Toivo, W., Scheepers, C. and Dewaele, J.-M. (2023) RER-LX: A new scale to measure reduced emotional resonance in bilinguals' later learnt language. *Bilingualism: Language and Cognition*, (doi: [10.1017/S1366728923000561](https://doi.org/10.1017/S1366728923000561))

RER-LX:

A new scale to measure reduced emotional resonance in bilinguals' later learnt language

Wilhelmiina Toivo[1], Christoph Scheepers[1] and Jean-Marc Dewaele[2]

[1] School of Psychology and Neuroscience, University of Glasgow, UK

[2] Department of Languages, Cultures and Applied Linguistics, Birkbeck University of London, UK

Corresponding Authors:

Wilhelmiina Toivo
University of Glasgow
School of Psychology and Neuroscience
Room 518B Boyd Orr building
University Avenue, Glasgow G12 8QW
wilhelmiina.toivo@glasgow.ac.uk

Christoph Scheepers
University of Glasgow
School of Psychology and Neuroscience
Room 550
62 Hillhead Street, Glasgow G12 8QB
christoph.scheepers@glasgow.ac.uk

Abstract

In two online survey studies (N = 688 and N = 247, respectively) we developed and validated a new psychometric scale for measuring emotional resonance reduction in bilinguals' LX ("later learnt language") relative to their L1 ("first language"). The final scale, dubbed RER-LX (for *Reduced Emotional Resonanc*e in *LX*), comprises 15 items and possesses a number of desirable psychometric properties. It yields good test reliability (expected alpha between 0.8 and 0.9), produces near-normally distributed test scores, and exhibits content validity in terms of its underlying factor structure. Moreover, it correlates well with the only other instrument previously used for the same purpose (BEQ subscale comprising *BEQ-swearing*, *BEQ-feelings* and *BEQ-anger*). However, compared to the BEQ items, RER-LX has significantly better discriminant validity in relation to LexTALE, a widely used measure of proficiency in English as a second language. Our new scale will be useful to researchers studying bilingualism and emotion.

Keywords: Bilingualism, Multilingualism, Emotion, Psychometrics

1.   Introduction

Barbara, a trilingual woman from Germany who is married to an English speaker and is living in the UK, pointed out in Dewaele (2013) that despite the fact of being equally proficient in her German L1 and English L2, emotion words in both languages felt very different: "The German words have (…) much more of a physical connection (…) like I tell off my children and if I do it in German, I get involved but if I do it in English it's a purely rational disciplinary thing. Expressing emotions in English to my English friends often feel as if I'm only pretending to have these emotions" (p. 94). Barbara's experience reflected a broad pattern in the data of 1,579 adult multilinguals from all over the world collected through the *Bilingualism and Emotions Questionnaire* (BEQ, Dewaele & Pavlenko, 2001-2003): emotional resonance of L1 emotion words was significantly stronger than that of emotion words from languages acquired later in life (LX), even among participants who rated themselves as being equally proficient in their L1 and LX (Dewaele, 2010).

The difference in emotional resonance between L1 and LX has been linked to the fact that they are acquired in very different contexts of acquisition: the L1 being absorbed in the multimodal and emotionally rich home context versus the LX typically being learnt in a relatively de-contextualised way between classroom walls (Harris, Gleason, & Ayçiçeği, 2006). Age of onset of acquisition, often hypothesised to be causally related to reduced emotional resonance in LX, shows some overlap with the emotional context of language learning but does not seem to be the cause of the L1-LX difference in itself (Harris et al. 2006).

L1 feels more strongly embodied because it is acquired during a period of intense affective socialisation (starting from birth) during which the child goes through "a process of integration of phonological forms of words and phrases with information from visual,

auditory, olfactory, tactile, kinesthetic, and visceral modalities, autobiographical memories, and affect" (Pavlenko, 2012, p.421). All the connected information is stored in the child's implicit memory. Emotion words and expressions thus gain a "physical connection" within the individual (cf. Barbara's personal account quoted above). Multilinguals who grow up with multiple L1s link the words swirling around them with rich autobiographical memories. They acquire the socio-pragmatic skills that allow them to gauge the exact impact of emotion words and to use these words appropriately. They might remember the startled look on the face of a teacher when using a taboo word in class, or the grin on the face of their friends when using that same word away from adults. This integration of emotion words and the emotional reactions they elicit will be much looser in the LX (typically acquired after the age of 3), and especially for LXs acquired through formal school instruction only. The main reason is "the decontextualized nature of the language classroom, which does not provide many opportunities for integration of all sensory modalities and verbal conditioning (other than foreign language anxiety) and thus leads to development of 'disembodied' words, used freely by speakers who do not experience their full impact" (Pavlenko 2012, p.421). LX words are more likely to be stored in the learner's declarative memory, meaning the user can access them and translate them, but multimodal connections remain sparse and weak. This has consequences for the way L1 and LX words are processed.

The processing of emotion words is automatic in L1(s) while in LXs there is semantic processing but no affective processing (Pavlenko 2012), which may explain why Barbara (quoted above) felt that telling her children off in English was a "purely rational disciplinary thing", in contrast to the extra depth when expressing her heartfelt anger in German. In other words, Barbara had a deep emotional connection with her German emotion words, and a much more superficial and intellectual one with her English emotion words.

Following on from more qualitative research into the matter, recent years have also seen a growing number of more *quantitative* attempts to measure differences in emotional resonance between L1 and LX (see Toivo & Scheepers, 2022 for a review). Here, we will briefly outline some of the key findings.

1.1 (Neuro-) physiological and behavioural measures

Physiological measures utilise techniques that are sensitive to the activation of the Autonomous Nervous System (ANS); pupillometry and the recording of skin-conductance responses are two prominent examples. Both techniques have provided solid support for the notion of reduced emotional resonance in LX. A general finding is that when perceivers encounter emotionally arousing stimuli, there will be greater activation in the ANS, as is measurable via increases in pupil size or skin-conductance, respectively. In participants' LX, the difference in the magnitude of physiological activation to arousing vs. neutral stimuli is typically smaller (for example, Harris, Ayçíçeǧí, & Gleason, 2003; Toivo & Scheepers, 2019), which can be interpreted as an indication of reduced emotional resonance in LX.

Caldwell-Harris and colleagues have conducted skin-conductance response studies across a number of different language combinations, modalities of stimuli, and types of bilinguals (e.g. Caldwell-Harris, Tong, Lung, & Poo, 2010; Harris, 2004; Harris et al., 2003). They have found strong support for a reduced physiological response in LX. However, the effect seemed to depend on the types of stimuli used; for example, Caldwell-Harris et al. (2010) found that participants who spoke Mandarin as their L1 responded *more strongly* to endearments in their LX. Further, late English L2 learners had stronger skin-conductance responses for childhood reprimands in their L1, but for early bilinguals, responses were comparable between L1 and L2 (Harris, 2004).

Pupillometry studies have established reduced emotional resonance using designs that involve single words (Toivo & Scheepers, 2019) and sentences (Iacozza, Costa, & Duñabeitia, 2017). Iacozza and colleagues, for example, tested Spanish (L1)-English (L2) late bilinguals and found that the difference in pupillary responses to high vs. low arousing stimuli was smaller in L2 as opposed to L1. Similarly, Toivo and Scheepers (2019) found that the difference in pupillary responses to high vs. low arousal words was smaller in participants' L2 than in L1, both for German (L1)-English (L2) and for Finnish (L1)-English (L2) late bilinguals.

Behavioural measures, on the other hand, utilise standard paradigms from cognitive psychology, such as the Lexical Decision Task (e.g. Conrad, Recio, & Jacobs, 2011; Ponari et al., 2015), different variations of the Stroop Task (e.g. Fan et al., 2016; Winskel, 2013), or other measures such as the Implicit Association Task (Segalowitz, Trofimovich, Gatbonton, & Sokolovskaya, 2008), or the Affective Simon Task (Altarriba & Basnight-Brown, 2010). These paradigms are perhaps most commonly used in studying bilingual emotions, as they are widely validated and relatively easy to administer. In the context of emotional resonance in bilinguals, such studies are built on the assumption that emotional activation is more automatic in speakers' L1. Correspondingly, reduced emotional resonance in LX would be reflected in an L1 advantage effect (faster responses to emotional stimuli in L1) or in an L2 advantage effect (faster responses to emotional stimuli in L2), depending on the type of paradigm used.

For example, testing Chinese English speakers with the Emotional Stroop Task, Fan and colleagues found that the Stroop effect was stronger in L1 as opposed to L2 (Fan et al., 2018; Fan et al., 2016). Winskel (2013) found a similar pattern in Thai speakers, where the expected interference effect was again stronger in L1. On the other hand, several studies have failed to detect differences in emotional interference/facilitation between L1 and L2 and

concluded that they did not find evidence for reduced emotional resonance in L2 (e.g. Dudschig, de la Vega, & Kaup, 2014; Eilola & Havelka, 2010; Kazanas & Altarriba, 2016; Sutton, Altarriba, Gianico, & Basnight-Brown, 2007). Interestingly, Eilola and Havelka (2010) found differences in skin-conductance responses between L1 and L2, but the interference effect was equal between L1 and L2 in a Stroop Task.

