos el tipo de escala. Comparamos dos tipos de escalas de valoración: verbales y

de brillo (gris). Encontramos que, si bien la lengua no tuvo un efecto en la evaluación, el tipo de escala sí lo tuvo: la escala de brillo llevó a calificaciones más altas en general. Es decir, los participantes evaluaron todas las caras como algo más felices con la escala de brillo. Esto ofrece una limitación al impacto de los efectos de lenguas extranjeras, proporcionando evidencia sobre la influencia que tienen las demandas cognitivas de la escala en las evaluaciones de emocionalidad.

REFERENCES

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Adams, F. M., & Osgood, C. E. (1973). A cross-cultural study of the affective meanings of color. *Journal of Cross-Cultural Psychology*, *4*, 135–156. https://doi.org/10.1177/002202217300400201

Au, E., Loprinzi, C. L., Dhodapkar, M., Nelson, T., Novotny, P., & Hammack, J. (1994). Regular use of a verbal pain scale improves the understanding of oncology inpatient pain intensity. *Journal of Clinical Oncology*, *12*, 2751–2755. https://doi.org/10.1200/jco.1994.12.12.2751

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*(328), 3–4. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blair, R. J. R., Murray, L., & Mitchell, D. G. V. (2001). A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Child Psychology*, *29*, 491–498. https://doi.org/10.1023/A:1012225108281

Brunelli, C., Zecca, E., Martini, C., Campa, T., Fagnoni, E., Bagnasco, M., … Caraceni, A. (2010). Comparison of numerical and verbal rating scales to measure pain exacerbations in patients with chronic cancer pain. *Health and Quality of Life Outcomes*, *8*(42), 1–8. https://doi.org/10.1186/1477-7525-8-42

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: general principles and empirical evidence. *Psychology of Learning and Motivation - Advances in Research and Theory*, *32*, 275–318. https://doi.org/10.1016/S0079-7421(08)60313-8

Caldwell-Harris, C. L. (2009). Emotion-memory effects in bilingual speakers: A levels-of-processing approach. *Bilingualism: Language and Cognition*, *12*, 291–303. https://doi.org/10.1017/S1366728909990125

Caldwell-Harris, C. L., & 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

Corey, J. D., Hayakawa, S., Foucart, A., Aparici, M., Botella, J., Costa, A., & Keysar, B. (2017). Our moral choices are foreign to us. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1–20. https://doi.org/10.1037/xlm0000356

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). "Piensa" twice: On the foreign language effect in decision making. *Cognition*, *130*, 236–254. https://doi.org/10.1016/j.cognition.2013.11.010

De Langhe, B., Puntoni, S., Fernandes, D. H., & van Osselaer, S. (2011). The anchor contraction effect in international marketing research. *Journal of Marketing Research*, *48*, 366–380. https://doi.org/10.2307/23033437

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity*, *52*, 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

Dewaele, J.-M. (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, J.-M. (2010). Multilingualism and affordances: Variation in self-perceived communicative competence and communicative anxiety in French L1, L2, L3 and L4. *International Review of Applied Linguistics in Language Teaching*, *48*, 105–129. https://doi.org/10.1515/iral.2010.006

Dewaele, J. M. (2011). Reflections on the emotional and psychological aspects of foreign language learning and use. *International Journal of English Studies*, *22*(1), 23–42.

Drolet, A., & Frances Luce, M. (2004). The rationalizing effects of cognitive load on emotion-based trade-off avoidance. *Journal of Consumer Research*, *31*, 63–77. https://doi.org/10.1086/383424

Duñabeitia, J. A., & Costa, A. (2015). Lying in a native and foreign language. *Psychonomic Bulletin and Review*, *22*, 1124–1129. https://doi.org/10.3758/s13423-014-0781-4

Elder, A. M. (2018). What words can't say: Emoji and other non-verbal elements of technologically-mediated communication. *Journal of Information, Communication and Ethics in Society*, *16*, 2–15. https://doi.org/10.1108/JICES-08-2017-0050

García-Palacios, A., Costa, A., Castilla, D., Del Río, E., Casaponsa, A., & Duñabeitia, J. A. (2018). The effect of foreign language in fear acquisition. *Scientific Reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-19352-8

Gollan, T. H., & Ferreira, V. S. (2009). Should I stay or should I switch? A cost-benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 640–665. https://doi.org/10.1037/a0014981.Should

Harris, C. L., Gleason, J. B., & Ayçiçeği, A. (2006). When is a first language more emotional? Psychophysiological evidence from bilingual speakers. In A. Pavlenko (Ed.), *Bilingual minds: Emotional experience, expression, and representation* (Vol. 56, pp. 257–283). Clevedon, UK: Multilingual Matters.

