Bilingualism: Language and Cognition

cambridge.org/bil

Examining the reliability and validity of bilingual language use and switching measures

Angela de Bruin

Department of Psychology, University of York, York, UK

Research Article

Cite this article: de Bruin, A. (2026). Examining the reliability and validity of bilingual language use and switching measures. *Bilingualism: Language and Cognition* 1–13. https://doi.org/10.1017/S1366728926100996

Received: 22 January 2025
Revised: 15 December 2025
Accepted: 19 December 2025

Keywords:
bilingualism; language use; reliability; validity; questionnaires

Corresponding author:
Angela de Bruin;
Email: angela.debruin@york.ac.uk

This research article was awarded Open Data badges for transparent practices. See the Data Availability Statement for details.

Abstract

Bilinguals vary in their daily-life language use and switching behaviours, which are also frequently studied in relation to other processes (e.g., executive control). Measuring daily-life language use and switching often relies on self-reported questionnaires, but little is known about the validity of these questionnaires. Here, we present two studies examining test–retest reliability and validity of language-use questionnaires (relative to Ecological Momentary Assessment, Study 1) and language-switching questionnaires and tasks (relative to recorded daily-life conversations, small-scale Study 2). Test–retest reliability and validity of the LSBQ (Anderson et al., 2018) were high and moderate, respectively, suggesting this questionnaire can capture daily-life language use well. Although only examined with a small sample size, Study 2 suggested relatively low validity of most language-switching questionnaires, with short language-production tasks potentially offering a more valid assessment. Together, these studies suggest that tools are available to reliably capture language use and switching with (a certain degree of) validity.

1. Introduction

Many differences exist between bilinguals in terms of (amongst others) age and context of acquisition, language proficiency, language use, switching and sociolinguistic contexts. These individual differences are studied in relation to language production and comprehension (e.g., Uchihara & Clenton, 2023), language choice and control (e.g., Bonfieni et al., 2019; de Bruin & Martin, 2022) and executive control (e.g., Gullifer & Titone, 2021; Hartanto & Yang, 2020; Kałamała et al., 2020), with a recent focus on daily-life language use and switching (Adaptive Control Hypothesis, Green & Abutalebi, 2013).

Very few articles published between 2005 and 2015 report detailed information about bilingual participants' language use (Surrain & Luk, 2019). While 79% of papers included some language-use information (e.g., the home languages), only 39% reported language-use *frequency* (e.g., 20% exposure to Language A on a weekly basis). Furthermore, bilinguals can vary in *how* they use their languages and how often and how they switch languages. Language switching was not examined in Surrain and Luk's review, but given the low number of studies reporting language use or sociolinguistic context, it can be assumed that language switching was typically not measured or reported in detail.

Since 2015 (i.e., the end of Surrain & Luk's review period), there has been an increase in measures assessing daily-life language use and switching (e.g., Anderson et al., 2018; Hartanto & Yang, 2020). However, the reliability and validity of language-use and -switching measures have only been partially examined. This is despite good reliability and validity being crucial to capture individual differences and to examine potential relationships with other processes. Here, we examine reliability and validity of language-use questionnaires (Study 1) and language-switching measures (Study 2).

1.1. Questionnaires

Most measures of language use and switching are based on self-reports (see the Study 2 Introduction for more objective switching measures). In general, self-reports are prone to biases (e.g., Furnham & Henderson, 1982; Okamoto et al., 2002). Focusing on language, biases have especially been discussed in relation to self-rated proficiency, where ratings can be influenced by confidence, language anxiety and motivation (e.g., MacIntyre et al., 1997). Tomoschuk et al. (2019) examined the relationship between self-rated proficiency and MINT scores (Multilingual Naming Test using picture naming). Within Chinese-English bilinguals, participants who grew up in the United States rated their Chinese proficiency higher relative to the objective score than participants who grew up in China. This suggests that the reference point (i.e., in this case, a bilingual comparing themselves to L1 speakers in China or to other bilingual/L2 speakers in the United States) can influence self-ratings of language proficiency.

Similar results were found by Hernández-Rivera et al. (2024), who looked at the role of language solidarity (personal identity and sense of belongingness). French-L1 participants with

greater L2 solidarity underestimated their L2 proficiency. In contrast, in English-L1 participants, those with lower L2 solidarity showed a greater difference between self-rated and objective proficiency. This suggests a complex interplay with feelings of solidarity, also depending on the participant groups tested.

These issues might not necessarily apply to self-ratings of language use, with Hernández-Rivera and colleagues showing that language-use questions were relatively unaffected by L2 solidarity. However, language-use and -switching estimates might still be influenced by comparisons to other people or by language status and attitudes (which vary between bilinguals, Dewaele & Wei, 2014). Furthermore, bilinguals might not always be aware of their switching behaviours (Rodriguez-Fornells et al., 2012).

1.2. Previous analyses of reliability and validity of language-use questionnaires

Although earlier questionnaires (e.g., Language Experience and Proficiency Questionnaire, LEAP-Q, Marian et al., 2007; Language History Questionnaire, LHQ, Li et al., 2006, 2020) also include language-use questions, the LSBQ (Language and Social Background Questionnaire; Anderson et al., 2018) is currently one of the most frequently used questionnaires to assess language use in detail. This includes questions about language use with different people (e.g., parents and friends), in different environments (e.g., at home and in shops and restaurants) and for different activities (e.g., emailing and reading). Using this questionnaire, Anderson et al. (2018) derived three factors in a study with bilinguals and monolinguals: "Non-English Home Use and Proficiency," "Non-English Social Use," and "English Proficiency/Use." Weighted scores per factor were correlated with behavioural data from other tasks, including the English PPVT as a measure of receptive vocabulary. This PPVT score correlated positively with the LSBQ English proficiency factor but negatively with the two non-English use and proficiency factors, with those (predicted) directions taken as support for the LSBQ's ecological validity. Mann and de Bruin (2022) also assessed test–retest reliability, with participants completing the LSBQ twice 2 weeks apart. Most items showed moderate-to-high correlations between sessions ($\rho s > .6$ for 24/29 items) and moderate-to-high agreement (weighted kappa scores >.5 for 22/28 items).