Lastly, there is also *neuro-physiological* evidence in support of reduced emotional resonance in L2. An ERP study by Wu and Thierry (2012), for example, suggested that words associated with more negative emotional valence tend to block simultaneous activation of L1 and L2. Similarly, Jonczyk and colleagues (2016) showed lower N400 amplitudes for negative valence sentences in L2 than in L1. Combining ERP with a Lexical Decision Task, other studies suggested weaker or delayed affective processing in L2 compared to L1 (Conrad, Recio, & Jacobs, 2011; Opitz & Degner, 2012). These findings suggest that there is a difference in how bilinguals process affective words, and that this may in turn affect access to such words in the mental lexicon. In an fMRI study where participants read passages of Harry Potter in L1 and L2, Hsu, Jacobs, and Conrad (2015) found stronger hemodynamic responses in the amygdala and the left pre-frontal cortex to *happy* vs. *neutral* passages, but this was only found when passages were read in L1. Again, this indicates stronger emotional involvement when reading emotional texts in L1 rather than L2.

While the different experimental paradigms – on a whole – appear to provide compelling evidence for the notion of reduced emotional resonance in bilinguals' LX, their feasibility in terms of operationalising and psychometrically quantifying this construct (e.g., to capture related inter-individual differences) is naturally somewhat limited. Many of these paradigms require specialist equipment and/or procedures that can only reasonably be applied in a laboratory setting. Moreover, while such paradigms tap into specific behavioural or physiological *manifestations* of affective processing in various languages, they are hardly

able to capture emotionality in LX across a wider range of contexts and experiences. This is where self-report methods come to their own, which we will discuss next.

1.2 Self-report methods

A key finding that emerged from studies using the BEQ or interviews was the significantly stronger emotional investment in L1 compared to LX which became manifest in a self-reported preference to express emotions in either L1 or LX, depending on emotional content (e.g., a preference for expressing endearments in L1, but a preference to talk about embarrassing topics in LX).

The emotional power of L1 swearwords was positively correlated with their frequency of use (Dewaele, 2004, 2010, 2011). One sub-group of bilingual participants that deviated from this pattern were Asian and Arab (especially female) participants, who were more likely to report that they preferred swearing in LX, to avoid the taboo of swearing in their L1. The use of LX for swearing allowed them to overcome this social constraint (Dewaele, 2013). Shakiba and Dewaele (2022) found a similar pattern among 204 Persian-English multilingual immigrants in Canada. Female participants reported swearing significantly less frequently in Persian than their male peers, but no gender difference existed for English. One female participant pointed out: "I cannot swear in Persian, I feel the heaviness of using those words and bad reaction from Persian culture to a woman who swears" (p. 15). Other studies that used the BEQ such as Resnik (2018), who collected data from 167 multilinguals, confirmed the overall preference of L1 for swearing and the superior emotional resonance of swear and taboo words in participants' L1 compared to their LX.

The closed items about language preference for swearing and for estimation of emotional resonance of L1 and LX swearwords were formulated very broadly in the BEQ. This sufficed for the identification of general patterns: "If you swear in general, what language do you

typically swear in?" (with a 5-point Likert scale ranging from "never" to "all the time") and "Do swear and taboo words in your different languages have the same emotional weight for you?" (with a 5-point Likert scale ranging from "not strong" to "very strong") (Dewaele 2013, p.230).

In order to gain a more granular view of multilinguals' swearing habits and perceptions, Dewaele (2016) made a list of 30 English negative emotion-laden words ranging from relatively mild levels of offensiveness such as "daft" and "fool" to much more offensive words and expressions such as "bitch" and "prick" which were integrated in a short sentence to provide some context. A total of 1,159 English L1 users and 1,165 English LX language users rated their understanding of each word, its offensiveness and their frequency of use on a 5-point Likert scale (Dewaele, 2016). The analysis of differences between L1 and LX users yielded some expected and some unexpected findings. For instance, L1 users reported a better understanding of the words and expressions and more frequent use of the highly offensive ones. Also, English LX users reported swearing significantly less in English than L1 users. Unexpectedly, however, LX users tended to significantly *overestimate* the offensiveness of 29 out of the 30 words and expressions. This could have been the result of foreign language teachers warning them that these were "red flag" words that could get them into trouble and that it was safer to stick to more neutral words. Lacking sufficient exposure to authentic use of the negative emotion-laden words, LX users may have struggled to differentiate between them and calibrate their affective power accurately. The expected pattern of *underestimation* of the emotional weight of LX swearwords was found only for a single word, starting with "c" and ending in "t", the most offensive word in the list. This could be due to the fact that this word is truly taboo, and as such used much less frequently by L1 users than a word like (say) "fuck", which in turn may make it harder for LX users to pick up its offensiveness (Dewaele, 2018).

Emotional resonance is not just a feature of negative emotion words. Positive emotion words and declarations of love can be particularly resonant depending on the language in which they are uttered. Dewaele (2008) found that half of the participants who filled out the BEQ felt that the words "I love you" were most powerful in their L1, around a third of participants rated them equally strong in their L1 and an LX, while the remaining participants described the expression as being stronger in their LX – the language used with their partner. Many participants admitted that even after years of using LX they still remained a little unsure about the emotional resonance of "I love you" in LX.

Caldwell-Harris, Kronrod, and Yang (2013) focussed on the Mandarin translation of "I love you" ("Wo ai ni") among 66 bilingual Chinese university students and the perception and use of "I love you" by 71 monolingual English American students. The expressions were found to be pragmatically different in both languages. Compared to the American students, the Chinese students were significantly less expansive in expressing love, and they actually preferred nonverbal expressions of love. Ożańska-Ponikwia (2017) focused on the emotional resonance of "I love you" and the equivalent Polish expression "Kocham cię" among 72 Polish-English bilinguals living in the UK and Ireland. "Kocham cię" was judged to be emotionally stronger that "I love you", but socialisation in English culture, a high self-perceived English proficiency and a high frequency of English use, were linked with a stronger emotional resonance of "I love you". Pursuing the investigation of love in an LX, Dewaele and Salomidou (2017) investigated whether expressing love to a partner in an LX caused linguistic and psychological challenges. A third of 429 participants from all over the world claimed not to have experienced any difficulty, while half mentioned lexical and conceptual limitations in the LX which they claimed had hampered their communication of emotion. An apparent lack of emotional resonance in LX was also mentioned by participants who were in an intercultural romantic relationship (Dewaele & Salomidou, 2017). Many

complained about a lack of emotional intensity when communicating with their partner in LX. Around a quarter of participants linked this to a lack of 'genuineness', at least at the start of the relationship. Most participants went through a process of emotional socialisation in LX and for a majority, their LX eventually became their 'language of the heart'.

Dewaele et al., (2021) developed a composite measure reflecting participants' self-reported emotional reactions when watching television news and films in L1 or LX. These include self-reported frequency and intensity of feeling emotional, frequency of laughter and degree of trust. Participants were 271 British English L1 users, 282 Greek and 271 Hungarian English LX users living in their home country. The study revealed an expected pattern of significantly weaker emotional reactions for English LX users compared to English L1 users. Unexpected differences emerged between the two LX user groups: the Greek-English bilinguals reported significantly stronger emotional reactions than the Hungarian-English bilinguals, despite similar levels of English (LX) proficiency. This could be linked to the quality of the foreign-language English teaching in the Greek education system, combined with a tradition of watching foreign films in the original version with subtitles rather than dubbing like in Hungary.

The reduced emotional resonance in LX may cause LX users to feel inauthentic, but it can also allow them to disclose and process traumatic experiences that would be too painful to discuss in their L1. Switching to the LX allows multilingual clients in psychotherapy to either zoom in or out depending on whether they need to distance themselves from the traumatic event they are discussing (Costa & Dewaele, 2014, 2019; Rolland, Costa, & Dewaele, 2021; Rolland, Dewaele, & Costa, 2017). Cook and Dewaele (2021) multiple case study on the language preferences of three refugees who had been tortured in their homeland because of their sexual orientation before starting therapy in the UK showed that some experiences were

just too traumatic to retell in their L1 but that the use of English (LX) enabled them "to visit their pain" as one participant put it.

1.3   The present research

As the above illustrates, quantitative studies that rely on self-reports have the potential to reach a larger number of participants with more different language combinations and cultural backgrounds than would be feasible in a laboratory experiment. Self-report methods are also able capture a wider range of emotional experiences than an experiment with a narrow focus on behavioural or physiological manifestations of affect would ever allow to measure.

To this date, however, there exists no self-report survey instrument that would demonstrably satisfy key psychometric standards for a reliable and valid measurement of reduced emotional resonance in LX. This is precisely the gap that the current study is going to address. Specifically, we aim to operationalise reduced emotional resonance in bilinguals' LX and to develop a psychometrically validated measurement tool to quantify this construct for the purpose of resolving inter-individual differences among bilinguals with diverse language combinations and cultural backgrounds.

2.   Study 1

2.1   Method

Conceptualisation of reduced emotional resonance in LX

The purpose of the scale we develop here is to measure reduced emotional resonance in the LX of bilingual speakers. The scale is aimed at speakers who use two or more languages in their daily lives, but consider their L1 to be dominant. They will be referred to as bilingual speakers here. The term LX is used to describe the "later learnt language" of these speakers. This term is chosen to reflect bilingualism as a spectrum of different language backgrounds.

This is to move away from describing bilingualism in terms of a clear "order" of languages (Dewaele, 2017). This idea allows for the concept of first and consequent languages to be more dynamic, based on the context of learning, context of current use and which language the speaker feels to be more dominant or their "main language".