Hemphill, M. (1996). A note on adults' color-emotion associations. *The Journal of Genetic Psychology*, *157*, 275–280. https://doi.org/http://dx.doi.org/10.1080/00221325.1996.9914865

Hjermstad, M. J., Fayers, P. M., Haugen, D. F., Caraceni, A., Hanks, G. W., Loge, J. H., … Kaasa, S. (2011). Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: A systematic literature review. *Journal of Pain and Symptom Management*, *41*, 1073–1093. https://doi.org/10.1016/j.jpainsymman.2010.08.016

Iacozza, S., Costa, A., & Duñabeitia, J. A. (2017). What do your eyes reveal about your foreign language? Reading emotional sentences in a native and foreign language. *PLoS ONE*, *12*(10), 1–10. https://doi.org/10.1371/journal.pone.0186027

Ivaz, L., Costa, A., & Duñabeitia, J. A. (2016). The emotional impact of being myself: Emotions and foreign-language processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*, 489–496. https://doi.org//10.1037/xlm0000179

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, *35*, 49–66.

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

Kinnunen, J. M., Malin, M., Raisamo, S. U., Lindfors, P. L., Pere, L. A., & Rimpelä, A. H. (2015). Feasibility of using a multilingual web survey in studying the health of ethnic minority youth. *JMIR Research Protocols*, *4*(2), 1–9. https://doi.org/10.2196/resprot.3655

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, *9*, 75–82. https://doi.org/10.1016/j.tics.2004.12.004

Lavie, N., & Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, *8*, 395–398. https://doi.org/10.1111/j.1467-9280.1997.tb00432.x

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lorette, P., & Dewaele, J.-M. (2015). Emotion recognition ability in English among L1 and LX users of English. *International Journal of Language and Culture*, *2*, 62–86. https://doi.org/10.1075/ijolc.2.1.03lor

Lundwist, D., Flyict, A., & Ohman, A. (1998). *The Karolinska Directed emotional faces–KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*. Stockholm.

Maranges, H. M., Schmeichel, B. J., & Baumeister, R. F. (2017). Comparing cognitive load and self-regulatory depletion: Effects on emotions and cognitions. *Learning and Instruction*, *51*, 74–84. https://doi.org/10.1016/j.learninstruc.2016.10.010

Martinez, A. M., & Du, S. (2012). A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *Journal of Machine Learning Research*, *13*, 1589–1608. https://doi.org/10.1007/978-3-319-57021-1_6

Martire, K. A., & Watkins, I. (2015). Perception problems of the verbal scale: A reanalysis and application of a membership function approach. *Science and Justice*, *55*, 264–273. https://doi.org/10.1016/j.scijus.2015.01.002

Moradi, T., Sidorchuk, A., & Hallqvist, J. (2010). Translation of questionnaire increases the response rate in immigrants: Filling the language gap or feeling of inclusion? *Scandinavian Journal of Public Health*, *38*, 889–892. https://doi.org/10.1177/1403494810374220

Niedenthal, P. M., Halberstadt, J. B., Margolin, J., & Innes-Ker, Å. H. (2000). Emotional state and the detection of change in facial expression of emotion. *European Journal of Social Psychology*, *30*, 211–222. https://doi.org/10.1002/(SICI)1099-0992(200003/04)30:2<211::AID-EJSP988>3.0.CO;2-3

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*, 107–120. https://doi.org/10.1007/s10919-008-0065-7

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rim Noh, S., & Isaacowitz, D. M. (2013). Emotional faces in context: Age differences in recognition accuracy and scanning patterns. *Emotion*, *13*, 238–249. https://doi.org/10.1037/a0030234.Emotional

Rudram, D. A. (1996). Interpretation of scientific evidence. *Science and Justice*, *36*, 133–138. https://doi.org/10.1016/S1355-0306(96)72587-X

Schrauf, R. W. (2000). Bilingual autobiographical memory: Experimental studies and clinical cases. *Culture & Psychology*, *6*, 387–417. https://doi.org/10.1177/1354067X0064001

Shor, Y., & Weisner, S. (1999). A survey on the conclusions drawn on the same footwear marks obtained in actual cases by several experts throughout the world. *Journal of Forensic Sciences*, *44*, 380–384. https://doi.org/10.1520/JFS14468J

Soriano, C., & Valenzuela, J. (2009). Emotion and colour across languages: Implicit associations in Spanish colour terms. *Social Science Information*, *48*, 421–445. https://doi.org/10.1177/0539018409106199

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. (J. M. Spector & S. P. Lajoie, Eds.). New York: Springer. https://doi.org/10.1007/978-1-4419-8126-4

Thong, I. S. K., Jensen, M. P., Miró, J., & Tan, G. (2018). The validity of pain intensity measures: What do the NRS, VAS, VRS, and FPS-R measure? *Scandinavian Journal of Pain*, *18*, 99–107. https://doi.org/10.1515/sjpain-2018-0012

Toet, A., Kaneko, D., Ushiama, S., Hoving, S., Kruijf, I. de, Brouwer, A. M., … van Erp, J. B. F. (2018). EmojiGrid: A 2D pictorial scale for the assessment of food elicited emotions. *Frontiers in Psychology*, *9*(2396), 1–21. https://doi.org/10.3389/fpsyg.2018.02396

Wexner, L. B. (1954). The degree to which colors (hues) are associated with mood-tones. *The Journal of Applied Psychology*, *38*(6), 6–9.