1.3. Previous analyses of reliability and validity of language-switching questionnaires

Continuing with language switching, one commonly used questionnaire is the 12-item BSWQ (Bilingual Switching Questionnaire, Rodriguez-Fornells et al., 2012). Four scores were derived from this questionnaire, reflecting switches to the L1/A, switches to the L2/B, contextual switches (e.g., frequent switching in specific settings) and unintended switches. Rodriguez-Fornells et al. (2012) showed that internal reliability of the raw scores was acceptable (alphas above .7) for switches to Language A, switches to Language B and contextual switch scores, but slightly lower (.58) for the unintended switching score. Correlations were observed between the L1/L2 switch scores and self-reported age of acquisition, proficiency, use and a fluency production task, reflecting that participants with a higher proficiency in and use of a language were less likely to switch to the other language. These correlations were often small and not observed for the overall BSWQ score or contextual switches score. However, some aspects of switching (like contextual switching) are not necessarily related to proficiency or even frequency of language use, making it difficult to draw strong validity conclusions based on these analyses.

Cox et al. (2020) compared the BSWQ to an autobiographical narrative task completed by Spanish-English participants. Bilinguals were free to switch but not explicitly instructed to do so. From these narratives, Cox and colleagues derived scores capturing overall switching to English, intra-sentential (within-utterance) and inter-sentential (between sentences) switching. Overall switching and multi-word intra-sentential frequency correlated positively with the BSWQ-Contextual switches and total score, but not significantly with the BSWQ-Unintended switches and BSWQ-switches to English scores. The one-word intra-sentential ratio also correlated positively with the BSWQ total score. No significant relationships were observed with the inter-sentential ratios. However, even when correlations were significant, *r*s were below .3. This suggests that participants who switch more often in narratives/conversations might also reflect this in their questionnaire answers, but those self-reported rates are not precise.

Testing English-Mandarin bilinguals, Lai and O'Brien (2020) asked participants to complete a questionnaire designed to capture time spent in different interactive contexts, as well as three production tasks. In the story recount task, participants first listened to an audio recording introducing a story using two English-only sentences, two Mandarin-only sentences, and two sentences with intra-sentential switches. They were asked to tell the rest of the story using English and Mandarin. In the un-cued naturalistic conversation task, participants discussed childhood stories with an English-Mandarin bilingual. Finally, participants completed a cued word-switching task in which they named words within a category while continuously alternating languages. These three switching tasks were included to capture different types of switching contexts, matching the contexts asked about in the questionnaire. Comparing switching frequency in these tasks to the self-reported time spent in the corresponding interactional contexts showed no significant correlations (all $rs < .2$; $ps > .05$). However, the contexts created in the production tasks might still differ from the questionnaire descriptions (or participants' interpretations), which could explain the relatively low correlations.

Contrary to these studies comparing questionnaires to lab-based tasks, Jylkkä et al. (2020) used Ecological Momentary Assessment (EMA). EMA is used frequently to sample participants' behaviours and experiences in real time, in the participants' own environment, and is argued to increase ecological validity (e.g., Shiffman et al., 2008). Jylkkä and colleagues asked participants to report their frequency of intended switches, frequency of unintended switches and percentage of contextual switches six times per day (approximately 2 hours apart), during a period of 14 days. Participants also completed the BSWQ. A strong correlation was observed between the EMA-Unintended switches and the BSWQ-Unintended switches ($r = .63$; $p < .001$), but the other correlations were low (<.30; *p*s not fully reported) or negative. This suggested low validity of the BSWQ, with the unintended switches perhaps standing out more and therefore being recalled more accurately. Another single question asking to report average switching frequency correlated moderately with the EMA-Intended switches ($r = .42$, $p < .05$) and EMA-Contextual switches ($r = .39$, $p < .05$). Test–retest reliability was moderate-to-high for the BSWQ-Unintended switches ($r = .80$; $p < .001$) and BSWQ-Language switches ($r = .62$; $p < .001$) but lower for the BSWQ-Contextual switches ($r = .27$; $p > .1$) and the single average-switching frequency question ($r = .40$, $p < .05$).

In summary, while reliability (test–retest and construct reliability) of language-use and -switching questionnaires has been studied

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

and appears at least moderate, relatively little is known about validity. Where validity is assessed, this is typically done by linking questionnaires to other tasks (e.g., vocabulary assessments) or to lab-based tasks (such as narrative/story telling tasks), with the exception of Jylkkä and colleagues (cf. also Arndt et al., 2023). The current research therefore examined the validity of language-use (Study 1, relative to EMA) and language-switching questionnaires (Study 2, relative to daily-life conversations at home).

2. Study 1

2.1. Introduction

In Study 1, we focused on the LSBQ, currently among the most frequently used language-use questionnaires. Compared to other commonly used questionnaires like the LEAP-Q or LHQ, it assesses language use in greater detail by addressing 25 contexts (people, activities and environments). Language use is also often still assessed through single-item questions (cf. Surrain & Luk's review, 2019). As pointed out in the original LSBQ manuscript (Anderson et al., 2018), single-item self-report assessments might be unreliable. We therefore also included single-item exposure and single-item use questions, to compare them to a more comprehensive assessment like the LSBQ.

We focused on test–retest reliability and validity relative to EMA data. Test–retest reliability is unlikely to be perfect as participants' responses might be influenced by their language behaviour immediately preceding a questionnaire and potential changes in language use between time points. Similarly, validity is unlikely to be *very* high because language use can vary substantially even on a daily basis, depending on the exact combination of circumstances (e.g., which activity in combination with which people). However, if participants' responses vary greatly depending on when a questionnaire is completed (low test–retest reliability) or deviate extensively from EMA responses (low validity), they are unlikely to capture real-life language-use patterns well. Therefore, for a questionnaire to be of practical value, we need at least moderate correlations and agreement.

We collected EMA data by asking participants five times per day throughout a period of 5 days to report their current language use, as well as the people/activity/environment that related to. EMA is a particularly suitable method here as it allows participants to provide language-use data in a range of contexts, as also assessed in the LSBQ.

2.2. Methods

Participants

Study 1 was completed by 65 participants. An additional 14 participants started but did not complete all parts and/or did not meet our inclusion criteria of only using two languages regularly (the LSBQ was developed for bilinguals rather than multilinguals). Eleven included participants knew another language but did not regularly use it at the moment of testing. All included participants completed the EMA at least 80% of the time (M = 98.1%, SD = 3.6) and passed the attention checks (e.g., being asked to select answer option "C"). Two participants reported having reading difficulties (e.g., dyslexia), but no participants reported having a language or communication disorder or hearing problems or vision problems that impacted the survey. In addition to English (most participants were living in the United Kingdom), participants spoke one of 27 different languages that we will refer to as Language X. The study was approved by the Ethics Committee of the Department of Psychology at the University of York, and participants provided informed consent. Participants received an Amazon voucher or course credits. The author asserts that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration 1975, as revised in 2008, apart from the study not being pre-registered.