Bilinguals often feel "less" when speaking in their LX. Despite being proficient and having an excellent command of the language, it may not feel the same as one's "first" language (L1) does (Pavlenko, 2006). In some cases, the LX can feel detached or even fake (Pavlenko, 2005). Bilinguals may also refer to an emotional distance in LX (Degner, Doycheva, & Wentura, 2011), as opposed to their L1, which usually feels more like the language of emotions (Dewaele, 2010, 2013).

This emotional distance may make it easier to discuss embarrassing topics (Bond & Lai, 1986) or swear in one's LX. Bilinguals may also feel that emotional words and phrases (such as the phrase "I love you" Ożańska-Ponikwia, 2017) have less weight in LX, as opposed to L1. Bilinguals may prefer to switch to their L1 when angry, so as to be able to express their anger (Dewaele, 2006). Indeed, emotional distance in L2 might actually be useful in therapy, when discussing emotional content or accessing memories (Rolland et al., 2017).

Here, we will refer to this emotional distance as *reduced emotional resonance of LX*. This term is chosen to reflect bilinguals' feeling of emotional intensity in a language, rather than expressing or experiencing different emotional states. We aim to capture bilingual speakers' perceptions of this emotional intensity, and how it might be reduced in their LX, *in comparison to their L1*. This provides flexibility in terms of which languages the researcher wants to measure and allows for the use of the scale in multilingual settings and with participants who have complex language backgrounds.

Candidate item generation

We began the process by generating candidate items. Each author contributed with items to the candidate list, and prior to the first phase of data collection, all items were reviewed and discussed with all the authors. Most candidate items were inspired by previous literature on reduced emotional resonance in LX, attempting to capture multiple modalities and situations. The pilot scale consisted of 22 candidate items. The full list of candidate items and supporting references for each can be found in Appendix 2.

Participants

Participants were recruited through social media, bilingualism/multilingualism-related mailing lists, and the subject pools at Glasgow University and Birkbeck, University of London. Participants were not paid for their participation, but 74 participants who were recruited through subject pools and completed the study as a course requirement were given course credit for their participation. In total, 1120 participants started the survey, of which 688 (61%) completed the study in full. Only the latter were included in subsequent analyses.

Of the final 688 participants, 527 (77%) identified themselves as female, 134 (19%) as male, and 17 (2.5%) as nonbinary. Ten participants (1.5%) preferred not to reveal their gender. The average reported age was 32.4 years ($SD = 13.5$ years), ranging from 18 to 80 years.

Participants were from very diverse L1 backgrounds, encompassing 87 different languages (Appendix 1a). All participants reported to actively speak between 2 and 10 different languages; 213 (31%) reported to use two languages and the remaining 475 (69%) to use more than two languages regularly. Eighty-four different languages were listed among participants' LXs (Appendix 1b).

Participants' average self-reported length of stay in the country where their LX is spoken was 10.98 years (SD = 12.2 years), ranging from 0 to 65 years.

Procedure

All data collection took place online through the Experimentum data collection platform (DeBruine, 2019). Participants were sent a link to access the study. First, participants were asked to provide demographic information (age, gender, languages acquired before the age of 3 [L1s] and languages acquired after the age of 3 [LXs], Length of stay in a country where LX is spoken, and which LX they use the most frequently). After completing the demographic questions, participants were asked to respond to all the candidate items (see Appendix 2), the order of which was randomised individually for each participant. Participants were asked to think of the L1 and LX that they use the most frequently. Each candidate item was presented with a Likert scale ranging from 1 (strongly disagree) to 6 (strongly agree), and participants were asked to use this scale to indicate, as best as they could, their degree of agreement with the statement in each given item.

2.2 Results

For the analyses reported in this section, all data and scripts are available at https://osf.io/v4k32/.

Test reliability and internal consistency

Analyses of test reliability were performed using the `alpha()` function of the R package *psych* (Revelle, 2022). By setting the `check.keys` argument to `TRUE`, the procedure automatically reverse-scored items with negative loadings on the first principal component. It emerged that the initial candidate set of 22 items (Appendix 2) already obtained a 'good' Cronbach's alpha of 0.815, with a 95% CI of [0.795, 0.832][1]. The median inter-item

---

[1] Here and in the following, CIs were determined *non-parametrically* via bootstrapping over 10,000 resamples. Only two-tailed CIs will be reported unless indicated otherwise.

correlation (MedIIC) was 0.178, with a 95% CI of [0.160, 0.205]. However, the output suggested that dropping some of the items would further improve reliability.

We therefore successively removed individual items from the scale (at each step dropping the item whose removal yielded the largest improvement in alpha) until alpha did not increase any further. Following this procedure, items Q2, Q3, Q6, Q8, Q9, Q11, and Q13 were excluded.

The resulting *final scale* includes 15 items. It obtains a Cronbach's alpha of 0.866 (95% CI = [0.849, 0.880]) and a MedIIC of 0.302 (95% CI = [0.269, 0.332]). Compared to the initial 22-item scale, the final scale significantly improves on alpha by 0.051 units (95% CI = [0.030, 0.056]) and on MedIIC by 0.124 units (95% CI = [0.089, 0.141]).

Composite test scores

Composite test scores were calculated via averaging across the final 15 items after reverse-scoring item Q7 (as suggested in the previous analyses). Higher values on the composite score indicate greater emotional resonance reduction in LX relative to L1. The 688 bilingual participants in our sample were roughly normally distributed on this metric (Figure 1), with an overall mean of 3.575 (95% CI = [3.506, 3.643]) and a standard deviation of 0.933 (95% CI = [0.887, 0.977]).

(a)

Distribution of mean composite test scores (N = 688)

(b)

Normal Q-Q Plot
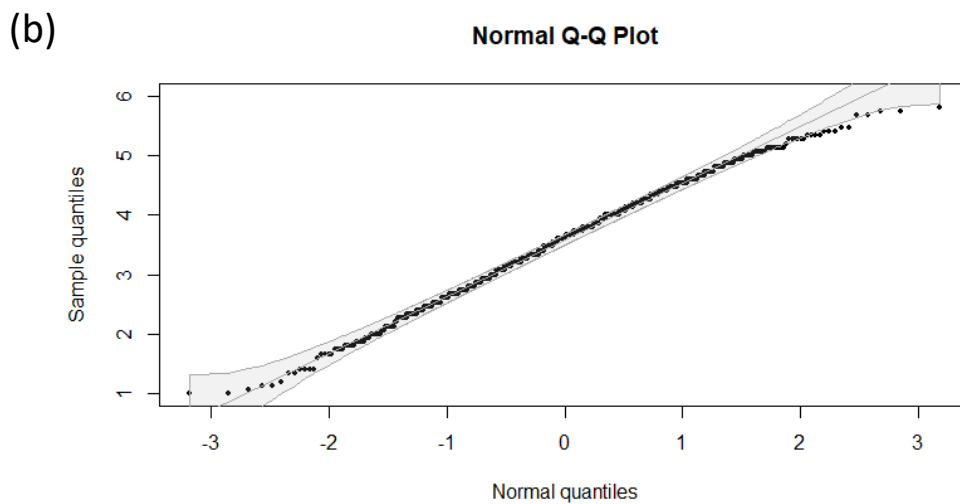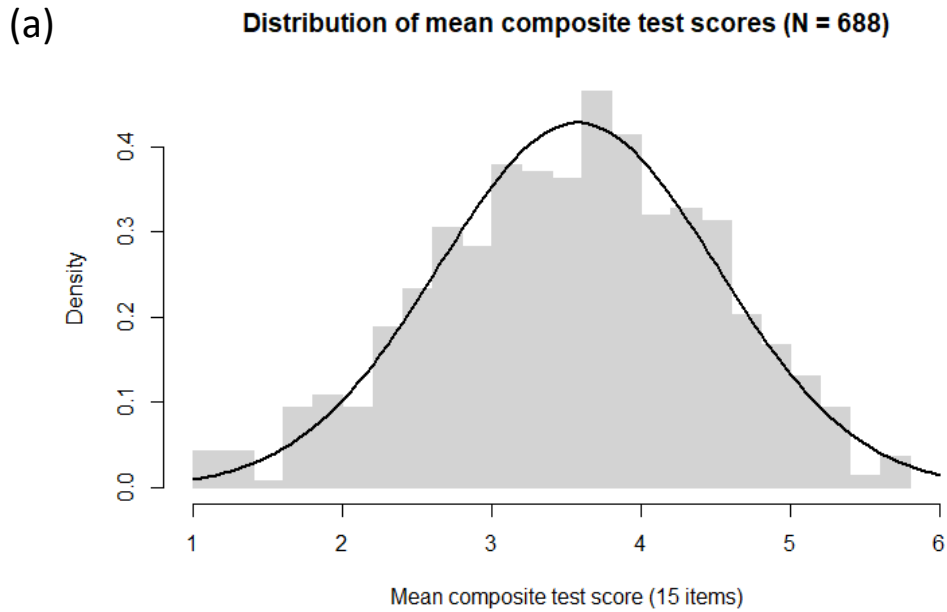
Figure 1. (a) Density distribution of the mean composite test scores from the final 15-item scale. Test scores can theoretically range from 1 to 6, in line with the 6-point Likert scales per item. Also shown is a hypothetical normal distribution curve for this value range, as predicted from the sample *mean* and *SD*. **(b)** Normal Q-Q plot for the mean composite test scores, with robustly estimated prediction line and 95% confidence bands. Only 20 of the 688 observations (3%) are not within the confidence envelope of a hypothesised normal distribution, due to somewhat under-dispersed test scores towards the upper end of the predicted quantile range.