Wissmath, B., Weibel, D., & Mast, F. W. (2010). Measuring presence with verbal versus pictorial scales: A comparison between online- and ex post-ratings. *Virtual Reality*, *14*, 43–53. https://doi.org/10.1007/s10055-009-0127-0

(Manuscript received: 28 November 2018; accepted: 20 August 2019)

APPENDIX:

The following variables were equated between groups. Below are the means and standard deviations for each of them, as well as the Bayes factors for the continuous variables.

	Experiment 1			Experiment 2		
	EN [*Mean* (*SD*)]	ES [*Mean* (*SD*)]	**BF_{01} (error %)**	EN [*Mean* (*SD*)]	ES [*Mean* (*SD*)]	**BF_{01} (error %)**
Age	35.26 (8.99)	37.50 (9.69)	**2.6 (0.03)**	34.03 (9.28)	35.00 (7.74)	**4.41 (2.48 x 10^{-6})**
Overall Level	5.91 (2.36)	5.98 (2.50)	**4.36 (0.03)**	6.71 (2.10)	6.29 (2.26)	**3.12 (0.03)**
Listening	5.79 (2.63)	5.48 (2.93)	**3.92 (0.03)**	6.45 (2.33)	6.40 (2.41)	**5.30 (2.59 x 10^{-6})**
Reading	6.83 (2.57)	6.57 (2.62)	**4.00 (0.03)**	7.19 (2.21)	7.00 (2.22)	**4.81 (2.53 x 10^{-6})**
Speaking	5.86 (2.62)	5.43 (2.75)	**3.48 (0.03)**	6.79 (2.34)	6.05 (2.60)	**1.44 (2.00 x 10^{-3})**
Writing	6.17 (2.58)	5.67 (2.63)	**3.13 (0.03)**	6.92 (2.29)	6.35 (2.50)	**2.32 (0.01)**
Daily English	36.88 (27.3)	35.98 (27.67)	**4.35 (0.03)**	43.17 (28.91)	38.23 (24.87)	**3.25 (0.03)**
AOA	15.26 (8.46)	13.55 (7.12)	**2.83 (0.03)**	13.02 (6.37)	13.39 (5.37)	**5.03 (2.56 x 10^{-6})**
Spanish LexTALE	0.91 (0.06)	0.90 (0.07)	**3.75 (0.03)**	0.90 (0.07)	0.91 (0.06)	**4.02 (2.43 x 10^{-6})**
English LexTALE	0.75 (0.12)	0.73 (0.10)	**3.32 (0.03)**	0.74 (0.10)	0.72 (0.09)	**3.45 (0.04)**

Note: The level assessments (both overall and of specific skills) were self-assessments of English given by participants on a 1 to 10 scale, where 10 was defined as the level expected in a native English speaker. Daily English refers to the percentage of the day the participant spends using English. AOA refers to Age of Acquisition of English. LexTALE scores are on a scale from 0 to 1.

Below are the contingency tables for the categorical variables.

	Experiment 1		Experiment 2	
Gender	English	Spanish	English	Spanish
Female	31	30	43	43
Male	11	12	22	22
BF_{01}	**4.07**		**4.885**	

	Experiment 1		Experiment 2	
Level of Schooling	English	Spanish	English	Spanish
Some High School	0	0	0	1
Practical Training	3	2	5	7
University Degree	33	35	54	55
Masters Degree	5	5	6	2
PhD	1	0	0	0
BF_{01}	**78.81**		**75.30**	

	Experiment1		Experiment 2	
Living Abroad	English	Spanish	English	Spanish
Never	30	27	48	49
< 3 Months	10	10	9	11
3 - 6 Months	1	2	3	4
6 - 12 Months	1	3	5	1
> 12 Months	0	0	0	0
BF_{01}	**78.81**		**75.30**	

Note: This refers to the question whether participant had lived abroad, in an English speaking country, and for how long.

	Experiment1		Experiment 2	
English School	English	Spanish	English	Spanish
No	42	42	58	58
Yes	0	0	7	7
BF_{01}	**NA**		**7.37**	

Note: This question refers to whether the participant attended an English or bilingual (with English) school as a child.