The target sample size was a minimum of 50 participants. This was based on Jylkkä et al. (2020), who conducted a similar EMA study and also aimed for at least a moderately sized correlation and estimated that 30 participants should yield sufficient power (.80). While lower effect sizes might be of theoretical interest, for a questionnaire to have practical value, correlations should be at least moderate. We aimed for a higher sample size to have enough participants left after attrition and exclusion and indeed exceeded the intended sample size.

Table 1. Overview of the Study 1 participants' age of acquisition, self-rated proficiency and time spent using Language X and English

Measure	Language X – Mean (SD)	English – Mean (SD)
Age of acquisition (AoA, years)	1.8 (4.3)	2.8 (3.4)
Self-rated proficiency (0–10)		
Speaking	8.8 (1.8)	8.8 (1.3)
Understanding	9.1 (1.5)	9.2 (1.0)
Reading	7.9 (2.4)	9.3 (1.0)
Writing	7.4 (2.6)	8.8 (1.5)
LexTALE (Lemhöfer & Broersma, 2012, 0–100%)	X	84.9% (9.1)
Time spent using each language (1 = none of the time; 5 = all the time)		
Speaking	3.7 (0.8)	3.9 (0.7)
Listening	3.6 (0.9)	4.0 (0.6)
Reading	2.6 (1.0)	4.2 (0.6)
Writing	2.4 (1.1)	4.1 (0.8)
General use (relative to other language, 0–100%)	35.8% (15.1)	64.2% (15.1)
General use (relative to other language, 0–100%)	37.5% (17.0)	62.5% (17.0)
Language preference (0–100%)[a]		
Speaking	54.2% (21.6)	45.8% (21.6)
Reading	22.0% (21.0)	78.0% (21.0)
Language comfort[a] *across languages* (1, most comfortable in English; 5, most comfortable in Language X)		
Speaking	3.2 (1.3)	
Listening	3.1 (1.1)	
Reading	2.2 (1.1)	
Writing	2.1 (1.1)	

Note: The self-rated proficiency and time spent per language questions that are included here are from parts 16 and 17 of the LSBQ, as answered in the first survey session.
[a]These questions were based on the LEAP-Q (Marian et al., 2007). The two preference questions asked about language preference when reading a text originally written in another language not known to the participant and when considering speaking with another person equally fluent in all languages. At an individual level, 33/65 participants reported a very balanced language preference for speaking, falling between 40% and 60% for both languages.

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

Participants' mean age was 22.4 years old (SD = 4.6). As reported in Table 1, participants had a high proficiency in both languages. They regularly used both languages for speaking and listening but showed slightly higher use of English than Language X for reading and writing, as well as a preference for using English for reading and writing. Seventeen participants reported acquiring both languages from birth; 35 reported acquiring Language X first; and 13 reported acquiring English first. Most participants reported learning both languages at home or in the community (61/65 for Language X and 55/65 for English). Overall, participants had a positive attitude towards code-switching (M = 3.7, SD = 0.6, scale 1 = very negative to 5 = very positive) and towards both languages (all mean scores above 4 on a scale of 1 = not very beautiful/important to 5 = very beautiful/important).

Materials

We included all "Language Background" questions from the LSBQ (i.e., the four questions about childhood/teenager language use; the eight questions about language use with different people; the eight questions about different environments; and nine questions about different activities). We left out the general background questions that were not relevant (e.g., parental education). We made two changes to the original LSBQ. We allowed participants to indicate N/A if a person/context/activity did not apply (e.g., if they did not have a partner). We also included an answer option "Using another language than English or Language X." In the analysis, these answer options (N/A and other language) were not included.

In addition to the LSBQ, we included two "general use" and "general exposure" questions, similar to the type of questions often asked in the literature. The exposure question asked participants how much time they were exposed to Language X and English. Participants had to use sliders from 0% to 100% of the time and had to make sure the two percentages across the languages added up to 100% (where this was not the case, we corrected the scores). The use question had the same format, now asking about participants' own language use. Finally, we included some questions about language switching for the purpose of Study 2 (see Study 2).

As part of the EMA, participants were asked four questions 25 times during five consecutive days. All questions asked participants to think about their language use in the past 10 minutes and to also indicate which people they had communicated with, the environment they had been in and the activities they had been doing. Participants could always choose from the same categories (e.g., "parents," "home," and "reading") also used in the main LSBQ questionnaire (i.e., the EMA and LSBQ question "categories" were identical). This way, we could easily map the EMA data to the questionnaire data. Participants could choose "none of the above" if none applied or if they preferred not to answer. They could always choose multiple answer options if multiple people, contexts, or activities had been part of the last 10 minutes. The answer options of the question asking the participant to choose which option described their language use in the past 10 minutes best also corresponded to the options in the LSBQ.

Procedure

Participants completed the questionnaire twice, once on Day 1 and once on Day 7 (one participant completed the final questionnaire one day later). During Days 2–6, participants received five requests per day to complete a short questionnaire as part of the EMA. The EMA requests were sent once every 2 hours between 10 am and 6 pm via email and (where preferred) text message. In total, each participant received 25 requests, of which 98% were completed. All EMA and LSBQ questionnaires were completed through Qualtrics. Each survey (Day 1 and Day 7) took approximately 30 minutes, while each EMA point took a couple of minutes to complete. At the very end, we also asked participants how they answered the questions. Thirty participants (out of 65) reported not comparing their language use to anyone else. Of those who compared themselves, 16 compared their use to parents/family, 5 to their partner, 22 to friends and 8 to classmates or colleagues (multiple answers were allowed).

Approximately half of the participants were asked to consider the preceding 2 months while answering the LSBQ, while the other half were asked to consider the preceding 6 months. Most participants said that they kept the indicated time frame in mind, with 10 participants reporting that they did not think about a concrete time frame at all, and four participants focused on the past weeks rather than months. Exploratory checks indicated no clear influence of the instructed time frame, and we therefore did not conduct further analyses comparing the groups with 2-month versus 6-month instructions.