Test reliability for smaller samples

The previous figures for Cronbach's alpha and MedIIC suggest good test reliability of the final 15-item scale, based on a sample size of 688 participants. However, in more practical terms, it is also useful to know how the scale will perform when sample sizes are much smaller – say, if one wants to measure emotional resonance reduction in LX for use as a participant-specific control variable in an experiment with only 20-60 participants.

Indeed, since sampling error is expected to increase with smaller Ns (higher likelihood of "extreme" samples) and assuming that sampling distributions for alpha and MedIIC are skewed (both measures can only take values between 0 and 1), there will likely be a *bias* in the average alphas and MedIICs for smaller Ns, combined with wider distributional spreads around those averages.

To examine this issue, and to provide some benchmarks for future applications of our scale, we performed a Monte Carlo analysis (based on bootstrapping) in which we simulated the sampling distributions for alpha and MedIIC at various 'small' settings of N, ranging from 15 to 60 participants by increments of 5 participants. Fifteen participants (same as the number of items) was considered the minimum feasible sample size for determining test reliability. For each setting of N, we took 10K resamples from our data, each time by randomly drawing (with replacement) N participants from the pool of 688 available. For each of these resamples, we then determined our test statistics of interest (Cronbach's alpha and MedIIC).

The results are summarised in Table 1. Across the 10K resamples per N, we calculated the means as well as 5% and 95% quantiles per test statistic. These quantiles represent the lower and upper bound for the central 90% of the bootstrapped sampling distribution per measure and N (or if one prefers, estimates for the upper and lower bound of the two-tailed 90% CI per measure and N).

Table 1. Bootstrap results for small(ish) participant samples ranging from N = 15 to N = 60. Shown are the means, 5% quantiles (Q .05), and 95% quantiles (Q .95) for Cronbach's alpha and median inter-item correlation (MedIIC). Figures are based on 10,000 resamples per N.

N	Cronbach's Alpha			MedIIC		
	Mean	Q .05	Q .95	Mean	Q .05	Q .95
15	0.844	0.705	0.927	0.312	0.139	0.484
20	0.851	0.737	0.922	0.311	0.162	0.460
25	0.855	0.764	0.917	0.309	0.181	0.442
30	0.857	0.778	0.913	0.308	0.191	0.427
35	0.859	0.786	0.911	0.308	0.198	0.419
40	0.859	0.791	0.909	0.306	0.204	0.410
45	0.860	0.799	0.906	0.306	0.210	0.404
50	0.861	0.803	0.905	0.306	0.214	0.399
55	0.861	0.807	0.904	0.306	0.219	0.394
60	0.862	0.810	0.902	0.305	0.222	0.389

The table shows that even with relatively small Ns, mean alphas and MedIICs stay reasonably close to their expected values of alpha = 0.866 and MedIIC = 0.302, respectively (cf. 2.2 *Test Reliability and Internal Consistency*). However, it is also evident that as N gets smaller, the mean alpha tends to deviate more negatively and the mean MedIIC more positively from the relevant expected value. These biases are due to increasing sampling error for smaller Ns combined with differently-skewed sampling distributions per test statistic (negative skew for alpha, positive skew for MedIIC).

The bootstrapping results also confirm that larger Ns lead to systematically narrower spreads in the sampling distributions for each test statistic (distances between the 5% and 95%

quantiles become consistently smaller as N increases). Again, this is predicted from the fact that larger Ns lead to a decrease in sampling error.

Given that the Q .05 figures in the table also indicate estimated lower bounds for the one-tailed 95% CIs per measure and N, we can be confident that our scale will achieve at least 'acceptable' test reliability (alpha ≥ 0.7) even with as few as only 15 participants per study. With between 45 to 50 participants, we can be 95% confident that the scale achieves at least 'good' reliability (alpha ≥ 0.8). Additional Monte Carlo runs (not shown in the table) indicated that from around 150 participants per study, confidence in alphas of 0.8 or better increases to at least 99.9%.

Dimensionality

High test reliability of the final scale reveals little about its underlying factor structure. Indeed, Cronbach's alpha must *not* be interpreted as a test of unidimensionality. We therefore also conducted an *Exploratory Factor Analysis* (EFA) over the final 15 items, using data from all 688 participants. As with the previous analyses, we employed the R package *psych* (Revelle, 2022) for this kind of modelling.

Given the ordinal nature of the original item responses (6-point "agreement" ratings), factors were determined on the basis of *polychoric* correlations using the *maximum likelihood* factoring method. To estimate how many factors are needed, we performed a *Parallel Analysis* (Horn, 1965) over 1000 iterations. In addition, we considered *Bayesian Information Criteria* (BIC and SABIC) for unrotated factor solutions. All of these indices converged on *five* as the optimum number of factors to extract.

To interpret the five extracted factors, we explored a range of different oblique rotation methods. Two of these (*simplimax* and *biquartimin*) failed to converge; the remaining ones (*promax*, *oblimin*, *bentlerQ*, *geominQ*, and *cluster*) all converged on the same factor

structure, but with minor differences in the magnitudes of the loadings and factor correlations. Figure 2 shows the *promax*-rotated solution.
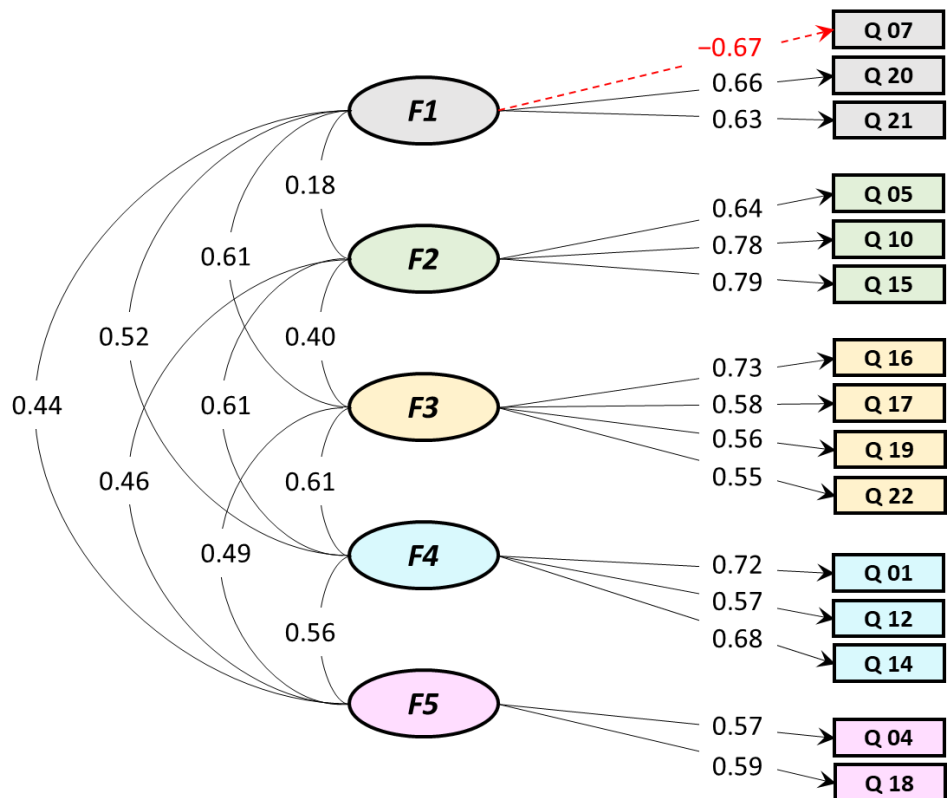
Figure 2. Polychoric EFA diagram illustrating how the 5 extracted factors (coloured ellipses on the left) load on the final 15 items (coloured rectangles on the right) after *promax* rotation. The corresponding arrows only show 'substantial' loadings with absolute values ≥ 0.3. The numbered arcs on the left represent between-factor correlations. Item Q7 was entered into the EFA using its original (non-reversed) scoring, hence the negative loading from factor *F1*. Model fit was good (RMSEA = 0.037, fit = 0.924, off-diagonal fit = 0.998), and so was factoring reliability (TLI = 0.976).

As can be seen from the figure, factor *F1* (Eigenvalue = 1.921) loads negatively on item Q7 ("I find it easier to talk about sex in my L1 than in my LX"), but positively on items Q20 ("I emotionally connect with my conversation partner better and faster in my L1 than in my

LX") and Q21 ("I have a better sense of what my conversation partner is thinking or feeling when using my L1 than my LX"), thus lending itself to a *Reduced Emotional Connection in LX* interpretation.

Factor *F2* (Eigenvalue = 1.803) is interpretable as *Reduced Vulnerability in LX*. It loads positively on items Q5 ("Swear words affect me more in my L1 than in my LX"), Q10 ("Being criticised feels more unpleasant in my L1 than in my LX"), and Q15 ("Insults feel more hurtful in my L1 than in my LX").

Factor *F3* (Eigenvalue = 1.676) loads positively on items Q16 ("I prefer my L1 over my LX when reading for pleasure"), Q17 ("A sad film is more likely to make me cry when I watch it in my L1 rather than my LX"), Q19 ("Poetry in my LX has less of an effect on me than poetry in my L1"), and Q22 ("Romantic songs feel more intense in my L1 than in my LX"). We therefore interpret it as *Reduced Emotional Engagement with Media/Art in LX*.

Factor *F4* (Eigenvalue = 1.667) loads positively on items Q1 ("I feel less emotional when using my LX than when using my L1"), Q12 ("Saying the equivalent of "I love you" has more weight in my L1 than in my LX"), and Q14 ("Compared to L1, I feel like there is more of an emotional distance when I use my LX"). Hence, we label this factor *Perceived Emotional Distance in LX*.

Finally, factor *F5* (Eigenvalue = 1.167) loads positively on items Q4 ("When I'm really angry, I tend to use my L1 more than my LX") and Q18 ("When I want swear words to have real weight, I use my L1 rather than my LX"), therefore lending itself to a *Reduced Offensiveness in LX* interpretation.

Very importantly, all between-factor correlations (numbered arcs in Figure 2) are significantly positive (Bonferroni-corrected ps < .05) and the majority of them quite substantial (≥ 0.3) – a pattern that is consistent across all the oblique rotation methods

considered (*promax*, *oblimin*, *bentlerQ*, *geominQ*, and *cluster*). This suggests that the five extracted factors represent more detailed (but related) facets of a single underlying meta-factor, which we may call *Reduced Emotional Resonance in LX*. Indeed, additional parallel analyses using factor scores from the different oblique rotation methods as input vectors consistently suggested only a single principal component for the five extracted factors. Our scale therefore measures *one* general construct which is composed of five interrelated sub-domains (or *facets*) that are more specific.

3.   Study 2

The second study was designed to assess the validity of our final 15-item scale, which we will henceforth refer to as *RER-LX* (for *Reduced Emotional Resonance in LX*). Specifically, we aim to evaluate its *discriminant validity* in relation to LexTALE (a measure of linguistic proficiency) as well as its *convergent validity* in relation to an item set that was used to measure reduced emotional resonance in LX in previous research.

3.1   Method

Materials

<u>Reduced Emotional Resonance in LX (RER-LX).</u> The final version of our RER-LX scale included 15 items, as shown in Appendix 3. Composite scores (for correlation with other measures) were computed as described in section 2.2 *Composite Test Scores*.

<u>Proficiency in LX.</u> The brief LexTALE test (Lemhöfer & Broersma, 2012) was used to assess participants' proficiency in English (the target LX in this study). The LexTALE is a 60-item lexical decision test, which takes about 5 minutes to complete. Participants are asked to distinguish real English words from English-looking words that do not exist in English. Following the authors' suggestions, the test scores from this scale were calculated as the percentage of correct responses, adjusted for unequal proportions of words and non-words in

the test. In our final sample of participants, the mean score was 85.75% (SD = 11.20%), ranging from 35% to 100% (fourteen participants in our sample, i.e. 6%, achieved the top proficiency score).

BEQ sub-scale. There are no existing scales which directly measure reduced emotional resonance in LX and return a single score. To test the convergent validity of RER-LX, three questions from the Bilingualism and Emotions Questionnaire (BEQ, Dewaele & Pavlenko, 2001-2003) were chosen for comparison, as they had been used as a proxy for emotional resonance in LX in previous research (Dewaele, 2013, Pavlenko, 2005).

1. BEQ-swearing: *If you swear in general, what language do you typically swear in?*
For this question, participants were asked to estimate how often they swear, both in their L1 and in their LX, on a scale from 1 (never) to 5 (all the time).

2. BEQ-feelings: *What language do you express your deepest feelings in?*
Here, participants were asked to estimate how often they express their feelings, both in L1 and in LX, on a scale from 1 (never) to 5 (all the time) across four situations (when alone; with parents/partner; when talking to friends; in letters and emails). Their responses to the four situations were summed up when calculating the score for BEQ-feelings.

3. BEQ-anger: *If you are angry, what language do you typically use to express your anger?*
For this question, participants were asked to estimate how often they express their anger, again both in L1 and in LX, on a scale from 1 (never) to 5 (all the time) across five situations (when alone; with parents/partner; when talking to friends; in letters and emails; when talking to strangers). Their responses to the five situations were summed up when calculating the score for BEQ-anger.

For each of the three questions, we then calculated the difference between L1 and LX, with a higher difference score reflecting a stronger preference for L1 in the given item (interpretable as reduced emotional resonance in LX relative to L1).

Since the ratings for BEQ-feelings were summed up across four situations, we divided the relevant difference scores by 4; likewise, difference scores for BEQ-anger were divided by 5, in line with the five situations of assessment for this question. This way, we ensured that the three BEQ items contributed equally to the mean composite test scores for the BEQ subscale.

Finally, a small number of missing data points (6 for BEQ-swearing, 9 for BEQ-feelings, and 2 for BEQ-anger) were conservatively replaced with the relevant item means.

Participants

As before, participants were recruited through social media and the respective subject pools at the authors' universities. In our advertisements, we specifically targeted L2 speakers of English. A total of 414 participants started the study, out of which 247 participants (60%) completed the study in full and were included in subsequent analyses. This sample size should give us sufficiently stable correlation estimates (see in particular Schönbrodt & Perugini, 2013).

All participants were LX speakers of English. Participants were not paid, but 103 participants were given course credit for their participation.

Participants' mean age was 29.64 years (*SD* = 12.91 years), ranging from 17 to 86 years. One-hundred-forty-four (58%) identified themselves as female, 43 (17%) as male, 5 (2%) as non-binary, and 55 (22%) preferred not to reveal their gender identity.

The average self-reported age at which participants acquired LX (English) was 7.33 years (SD = 4.44 years), ranging from 0 to 40 years. The average self-reported length of stay in an English-speaking country was 8.54 years (SD = 9.14 years), ranging from 0 to 45 years.

There were 52 different languages among the spoken L1s reported by the participants in our sample (Appendix 1c).

Participants reported speaking between one and eight different LXs. The most commonly reported number of LXs spoken was two (80 participants) or three (80 participants).

Procedure

All data collection took place online through Experimentum (DeBruine et al., 2020). Participants were sent a link to access the study. First, participants were asked to provide demographic information (Age, Gender, L1, length of stay in an English-speaking country, age of acquisition for English). After completing the demographic questions, participants were asked to complete the LexTALE proficiency test, after which they completed a brief distractor task. Then they were asked to respond to all RER-LX scale items (presented in an individually determined random order), and finally to the three BEQ items. Each of the measures included in the study appeared on a separate page. Participants were instructed to think of English as their LX when completing RER-LX and BEQ.

3.2   Results

For the analyses reported in this section, all data and scripts are available at https://osf.io/v4k32/.

Test reliabilities for RER-LX and BEQ

For the new sample of participants (N = 247), our 15-item RER-LX scale obtained a Cronbach's alpha of 0.867 (95% CI = [0.837, 0.890]) and a MedIIC of 0.294 (95% CI = [0.253, 0.357]). These figures closely replicate those from Study 1.

The 3-item BEQ subscale obtained a Cronbach's alpha of 0.850 (95% CI = [0.805, 0.885]) and a MedIIC of 0.604 (95% CI = [0.523, 0.698]).

With a difference in alpha of only 0.017 (95% CI = [−0.023, 0.061]), test reliabilities for RER-LX and BEQ are roughly the same. This is in spite of considerably lower inter-item correlations for RER-LX than for the BEQ subscale (difference in MedIIC = −0.310, 95% CI = [−0.398, −0.214]). The much larger number of (more unique) items in RER-LX obviously compensates for the latter (cf. de Vet et al., 2017). Also note that the high MedIIC for BEQ (> 0.5) may actually indicate some undesirable redundancy among the three BEQ items.

Convergent and discriminant validity

To establish *convergent validity*, we correlated the mean composite scores from RER-LX with the mean composite scores from the BEQ subscale. Since the two scales are designed to measure the same construct and are scored in the same direction (*higher* meaning *more emotional resonance reduction in LX relative to L1*), we expected a positive correlation.

To establish *discriminant validity*, we correlated the RER-LX and BEQ measures with the scores from LexTALE, which is designed to measure linguistic proficiency in LX (in this case, English). To exhibit discriminant validity, we expected correlations with LexTALE to be around zero. We provide further theoretical motivation for this in the General Discussion.

Table 2 shows the inter-correlations between the three measures (NB, histograms for each measure suggested no obvious outliers), together with bootstrapped 95% confidence intervals.

Table 2. Inter-correlations between RER-LX, BEQ, and LexTALE, with two-tailed 95% CIs in square brackets. The latter were determined via bootstrapping over 10,000 resamples.

	BEQ	LexTALE
RER-LX	+0.453 [+0.345, +0.549]	−0.139 [−0.274, +0.001]
BEQ		−0.318 [−0.417, −0.207]

As becomes evident, RER-LX clearly exhibits convergent validity in relation to BEQ – the relevant correlation is positive, appreciable in size, and significant (the lower bound of the associated 95% CI is markedly above zero). At the same time, RER-LX is only weakly negatively related to LexTALE and the corresponding 95% CI encloses zero (meaning this correlation is not significant). In comparison, the BEQ subscale exhibits a more substantial negative correlation with LexTALE, which is indeed significant (the upper bound of the associated 95% CI is clearly below zero). A direct comparison between the two negative correlations confirms that the relation between RER-LX and LexTALE is significantly less negative (i.e., closer to zero) than the relation between BEQ and LexTALE (difference = 0.179, 95% CI = [0.053, 0.302]). RER-LX therefore yields better discriminant validity in relation to LexTALE than the BEQ subscale.