Finally, we asked participants if they had undergone large changes in their language use in the months preceding the study. Fifteen participants reported some variations in language use in the preceding months, but open-text responses suggested that these changes were minimal, related to specific people or contexts (e.g., different language use with housemates after moving). Participants also reported that their language use during the week of the study was representative of their general language use (M = 7.2, SD = 2.5; scale 0 = not representative at all, 10 = extremely representative).

Data analysis

The data are available on https://osf.io/6bvk5/.

Study 1 had two key aims: to examine the validity of the LSBQ language use questions relative to EMA (Analysis 1–3; all reporting Spearman-rho correlations) and to examine test–retest reliability (Analysis 4). Analyses were conducted with JASP v19.3 (JASP Team, 2024).

Analysis 1: Examine validity of mean LSBQ scores relative to the mean EMA score. We compared the EMA data to the Day-1 LSBQ data (provided before the EMA was completed), to avoid the EMA enhancing participants' awareness of their language use. For the first analysis, we correlated the mean EMA score (averaged language use score across all responses per participant, regardless of the context/activity/people; excluding responses that said not applicable or another language than the two main languages) with the mean LSBQ score. The mean, unweighted, LSBQ score was taken across all "current language use" questions (sections 19–21), to focus on current use and only include questions that were comparable in format to the EMA questions. We did not include questions about language use during childhood, the final questions about language switching (see Study 2), the self-rated proficiency questions or questions asking about time spent speaking, listening, reading and writing in Language X/English. We also correlated the mean EMA score with the general English language use and exposure scores (based on the two single questions asking about general use/exposure).

Analysis 2: Examine validity of weighted LSBQ factors and EMA factors. Next, we conducted a factor analysis based on the Day-1 LSBQ data, again only including the "current language use" LSBQ questions that were identical to the EMA questions. To avoid too many missing variables (mostly "not applicable" responses), we removed items with fewer than 50 responses (language use with a partner, at work, in religious settings and for praying). All correlations were below .9, and all items had at least one correlation above .3 with another item. We used a principal component analysis (PCA),

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

using promax as the rotation method. We suppressed loadings below .4 and excluded missing values pairwise. The number of components was based on a parallel analysis. None of the items loaded on multiple factors (above .4), but the item "Friends" did not load (above .4) onto any factor and was removed. We then computed participants' weighted factor scores based on the approach described in Anderson et al. (2018). For each participant, we standardised their raw score per item and multiplied this by the item weight derived from the factor analysis. For each factor, these weighted standardised scores were summed across the included items to produce the participant's factor score.

Per EMA response, participants also indicated the context, activity and people included in the preceding 10 minutes. Based on these answers, we also computed weighted EMA scores per LSBQ-derived factor. We used the LSBQ factor structure (i.e., items per factor) and weights to compute these three EMA weighted scores. For instance, if an LSBQ factor included the three questions about language use with parents, siblings and grandparents, the corresponding EMA factor would include the language-use scores when participants indicated that they had spent time with their parents, siblings and/or grandparents. We then correlated the weighted EMA and LSBQ scores per factor.

The use of weighted factor scores is in line with the approach used in the original LSBQ paper (Anderson et al., 2018). We did not use their determined factors as these factors can vary depending on the participants tested (cf. also Mann & de Bruin, 2022). In the Anderson et al. study, both monolinguals and bilinguals were included, which can result in different factors than when only bilinguals are tested, as in the current study. While our current sample size might be on the low side for a PCA analysis, studies in the literature conduct similar analyses with comparable sample sizes, and it is important to therefore examine the validity of this approach.

Analysis 3: Examine validity of individual LSBQ questions relative to corresponding EMA contexts/activities/people.

Analysis 4: Examine test–retest reliability for the following Day 1–Day 7 comparisons (same scores as included in Analyses 1–3): mean LSBQ score, general English exposure, general English use, the three weighted LSBQ factor scores and the individual items included in the three factors. We report Spearman's rho values to reflect correlations and weighted kappas and ICC scores as measures of agreement. For completeness (reported in the Supplementary Materials), we also conducted these analyses with the additional LSBQ items: self-rated proficiency and time spent speaking, listening, reading and writing in English and Language X.

2.3. Results

Analysis 1: General LSBQ, use and exposure – validity

The mean LSBQ score correlated positively with the mean EMA score ($\rho = .556$, $p < .001$, see Supplementary Figure 1). Both general use ($\rho = .539$, $p < .001$) and general exposure questions ($\rho = .446$, $p < .001$) also correlated with mean EMA scores. Finally, the general English exposure and use scores also showed positive correlations with the mean LSBQ score (general use: $\rho = .662$, $p < .001$; general exposure: $\rho = .541$, $p < .001$) and with each other ($\rho = .859$, $p < .001$).

Analysis 2: Weighted LSBQ and EMA factors – validity

The PCA analysis identified three factors, with eigenvalues above 2.7. The overall KMO value was .648, with all included items above .5. KMO was likely slightly lower than ideal because participants could indicate that a context was not applicable to them, leading to missing responses. Bartlett's test was significant ($\chi^2(190) = 788.980$, $p < .001$). The first factor can be summarised as "interactions outside the family environment" and included housemates, extracurricular activities (hobbies/sports), neighbours, social activities, shopping/restaurants/commercial services, TV/radio, official instances (health care, public offices, banks), movies and school/university. The second factor focused on (online) activities without in-person interaction and included browsing on the internet, social media, emailing, reading, writing lists/notes and texting. Finally, the third factor was best summarised as interaction with family members/at home and included parents, grandparents, siblings, other relatives and home.

We computed weighted factor scores for EMA and LSBQ responses based on the three factors described above. Significant EMA-LSBQ correlations were observed for the first factor (interactions outside the family environment: $\rho = .633$, $p < .001$) and for the third factor (interaction with family/at home: $\rho = .439$, $p < .001$). However, Factor 2 EMA and LSBQ responses did not correlate (activities without an in-person interaction component: $\rho = .078$, $p = .538$, see Figure 1).

Analysis 3: Item-specific correlations LSBQ and EMA – validity

Correlations between *individual* LSBQ items and EMA responses to the corresponding activity/person/context were only computed for items with sufficient data points (i.e., >50% of participants), as not all individual LSBQ items came back in the EMA responses for all participants (e.g., not all participants talked with their neighbours during the EMA period). Correlations were moderate for items referring to concrete people (parents, siblings and friends), social activities, tv/radio and texting but not for the other activities and environments (see Supplementary Table 1).