To further corroborate this point, we conducted a Principal Component Analysis across the three measures. We extracted two orthogonal components which together accounted for 83% of the total variance. Considering the loadings after Varimax rotation (Table 3 below), the

first principal component (PC1, Eigenvalue = 1.414) clearly captures emotional resonance reduction in LX, as it receives strong positive loadings from both RER-LX and BEQ, but only a weak negative loading from LexTALE. The second component (PC2, Eigenvalue = 1.081) mostly reflects proficiency in LX, with a very strong positive loading from LexTALE, a weaker (but considerable) negative loading from BEQ, and only a very small positive loading from RER-LX. Again, this pattern supports the notion of RER-LX having better discriminant validity than the BEQ subscale. Loading on both PCs, the BEQ subscale obviously captures aspects of both emotional resonance and proficiency in LX. RER-LX, by contrast, is more orthogonal to LexTALE and therefore represents a 'purer' measure of emotional resonance reduction in LX than the BEQ subscale.

Table 3. Varimax loading matrix for the first two principal components (PC1 and PC2) extracted from RER-LX, BEQ, and LexTALE. Substantial loadings (absolute values ≥ 0.3) are highlighted with double asterisks.

	PC1	PC2
RER-LX	+0.906 **	+0.045
BEQ	+0.764 **	−0.370 **
LexTALE	−0.099	+0.970 **

4.   General Discussion

4.1   Summary

In this paper, we developed a psychometrically validated measurement instrument for the purpose of operationalising and quantifying emotional resonance reduction in bilinguals' LX: the RER-LX scale. While there is an abundance of cross-disciplinary research studying

bilingual emotions, a scale that explicitly quantifies how much bilinguals "feel less" in languages they acquired later in life has largely been missing in the field – a gap that the current paper tries to close.

As discussed in the introduction, there is ample support for the notion of reduced emotional resonance in bilinguals' LX from a variety of cognitive-behavioural, psycho-physiological, and self-report paradigms. However, there is as yet no validated psychometric scale to quantify the concept directly. To this date, the Bilingualism and Emotions Questionnaire (BEQ, Pavlenko & Dewaele, 2001-2003) is the best and largest-scale attempt at quantifying and trying to understand the underlying reasons for reduced emotional resonance – however, the BEQ is too long to be used concurrently with other measures as a whole. Moreover, the complete BEQ instrument is less specific than RER-LX, includes multiple question types, does not come with explicit scoring instructions, and lacks validation. We have attempted to build on the questions captured in the BEQ and expand the measurement to include other situations and modalities, which have been established since the original BEQ project.

In the two online studies reported here (N = 688 and N = 247, respectively), we established our 15-item RER-LX scale to be a reliable and valid measure of emotional resonance reduction in LX. RER-LX achieves very good test reliability on large samples (alphas close to 0.87 in both studies). Even with smaller samples that are more typical for experimental research, alpha will likely stay well above 0.7 (see 2.2 *Test Reliability for Smaller Samples*). Composite test scores from the scale are roughly normally distributed (see 2.2 *Composite Test Scores*), which is desirable from a normative measurement perspective and makes the RER-LX scores reasonably suitable for parametric testing (t-test, ANOVA, linear regression, etc.).

RER-LX exhibits convergent validity in relation to the 3-item BEQ-subscale that has previously been used as a proxy for the same construct. Importantly, RER-LX improves on the latter by having significantly better discriminant validity in relation to LexTALE, a measure of proficiency in LX (see 3.2 *Convergent and Discriminant Validity*). We will further elaborate on this point in section 4.2 below.

Lastly, the scale exhibits content validity by covering five related facets of emotional resonance in LX that seem to relate well with previous approaches to conceptualizing and measuring the construct: *Reduced Emotional Connection in LX*, *Reduced Vulnerability in LX*, *Reduced Emotional Engagement with Media/Art in LX*, *Perceived Emotional Distance in LX*, and *Reduced Offensiveness in LX* (see 2.2 *Dimensionality* and references in Appendix 2).

4.2   Proficiency and emotional resonance

In terms of *discriminant validity*, one may ask why we consider independence of proficiency a desirable property of a scale measuring emotional resonance reduction in LX. Indeed, the debate surrounding the relationship between proficiency and emotional resonance in LX is related to that between language and culture. Both constructs are entangled but are different types of skills and knowledge. Although they tend to co-occur, bilingualism and biculturalism are not isomorphous (Grosjean, 2014). The capacity to speak an LX to some degree does not *necessarily* imply that the user possesses sociocultural and socio-pragmatic competence to the same degree. For example, foreign language learners who study in a school context will become increasingly proficient thanks to the linguistic input of the teacher and the learning materials, but they will only start developing sociocultural and socio-pragmatic competence through contact with the users of the language. In other words, by observing and participating in authentic interactions, LX users start to understand how the choice of specific words combined with volume, pitch, intonation, facial expression and body language contribute to

the impact of emotion words. This will trigger a process of LX socialisation that could last a lifetime. An early start of that process is linked to both higher proficiency and higher emotional resonance in the LX. The degree of LX socialisation and the size of the LX networks were found to be strong predictors of LX emotional resonance in Dewaele (2013). LX users who had acquired their LX naturalistically and from a young age reported that the LX had higher levels of emotional resonance than those who had learnt it only in an instructed context and/or later in life.

Proficiency and emotional resonance in LX are therefore conceptually independent dimensions. Indeed, Dewaele et al. (2021) found a non-significant relationship ($r = 0.14$, $p = 0.057$) between self-reported emotional reactions and English proficiency in a group adult British L1 users of English. The picture was different for proficient English LX users, where a small but significant positive correlation was found ($r = 0.20$). However, much stronger correlations were found between self-reported emotional reactions and both frequency of English use ($r = 0.40$) and frequency of watching television in English ($r = 0.41$) (p. 353). In conclusion, high proficiency does not *automatically* imply greater emotional resonance in LX, as the relationship between the two appears to depend on frequency of use and cultural exposure to LX, among other factors. We therefore deem it advantageous if a scale of emotional resonance in LX remains in principle orthogonal to proficiency in LX, the latter of which can be assessed separately.

4.3   Limitations and future uses of RER-LX

Being a self-report scale, RER-LX may suffer from some of the problems that are associated with such scales in general. For example, participants not only need to understand the questions, but also have sufficient motivation and introspective ability to answer them truthfully and accurately. There may also be cross-cultural differences in the social

acceptability of certain questions (or the social desirability of certain answers, respectively) which in turn may bias participants' responses. These issues clearly require additional empirical scrutiny in the future. All we can say at present is that, across the two independent studies reported here (each with participants from very diverse linguistic and cultural backgrounds), internal consistency of RER-LX stayed largely the same.

It is also worth noting that he present research was carried out under the restrictions imposed by the COVID-19 pandemic, which meant that we had to limit ourselves to the use of online questionnaires to evaluate the reliability and validity of the RER-LX scale. Future research would also need to further assess the scale's *predictive* usefulness, specifically in terms of how well variation in bilinguals' RER-LX scores can predict variation in their cognitive-behavioural or (neuro-) physiological responses to affective stimuli. Indeed, our expectation is that bilinguals who score higher on RER-LX should exhibit a larger discrepancy between L1 and LX when processing emotional stimuli in those languages. Given the variety of tasks, designs and measures used in experimental research on the topic (see introduction and the review by Toivo & Scheepers, 2022) it seems unlikely that such an assessment of the predictive validity of RER-LX can be achieved within a single 'definitive' study. Hence, we encourage interested members of the research community to make extensive use of RER-LX in their own future research on questions related to emotional resonance reduction in bilinguals' LX. Over time, this should yield a more comprehensive sense of what the scale can and cannot predict, and in turn, may have useful implications for further theory development and refinement in this area.

4.4   Conclusion

Compared to previous psychometric attempts at measuring the same construct, RER-LX should allow for more reliable and accurate testing of theories about the underlying reasons

for reduced emotional resonance in LX, and how it may dynamically develop over time as bilinguals may experience changes in daily exposure to and/or usage of their languages. The scale was intentionally designed to be used alongside other measures, to enable the testing of complex research questions about bilingual emotions; with 15 items (Appendix 3), the scale is reasonably brief, and it provides a single score to give an indication of participants' emotional resonance reduction in LX relative to their L1. A higher composite score on the scale suggests that the participant feels a stronger emotional discrepancy between their L1 and LX, in the direction of reduced emotional resonance in LX.

The scale instructions can be modified to reflect participants' L1 and a specific LX (see, e.g., Study 2 where all participants were asked to respond by comparing their L1 with English as their LX), or to allow for more flexibility by having participants to compare their L1 with an LX they determine themselves (e.g., Study 1 where participants were prompted to think of their L1 and the LX they use *most frequently).* This not only accounts for specific research questions, which compare two fixed languages, but also is suitable for larger-scale studies with participants from a variety of different language backgrounds and supports the idea of studying bilingualism as a dynamic concept.

On a final note, it is important to keep in mind that the fifteen RER-LX items (Appendix 3) always encourage participants to make a *comparison* between L1 and LX. This means that the scale does not measure emotional resonance in LX in an absolute sense, but rather *in relation to L1*. For any given participant, the overall RER-LX test score is therefore to be interpreted as the amount of emotional resonance reduction in their LX relative to their L1.

To conclude, use of the RER-LX scale is recommended for any study examining the emotional resonance of the languages of bi- and multilinguals. We hope that this new scale will facilitate the investigation of emotional resonance in psychology and applied linguistics.