Analysis 4: Test–retest reliability

Finally, we conducted test–retest analyses. First, mean LSBQ scores correlated positively and significantly across Day 1 and Day 7 ($\rho = .845$, $p < .001$) and showed strong agreement in terms of weighted kappa (.835, 95% CI: .763–.907) and ICC (.852, 95% CI: .768–.907). Correlations across Day 1 and Day 7 were a little lower but still strong for the two general single-item questions: % English use ($\rho = .827$, $p < .001$) and % English exposure ($\rho = .708$, $p < .001$). Similarly, agreement was a little lower but still substantial for English use (weighted kappa: .752, 95% CI: .653–.851; ICC: .766, 95% CI: .644–.851) and for % English exposure (weighted kappa: .660, 95% CI: .479–.840; ICC: .658, 95% CI: .495–.777).

To examine the LSBQ-derived factors, we first conducted another PCA on the Day 7 data to examine whether similar factors would be identified as in Day 1. We only entered the items also included in the Day-1 factors. The item "watching movies" did not load on any factors and was removed. The overall KMO score was .732, with all items scoring above .5. Bartlett's test was significant ($\chi^2(171) = 633.810$, $p < .001$). The Day-7 analysis again showed three factors, similar to the Day-1 analysis. Those three factors again included the same items and reflected the same general constructs. The only key difference was that the order of factors was reversed (Day-1's Factor 3 on interaction with family now had a larger eigenvalue than Day-1's Factor 2 on online activities).

Next, we examined test–retest reliability for the three weighted factors (with factor numbers corresponding to the Day-1 analysis). Reliability was generally high: Factor 1: $\rho = .818$; ICC: .897, 95% CI: .836–.936; Factor 2: $\rho = .807$; ICC: .890, 95% CI: .825–.931; Factor 3: $\rho = .769$; ICC: .934, 95% CI: .893–.959. We ran similar analyses for the individual items included in these factors, for which test–retest reliability was generally moderate-to-high (see Supplementary Tables 2 and 3 for self-reported proficiency).

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

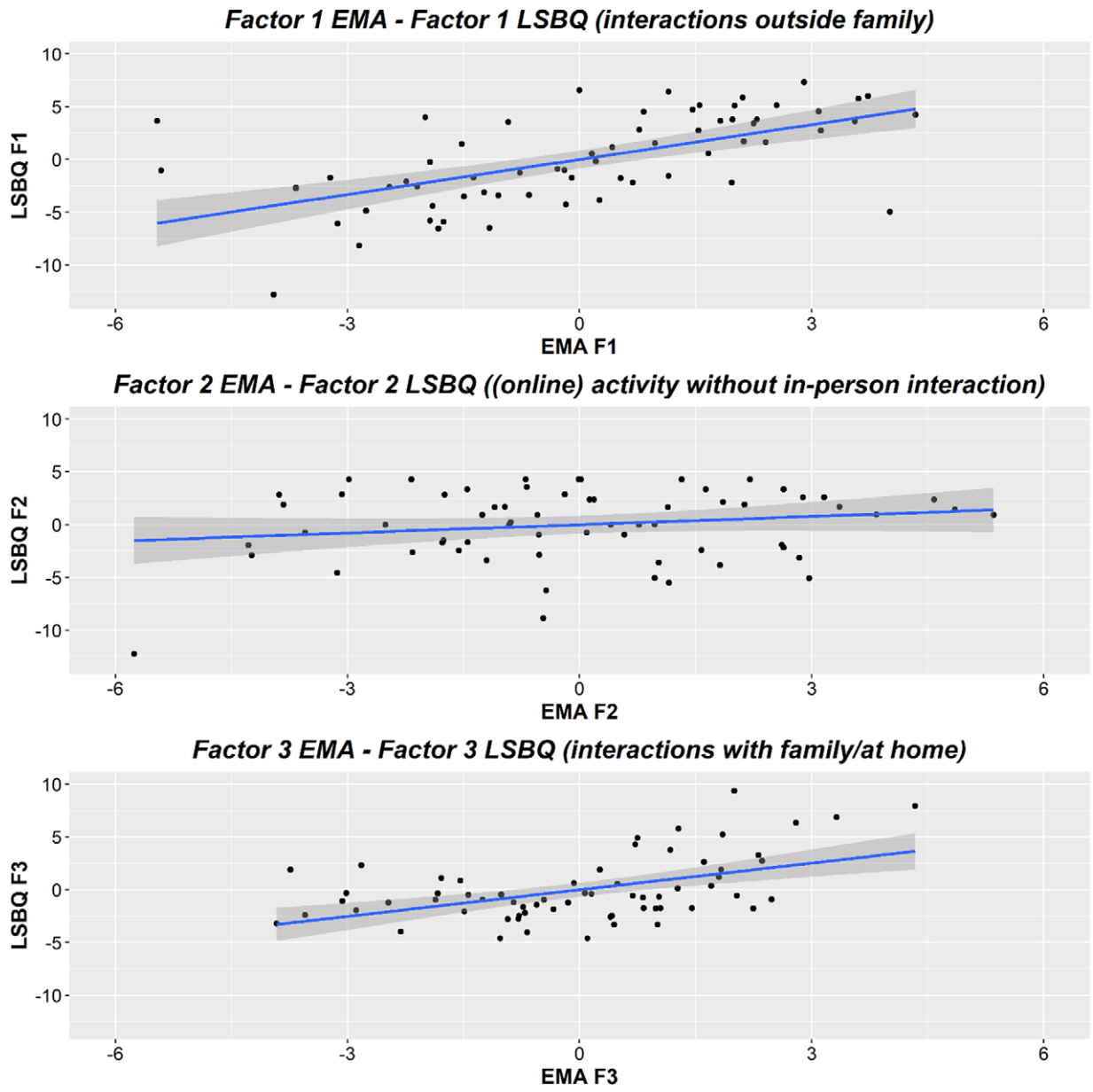

Figure 1. Scatterplots showing the relationship between EMA and LSBQ scores per factor (weighted). The top plot shows the relationship for Factor 1, reflecting interactions outside the family/home. The middle plot shows the (absence of a significant) relationship for Factor 2, (online) activities without an in-person component. The bottom plot shows the relationship for Factor 3, interactions with the family/at home.

2.4. Discussion

Study 1 examined test–retest reliability and validity (relative to EMA scores) of the LSBQ (Anderson et al., 2018), a commonly used measure of language use in bilinguals. The mean LSBQ score, as well as general exposure/use questions, showed a moderate relationship with the mean EMA scores. Weighted EMA and LSBQ factor scores also showed moderate relationships for factors related to language use with the family/at home and outside the family environment but not for the factor reflecting (online) activities without other people. At the individual item level, correlations between EMA and LSBQ scores varied more.