Appendix 1

Appendix 1a. The 87 L1s reported by participants in Study 1:

Afrikaans, Arabic, Armenian, Asante Twi, Bahasa Indonesia, Bahasa Melayu, Balinese, Basque, Bengali, Bosnian, Brunei, Malay, Bulgarian, Cantonese, Catalán, Common Moroccan language, Croatian, Czech, Danish, Dhivehi, Dutch, English, Estonian, Fante, Faroese, Farsi, Filipino, Finnish, French, Frisian, Galician, German, Greek, Hakka, Haya, Hebrew, Hindi, Hokkien, Hungarian, Indonesian, Italian, Japanese, Java, Komi, Korean, Kurdish, Latvian, Lithuanian, Louisiana Creole, Makassarese, Malayalam, Malay, Maltese, Mandarin, Mauritian creole, Mongolian, Moroccan Darija, Norwegian, Odia, Paltienski, Persian, Pidgin English, Polish, Portuguese, Punjabi, Romanian, Russian, Sarawak, Scottish Gaelic, Serbian, Silesian, Sinhala, Slovak, Slovene, Spanish, Sundanese, Suzhou dialect, Swahili, Swedish, Swiss German, Tamil, Teochew, Tswana, Turkish, Turkmen, Ukrainian, Urdu, Valencian, Vietnamese, and *Welsh.*

Appendix 1b. The 84 LXs reported by participants in Study 1:

Afrikaans, American Sign Language, Arabic, Bahasa Indonesia, Bahasa Malaysia, Basque, Bengali, British Sign Language, Bulgarian, Cantonese, Castilian, Catalan, Catalan sign Language, Chilean Sign Language, Creole, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Farsi, Finnish, French, Gaelic, Galician, German, German Sign Sanguage, Greek, Haya, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Java, Kannada, Kiswahili, Korean, , Latvian, Lithuanian, Lule Sami, Malay, Mandarin, Minang, Nepali, Norwegian, Occitan, Persian, Polish, Portuguese, Punjabi, Putonghua, Romanian, Russian, Russian Sign Language, Sepedi, Serbian, Sign Language of the Netherlands, Sindhi, Sinhala, Slovak, South Sotho, Spanish, Spanish Sign Language, Standard Malay, Sundanese, Swahili, Swedish, Tamil, Telugu, Thai, Tsonga, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Zaza, and *Zulu.*

Appendix 1b. The 52 L1s reported by participants in Study 2:

Albanian, Arabic, Bangla, Bengali, Bulgarian, Cantonese, Catalan, Chinese, Croatian, Danish, Dutch, Farsi, Finnish, French, Scottish Gaelic, German, Greek, Gujarati, Hindi, Hungarian, Indonesian, Italian, Japanese, Khmer, Korean, Latvian, Lebanese, Lithuanian, Luganda, Malay, Malayalam, Mandarin, Mongolian, Nepali, Persian, Piedmontese. Polish, Portuguese, Punjabi, Romanian, Russian, Serbian, Shona, Slovak, Somali, Spanish, Tamil, Thai, Turkish, Ukrainian, Urdu, and *Welsh.*

Appendix 2

The initial 22 candidate items for Study 1, with cited sources of inspiration. Each item came with a 6-point Likert scale on which participants had to indicate their degree of agreement (1 = *strongly disagree* to 6 = *strongly agree*) with the relevant item statement.

ID	Candidate item	Inspired by (Citation)
Q1	I feel less emotional when using my LX than when using my L1	Pavlenko (2005)
Q2	I find it easier to talk about embarrassing topics in my LX than in my L1	Bond and Lai (1986)
Q3	I find it harder to talk about awkward topics in my L1 than in my LX	Bond and Lai (1986)
Q4	When I'm really angry, I tend to use my L1 more than my LX	Dewaele (2006)
Q5	Swear words affect me more in my L1 than in my LX	Resnik (2018); Dewaele (2013)
Q6	I feel less inhibited about using swear words in my LX than in my L1	Discussion among authors
Q7	I find it easier to talk about sex in my L1 than in my LX	Bond and Lai (1986)
Q8	I find it harder to talk about emotional topics in my L1 than in my LX	Dewaele & Salomidou (2017)
Q9	I find it easier to express endearment in my LX than in my L1	Caldwell-Harris et al., (2010)
Q10	Being criticised feels more unpleasant in my L1 than in my LX	Gao, Luo, and Gou (2020)
Q11	Taboo words feel worse in my LX than in my L1	Resnik (2018); Dewaele (2013)
Q12	Saying the equivalent of "I love you" has more weight in my L1 than in my LX	Ozańska-Ponikwia (2017); Dewaele (2008)
Q13	My LX feels more emotional than my L1	Pavlenko (2005)
Q14	Compared to L1, I feel like there is more of an emotional distance when I use my LX	Degner et al. (2011)

Q15	Insults feel more hurtful in my L1 than in my LX	Caldwell-Harris et al., (2010)
Q16	I prefer my L1 over my LX when reading for pleasure	Hsu, Jacobs, and Conrad (2015)
Q17	A sad film is more likely to make me cry when I watch it in my L1 rather than my LX	Dewaele (2021)
Q18	When I want swear words to have real weight, I use my L1 rather than my LX	Dewaele (2006)
Q19	Poetry in my LX has less of an effect on me than poetry in my L1	Hsu, Jacobs, and Conrad (2015)
Q20	I emotionally connect with my conversation partner better and faster in my L1 than in my LX	Pavlenko (2005) - emotional detachment; Dewaele & Salomidou (2017)
Q21	I have a better sense of what my conversation partner is thinking or feeling when using my L1 than my LX	Dewaele & Salomidou (2017)
Q22	Romantic songs feel more intense in my L1 than in my LX	Ozańska-Ponikwia (2017); Dewaele (2008)

Appendix 3

The final 15 RER-LX items. Each item comes with a 6-point Likert scale on which participants have to indicate their degree of agreement (1 = *strongly disagree* to 6 = *strongly agree*) with the relevant item statement. Ratings for item Q4 (Q7 from the pilot scale) need to be reverse scored (R). Z-standardisation of items is not strictly necessary – raw and standardized alphas for the scale were virtually the same in all analyses, small Ns included.

Original ID	Final ID	Item
Q1	**Q1**	I feel less emotional when using my LX than when using my L1
Q4	**Q2**	When I'm really angry, I tend to use my L1 more than my LX
Q5	**Q3**	Swear words affect me more in my L1 than in my LX
Q7 (R)	**Q4 (R)**	I find it easier to talk about sex in my L1 than in my LX
Q10	**Q5**	Being criticised feels more unpleasant in my L1 than in my LX
Q12	**Q6**	Saying the equivalent of "I love you" has more weight in my L1 than in my LX
Q14	**Q7**	Compared to L1, I feel like there is more of an emotional distance when I use my LX
Q15	**Q8**	Insults feel more hurtful in my L1 than in my LX
Q16	**Q9**	I prefer my L1 over my LX when reading for pleasure
Q17	**Q10**	A sad film is more likely to make me cry when I watch it in my L1 rather than my LX
Q18	**Q11**	When I want swear words to have real weight, I use my L1 rather than my LX
Q19	**Q12**	Poetry in my LX has less of an effect on me than poetry in my L1
Q20	**Q13**	I emotionally connect with my conversation partner better and faster in my L1 than in my LX
Q21	**Q14**	I have a better sense of what my conversation partner is thinking or feeling when using my L1 than my LX
Q22	**Q15**	Romantic songs feel more intense in my L1 than in my LX

Note: In both of our studies, order of items was randomly determined for each participant.

References

Altarriba, J., & Basnight-Brown, D. M. (2010). The representation of emotion vs. emotion-laden words in English and Spanish in the Affective Simon Task. *International Journal of Bilingualism, 15*(3), 310-328. doi:10.1177/1367006910379261

Bond, M., & Lai, T. (1986). Embarrassment and Code-Switching into a Second Language. *The Journal of Social Psychology, 126*(2), 179-186.

Caldwell-Harris, C., Kronrod, A., & Yang, J. (2013). Do more, say less: Saying "I love you" in Chinese and American cultures. *Intercultural Pragmatics, 10*(1). doi:10.1515/ip-2013-0002

Caldwell-Harris, C. L., Tong, J., Lung, W., & Poo, S. (2010). Physiological reactivity to emotional phrases in Mandarin—English bilinguals. *International Journal of Bilingualism, 15*(3), 329-352. doi:10.1177/1367006910379262

Conrad, M., Recio, G., & Jacobs, A. M. (2011). The Time Course of Emotion Effects in First and Second Language Processing: A Cross Cultural ERP Study with German-Spanish Bilinguals. *Frontiers in Psychology, 2*, 1-16. doi: 10.3389/fpsyg.2011.00351

Cook, S. R., & Dewaele, J.-M. (2021). 'The English language enables me to visit my pain'. Exploring experiences of using a later-learned language in the healing journey of survivors of sexuality persecution. *International Journal of Bilingualism, 26*(2), 125-139. doi: 10.1177/13670069211033032

Costa, B., & Dewaele, J.-M. (2014). Psychotherapy across languages: beliefs, attitudes and practices of monolingual and multilingual therapists with their multilingual patients. *Counselling and Psychotherapy Research, 14*(3), 235-244. doi:10.1080/14733145.2013.838338

Costa, B., & Dewaele, J.-M. (2019). The talking cure—building the core skills and the confidence of counsellors and psychotherapists to work effectively with multilingual

patients through training and supervision. *Counselling and Psychotherapy Research, 19*(3), 231-240. doi: 10.1002/capr.12187

DeBruine, L., Lai, R., Jones, B., Abdullah, R., & Mahrholz, G (2020). Experimentum (Version v.0.2). Zenodo. doi:10.5281/zenodo.2634355

Degner, J., Doycheva, C., & Wentura, D. (2011). It matters how much you talk: On the automaticity of affective connotations of first and second language words. *Bilingualism: Language and Cognition, 15*(01), 181-189. doi:10.1017/s1366728911000095

de Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman-Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85, 45-49.