Test–retest reliability was substantial-to-strong, in line with previous research using a two-week time frame (Mann & de Bruin, 2022). Together, these studies suggest that participants provide comparable answers when asked to complete the LSBQ multiple times. In terms of validity, previous studies observed good validity relative to other measures (e.g., proficiency tasks, Anderson et al., 2018). Here, we show generally moderate validity of the LSBQ relative to EMA, when using mean or weighted factor scores. The validity of Factor 2 ("(online) activities without an in-person component"; and some of the included individual items, such as online browsing and emailing) was much lower than the others. This factor referred to online language behaviours, mostly without in-person interactions. It is likely that these types of behaviour vary more depending on the addressee (e.g., writing an email to your mother or friend) and topic (e.g., browsing topics related to studying in the United Kingdom). In the questionnaires, participants often reported only using English online, while EMA data showed considerably more variability. Similarly, a lower-scoring item like reading can refer to different types of texts. Future

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

studies might therefore want to break activities down into more concrete questions (e.g., "reading for pleasure" or "reading for school/work").

We also included two more general use/exposure questions. Although validity and reliability were a little higher for the LSBQ (especially compared to the exposure question), reliability and validity scores were relatively similar. The advantage of the LSBQ, however, is that it allows for more detailed examinations of language use, including different types of language use established through factor analyses. Indeed, we again showed that language use patterns differ by context, with these factors remaining consistent across the LSBQ testing points. This highlights the importance of considering different contexts (in the form of people, activities and environments) when asking about daily-life language use.

The relatively short EMA duration (5 days, with questions asked once every 2 hours) has limitations. Some items refer to activities that might vary more (e.g., reading, browsing). A longer EMA period could help to establish more stable scores for these activities. Furthermore, participants did not always report being in a certain context/doing a certain activity during the EMA period. Therefore, for individual participants, weighted scores did not always include the same number of EMA and LSBQ items. Nevertheless, correlations were generally moderate, suggesting good validity even if the comparisons were not perfect. EMA scores at an item level could furthermore be influenced by the number of times a participant reported being in that context. While this can play a role, however, it did not immediately appear to explain the results (e.g., validity for "home" was low, even though participants often reported that context). Despite these limitations, the present data strongly suggest that self-reported language use can provide a reliable and valid snapshot of daily-life language use.

3. Study 2

3.1. Introduction

Self-rating language switching might be more difficult than self-rating language use, with some studies questioning the validity of existing switching questionnaires (e.g., Jylkkä et al., 2020). In Study 2, a small-scale pilot-type study, we therefore provide an initial examination of the validity of commonly used switching questionnaires (BSWQ, Rodriguez-Fornells et al., 2012; Bilingual Interactional Context Questionnaire, BICQ, Hartanto & Yang, 2016, 2020; Language Mixing Scale, LMS, Byers-Heinlein, 2013; and a switching version of the LSBQ). In addition, we included a code-switching frequency-judgement task (e.g., Hofweber et al., 2016). This presents participants with sentences and asks them to indicate how often they encounter similar utterances in daily-life bilingual conversations. Although still relying on ratings, this task has been described as more ecologically valid and has been found to predict code-switching frequency in emails (Hofweber et al., 2018).

We also included two production tasks as a more objective switching-frequency measure. In the first, participants see pictures they can name in their language of choice (e.g., de Bruin et al., 2018; de Bruin & Xu, 2023; Gollan et al., 2014; Gollan & Ferreira, 2009). In these tasks, switching frequencies vary between participants and have been associated with self-reported daily-life switching frequency (e.g., Coumel et al., in press). The second production task included four short story-telling/narrative tasks, chosen in line with previously used narrative tasks (e.g., Cox et al., 2020; Lai & O'Brien, 2020). While this task did not include a conversation partner (making it less naturalistic), this also ensured participants' switching was not influenced by other bilinguals.

To assess validity of these tasks and questionnaires, we compared switching frequency against recorded daily-life conversations participants had at home with other bilinguals. We used recorded conversations (rather than, e.g., EMA) to also capture switches bilinguals might not be aware of themselves. These data only provide a first basis for future research to examine these questions in (much) larger sample sizes. The results should therefore be interpreted as equivalent to pilot data.

3.2. Methods

Participants

Study 2 was completed by 16 participants. A larger sample size was not feasible due to the study complexity and funding duration. An additional 13 participants started the study (i.e., completed the first questionnaire and/or some recordings) but were excluded because they either did not take part in the rest of the study (n = 4); indicated in the questionnaire that they had no knowledge of Spanish or they did not pass the attention checks (n = 6); or predominantly used a language other than Spanish or English in their recordings (n = 3). For the remaining participants, 12 took part "independently" (i.e., their conversation partner gave consent to be recorded but was not a study participant) and 4 took part with another participant (i.e., in two sets of recordings both conversation partners were participants). All participants (12 females, 4 males) spoke Spanish and English. Fourteen participants were born in a Spanish-dominant country and grew up speaking Spanish but were currently living in the United Kingdom; two grew up speaking English and acquired Spanish during adolescence. Participants on average had been living in the United Kingdom for approximately 7.9 years at the time of testing (SD = 8.0). All participants reported high proficiency in and frequent use of and exposure to both languages (see Supplementary Table 4; all participants reported using each language at least 25% of the time). Participants reported no language or reading difficulties. Table 2 reports the participants' self-reported daily-life language switching, measured through the Bilingual Switching Questionnaire (BSWQ, Rodriguez-Fornells et al., 2012), three switching questions, a switching version of the LSBQ and the Language Mixing Scale (LMS, Byers-Heinlein, 2013). It also reports language use as measured through the LSBQ (Anderson et al., 2018), switching frequency by context measured through the Bilingual Interactional Context Questionnaire (BICQ, Hartanto & Yang, 2016, 2020) and participants' code-switching (CS) attitudes (Dewaele & Wei, 2014). Participants varied in their self-reported switching, from very rare to very frequently.

Materials and scoring

We included the following questionnaires, commonly used in the literature (e.g., most have over 200 citations in Google Scholar):

Bilingual Switching Questionnaire (BSWQ, Rodriguez-Fornells et al., *2012):* Rodriguez-Fornells and colleagues conducted a factor analysis eliciting four factors: Switches to Language A, Switches to Language B, Contextual switches and Unintended switches. Each factor includes three questions, with each question answered on a scale from 1 ("never") to 5 ("always"). An Overall Switching score can be obtained by adding all items (min score = 12; max score = 60).

Bilingual Interactional Context Questionnaire (BICQ, Hartanto & Yang, 2016, 2020*):* This questionnaire asks participants how much time they spend at home, school/university, work and other contexts. Per context, it asks participants how much time they

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

Table 2. Overview of the Study 2 participants' self-rated daily-life language switching

	Mean (SD)	Min–Max scores
BSWQ (12–60; 60 = most frequent)	37.4 (6.9)	22–46
1 (switches to Spanish; 3–15)	9.1 (2.0)	5–12
2 (switches to English; 3–15)	10.3 (2.2)	7–15
3 (contextual switches; 3–15)	10.3 (3.1)	4–15
4 (unintended switches; 3–15)	7.7 (2.9)	4–13
Switching frequency (1 = never; 5 = all the time)		
On a daily basis	3.7 (1.1)	2–5
In a conversation	3.4 (0.8)	2–5
Within a sentence	2.9 (1.0)	1–5
LSBQ Language switching (1 = never; 5 = all the time)		
Mean LSBQ	2.4 (0.6)	1.6–3.2
Chosen conversation partner	3.9 (1.4)	1–5
When speaking	3.8 (1.0)	2–5
When listening	2.4 (1.0)	1–4
When reading	2.3 (1.2)	1–5
When writing	2.8 (1.6)	1–5
Language mixing scale[a] (0–30, 30 = most frequent)	16.0 (8.1)	2–26
Language use (1–5; 1 = all English; 5 = all Spanish)		
Mean LSBQ	2.6 (0.4)	1.7–3.5
Chosen conversation partner	3.6 (0.7)	3–5
For speaking	3.0 (0.6)	2–4
For listening	2.4 (0.5)	2–3
For reading	2.2 (0.7)	1–3
For writing	2.0 (0.6)	1–3
BICQ (0–100% time spent)		
Single-language context	62.8 (18.8)	28–90
Dual: switching between people/utterances	26.2 (19.3)	6–70
Dense CS within utterances	11.0 (12.1)	0–44
Attitudes towards CS (1–5; 5 = most positive)	4.2 (0.6)	3–5

[a]Participants also indicated their reasons for using words from the other language. They indicated borrowing English words when no Spanish translation exists (9/16), when they are not sure of the Spanish word (7/16) and when teaching new words (1/16). They also indicated borrowing Spanish words when no English translation exists (8/16), when unsure of the English word (7/16), when the English word is hard to pronounce (2/16) and when teaching new words (2/16). In the "other" category, participants indicated using words from the other language to speed up conversations, because they are easier to remember and when talking about a specific topic associated with that language.

spend speaking only one language and rarely switching, speaking two languages while conversing with different speakers within that environment but rarely mixing languages within a sentence ("BICQ_Dual"), and routinely mixing two languages within a sentence ("BICQ_Dense"). For each participant, we computed a mean score for each switching type across the four contexts (home, school, work, other), relative to the amount of time spent within each context. Where participants' scores did not add up to 100%, this was adjusted.

Language Mixing Scale (LMS, Byers-Heinlein, *2013):* Although this questionnaire was developed to ask parents about their switching with their children, the questions can also be asked about adults' own code switching. Byers-Heinlein (2013) showed the five questions correspond to one factor, which demonstrated high test–retest reliability across two sessions 6 months apart.

In addition, we included:

LSBQ (switching): The original LSBQ includes three questions about language switching (with family, friends and on social media). Here, we included the same people, contexts and activities as in the LSBQ used in Study 1 but now referred to language switching rather than language use. Participants' answer options were never, rarely, sometimes, often, all the time and not applicable. We computed a mean score; a factor analysis was not possible due to the sample size.

Three questions on language switching: We included three questions asking participants to indicate how often (on a scale from 1 to 5) they switched languages on a daily basis, within a conversation (with another bilingual) and within a sentence (with another bilingual). We computed one mean score across these questions ("3Q").

In addition to these questionnaires, we also included the following tasks:

Code-switching frequency-judgement task (e.g., Hofweber et al., 2016): Participants saw 40 sentences (order randomised; using English and Spanish equally): twenty without a switch and twenty with code switches (insertions or alternations). Participants were asked to rate (scale 1 = not at all, 7 = very often) how frequently they encounter similar utterances in informal conversations with Spanish-English speakers, regardless of the content. The sentences were based on Fricke and Kootstra (2016), adjusted to create sufficient examples per category. The switch and non-switch sentences contained a comparable number of words (M = 6.7 and 6.6 words, respectively). We computed a participant's mean score for sentences with and without switches, taking the difference as a measure of their code-switching frequency. More negative scores reflected less switching (most participants reported lower scores for switches than for sentences without a switch).

Picture-naming task: Participants saw pictures they could name in Spanish or English, completing 98 trials with 14 unique pictures (each picture being repeated seven times). The pictures corresponded to easy-to-name words, between one and three syllables long. We scored whether participants switched languages relative to the previous picture and then computed the percentage of switches relative to the number of correctly answered items that could be identified as a switch or non-switch (trials preceded by a break or mistake were excluded as trial type could not be determined).

Story-telling task: Participants were asked to talk about four different topics for 1 minute each: their favourite holiday, their hobbies and two fairy tales of their choice. The recordings were transcribed by L1 Spanish speakers with a high proficiency in English and were checked by another English-Spanish speaker. We counted the number of switches per story and then computed the percentage of switches relative to the total number of words per story. We excluded ambiguous hesitations (e.g., "um") and incomplete words (e.g., "re…") that could not be assigned to a language. We then computed the mean switching frequency across the four stories. Most switches were intra-sentential insertions. Insertions and dense code switches were counted as one switch (i.e., a switch from the matrix language to the language of the inserted word). For instance, in a fictional sentence like "John eats the *manzana* at the pool," the switch to Spanish "manzana" would count as one switch. Where participants alternated

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

between languages, either within or between sentences, these were counted as one switch to the new language and one switch back to the previously used language. Names of places or brands were only counted as switches if they differed across languages (e.g., London/Londres) or if there was a noticeable switch in pronunciation.