Dewaele, J.-M. (2004). The Emotional Force of Swearwords and Taboo Words in the Speech of Multilinguals. *Journal of Multilingual and Multicultural Development, 25*(2-3), 204-222. doi:10.1080/01434630408666529

Dewaele, J.-M. (2008). The emotional weight of I love you in multilinguals' languages. *Journal of Pragmatics, 40*(10), 1753-1780. doi:10.1016/j.pragma.2008.03.002

Dewaele, J.-M. (2010). *Emotions in multiple languages*. Basignstoke: Palgrave–MacMillan.

Dewaele, J.-M. (2011). Self-reported use and perception of the L1 and L2 among maximally proficient bi- and multilinguals: a quantitative and qualitative investigation. *2011*(208), 25-51. doi:doi:10.1515/ijsl.2011.011

Dewaele, J.-M. (2013). *Emotions in Multiple Languages* (2 ed.). Basingstoke: Palgrave Macmillan.

Dewaele, J.-M. (2016). Thirty shades of offensiveness: L1 and LX English users' understanding, perception and self-reported use of negative emotion-laden words. *Journal of Pragmatics, 94*, 112-127. doi:10.1016/j.pragma.2016.01.009

Dewaele, J.-M., Lorette, P., Rolland, L., & Mavrou, I. (2021). Differences in emotional

reactions of Greek, Hungarian, and British users of English when watching television

in English. *International Journal of Applied Linguistics, 31*(3), 345-361. doi:

10.1111/ijal.12333

Dewaele, J.-M., & Salomidou, L. (2017). Loving a partner in a Foreign Language. *Journal of*

Pragmatics, 108, 116-130. doi: 10.1016/j.pragma.2016.12.009

Dewaele, J. M. (2006). Expressing anger in multiple languages. In A. Pavlenko (Ed.),

Bilingual minds: Emotional experience, expression, and representation. Clevedon:

Multilingual matters.

Dewaele, J. M. (2017). Why the dichotomy 'L1 versus LX user'is better than 'native versus

non-native speaker'. *Applied Linguistics, 39*(2), 236-240.

Dewaele, J. M. (2018). "Cunt": On the perception and handling of verbal dynamite by L1 and

LX users of English. . *Multilingua, 37*(1), 53-81.

Dewaele, J. M., & Pavlenko, A. (2001-2003). Web questionnaire Bilingualism and Emotions.

University of London.

Dudschig, C., de la Vega, I., & Kaup, B. (2014). Embodiment and second-language:

automatic activation of motor responses during processing spatially associated L2

words and emotion L2 words in a vertical Stroop paradigm. *Brain & Language, 132*,

14-21. doi:10.1016/j.bandl.2014.02.002

Eilola, T. M., & Havelka, J. (2010). Behavioural and physiological responses to the

emotional and taboo Stroop tasks in native and non-native speakers of English.

International Journal of Bilingualism, 15(3), 353-369.

doi:10.1177/1367006910379263

Fan, L., Xu, Q., Wang, X., Xu, F., Yang, Y., & Lu, Z. (2018). The automatic activation of

emotion words measured using the emotional face-word Stroop task in late Chinese-

English bilinguals. *Cognition and Emotion, 32*(2), 315-324.

doi:10.1080/02699931.2017.1303451

Fan, L., Xu, Q., Wang, X., Zhang, F., Yang, Y., & Liu, X. (2016). Neural Correlates of Task-Irrelevant First and Second Language Emotion Words–Evidence from the Emotional Face–Word Stroop Task. *Frontiers in Psychology, 7*, 1672.

Gao, S., Luo, L., & Gou, T. (2020). Criticism in a foreign language hurts less. *Cognition and Emotion, 34*(4), 822-830. doi:10.1080/02699931.2019.1668751

Grosjean, F. (2014). Bicultural bilinguals. *International Journal of Bilingualism, 19*(5), 572-586. doi:10.1177/1367006914526297

Harris, C. L. (2004). Bilingual Speakers in the Lab: Psychophysiological Measures of Emotional Reactivity. *Journal of Multilingual and Multicultural Development, 25*(2-3), 223-247. doi:10.1080/01434630408666530

Harris, C. L., Ayçíçeğí, A., & Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics, 24*(4), 561-579.

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* (pp. 257-283): Clevedon, Multilingual Matters.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30 (2), 179–185. doi:10.1007/bf02289447

Hsu, C. T., Jacobs, A. M., & Conrad, M. (2015). Can Harry Potter still put a spell on us in a second language? An fMRI study on reading emotion-laden literature in late bilinguals. *Cortex, 63*, 282-295.

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), e0186027. doi:10.1371/journal.pone.0186027

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

Kazanas, S. A., & Altarriba, J. (2016). Emotion Word Processing: Effects of Word Type and Valence in Spanish-English Bilinguals. *Joural of Psycholinguistic Research, 45*(2), 395-406. doi:10.1007/s10936-015-9357-3

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*(2), 325-343. doi:10.3758/s13428-011-0146-0

Opitz, B., & Degner J. (2012). Emotionality in a second language: it's a matter of time. *Neuropsychologia*, 50(8):1961-7. doi: 10.1016/j.neuropsychologia.2012.04.021.

Ożańska-Ponikwia, K. (2017). Expression and perception of emotions by Polish–English bilinguals I love you vs. Kocham Cię. *International Journal of Bilingual Education and Bilingualism*, 1-12. doi:10.1080/13670050.2016.1270893

Pavlenko, A. (2005). *Emotions and multilingualism*. Cambridge, UK: Cambridge University Press. .

Pavlenko, A. (2012). Affective processing in bilingual speakers: disembodied cognition? *International Journal of Psychology, 47*(6), 405-428. doi:10.1080/00207594.2012.743665

Ponari, M., Rodríguez-Cuadrado, S., Vinson, D., Fox, N., Costa, A., & Vigliocco, G. (2015). Processing advantage for emotional words in bilingual speakers. *Emotion, 15*(5), 644-652.

Resnik, P. A. (2018). *Differences in Feeling – Feeling the Difference: Multilinguals' Verbalisation and Perception of Emotions.* Bristol: Multilingual Matters.

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.2.5, https://CRAN.R-project.org/package=psych.

Rolland, L., Costa, B., & Dewaele, J.-M. (2021). Negotiating the language(s) for psychotherapy talk: A mixed methods study from the perspective of multilingual clients. *Counselling and Psychotherapy Research, 21*(1), 107-117. doi: 10.1002/capr.12369

Rolland, L., Dewaele, J.-M., & Costa, B. (2017). Multilingualism and psychotherapy: exploring multilingual clients' experiences of language practices in psychotherapy. *International Journal of Multilingualism, 14*(1), 69-85. doi:10.1080/14790718.2017.1259009

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609-612. doi: 10.1016/j.jrp.2013.05.009

Segalowitz, N., Trofimovich, P., Gatbonton, E., & Sokolovskaya, A. (2008). Feeling affect in a second language: The role of word recognition automaticity. *The Mental Lexicon, 3*(1), 47-71. doi:10.1075/ml.3.1.05seg

Shakiba, N., & Dewaele, J.-M. (2022). Immigrants' language preferences for swearing in Persian and English: The effects of acculturation and sociobiographical background on language choice for swearing. In I. Mavrou, Pérez Serrano, M., Dewaele, J-M.

(Ed.), *Recent advances in second language emotion research* (pp. 191-215): Cizur

Menor: Civitas Aranzadi Thomson Reuters.

Sutton, T. M., Altarriba, J., Gianico, J. L., & Basnight-Brown, D. M. (2007). The automatic

access of emotion: Emotional Stroop effects in Spanish–English bilingual speakers.

Cognition & Emotion, 21(5), 1077-1090. doi:10.1080/02699930601054133

Toivo, W., & Scheepers, C. (2019). Pupillary responses to affective words in bilinguals' first

versus second language. *PLoS One, 14*(4), e0210450.

doi:10.1371/journal.pone.0210450

Toivo, W., & Scheepers, C. (2022). Methodological Approaches to Studying Reduced

Emotional Resonance in Bilinguals' Later Learnt Language (LX). In I. Mavrou, M.

Pérez Serrano, & J.-M. Dewaele (Eds.), *Recent advances in second language emotion

research*: Cizur Menor: Civitas Aranzadi Thomson Reuters.

Winskel, H. (2013). The emotional Stroop task and emotionality rating of negative and

neutral words in late Thai-English bilinguals. *International Journal of Psychology,

48*(6), 1090-1098. doi:10.1080/00207594.2013.793800

Wu, Y. J., & Thierry, G. (2012). How reading in a second language protects your heart.

Journal of Neuroscience, 32(19), 6485-6489. doi: 10.1523/JNEUROSCI.6119-

11.2012