Daily-life recorded conversations: Participants were asked to record their conversations at home during 3 days, with each recording about 20 minutes long. Participants received no detailed instructions, apart from asking them to have a normal conversation (e.g., while preparing dinner) and to not include any personal/sensitive information about themselves or others. They recorded their conversation with their partner (n = 11), family (n = 4) or friend (n = 1). All participants reported their language use to be quite similar (n = 3) or very similar (n = 13) to their common daily-life language use. For those who reported "quite" rather than "very," two reported using Spanish less than normal and one using English less than normal. All participants reported their switching to be similar to other daily-life conversations, although some participants noted that their switching frequency depends on the conversation partner. The conversations were scored similarly to the story-telling tasks described above. We only considered the language of the participant, not the language of the conversation partner. Again, a switching percentage was computed per conversation, and the mean was taken across the three conversations.

Procedure

Participants were first asked to read an information sheet describing the entire study and to complete a consent form. They were asked to choose one other Spanish-English bilingual that they regularly interacted with, to record the home recordings with. Participants were instructed to only choose one other bilingual and to record all conversations with the same person and at home, with no other people present. The study could only begin after receiving consent from both the participant and the conversation partner.

Participants first completed the questionnaires, as described above. Questionnaires were always completed on the first day to avoid participants just self-reporting their switching behaviour in the specific conversations they recorded. Prior to answering the questionnaires, we gave participants examples of switches between sentences and within sentences. At the end, participants were asked if they compared themselves to other bilinguals while rating their switching frequency. Most (11) participants reported that they did not compare themselves to other people. During the following 3 days, participants recorded their daily-life conversations. During the final day, participants completed the three switching tasks (frequency-judgement, picture-naming and story-telling tasks). Instructions for these tasks were given in a combination of Spanish and English, and participants were instructed that they could use both languages freely as they wanted. The order of tasks and questionnaires within each session was counterbalanced. Both sessions took less than 30 minutes to complete.

Data analysis

Three types of correlational analyses were run: 1) examining correlations between the questionnaires, 2) examining correlations between the tasks and questionnaires and 3) examining correlations between the tasks, questionnaires and the conversations as a measure of validity. The sample size was low, and we therefore focus on descriptive statistics. It should again be noted that these data are only intended to provide an initial estimate of validity of tasks and

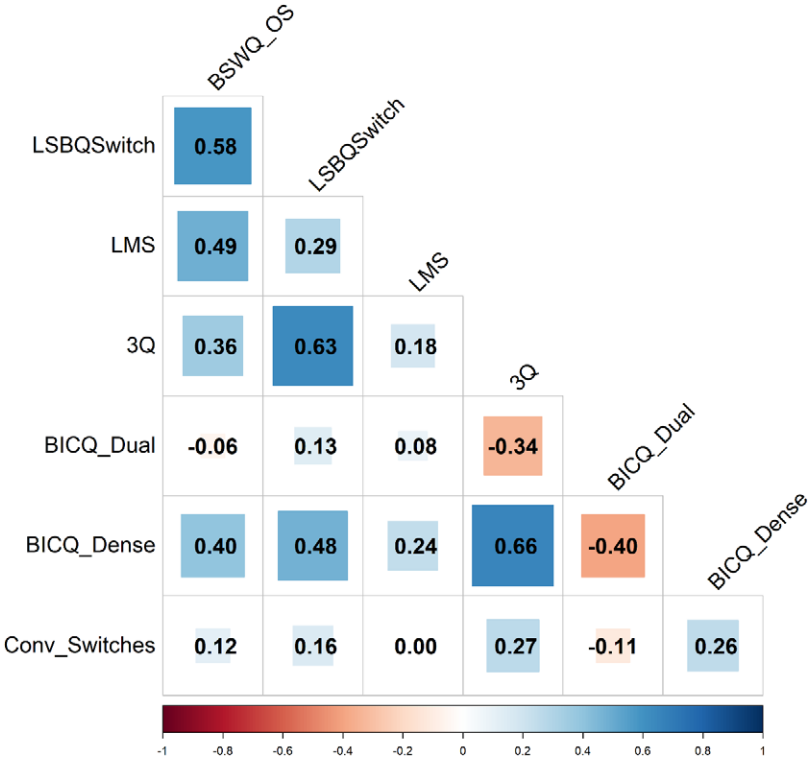

Figure 2. Correlation matrix for the included questionnaires (using Spearman's rho). BSWQ_OS refers to the overall score in the Bilingual Switching Questionnaire; LSBQSwitch to the LSBQ version asking about switching; LMS to the Language Mixing Scale; 3Q to the three short questions on language switching; BICQ_Dual and BICQ_Dense to the questions in the Bilingual Interactional Context Questionnaire asking about switching between versus within utterances, respectively. Conv_Switches refers to the correlations with the recorded daily-life conversations.

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

questionnaires. We also provided Bayes Factors to quantify evidence for the alternative hypothesis (correlation), in the format BF_{10}, with scores above 1 supporting a correlation and scores below 1 suggesting low validity. However, again these outcomes should purely be interpreted as initial pilot-like findings. Finally, we also report test–retest reliability for the switching measures included in Study 1: the LSBQ switching questions and LMS.

3.3. Results

Analysis 1: Correlations between the questionnaires

Correlations between the questionnaires varied (see Figure 2). The BSWQ, LSBQSwitch, 3Q and BICQ_Dense scores correlated positively with each other (ρs .36–.66; BF_{10} = 1.4–11.4).

The LMS showed slightly lower but still positive correlations with the other questionnaires (BF_{10} = 0.5–2.2). The BICQ_Dual showed *negative* correlations with the BSWQ, 3Q and BICQ_Dense scores (BF_{10} = 0.5–1.0). Correlations with the four BSWQ factors individually are shown in Supplementary Figure 2.

Analysis 2: Correlations with and between tasks

Switching frequency in the picture-naming task and the story-telling tasks was positively correlated (ρ = .50; BF_{10} = 1.9). The frequency-judgement task correlated with switching frequency in the story task (ρ = .56; BF_{10} = 2.2) but not strongly with the picture-naming task (ρ = .27; BF_{10} = 0.7).

The frequency-judgement task showed moderate to high correlations with the BSWQ_OS and LMS (ρ = .66 and .52 respectively; BF_{10} = 3.9–4.0) but low or negative correlations with the other questionnaires (ρ = −0.33 to 0.38; BF_{10} = 0.5–0.9).

Switching frequency in the picture-naming task versus the questionnaires generally showed only small correlations, not supported by the Bayes Factors (BSWQ_OS ρ = .08; BF_{10} = 0.5; LSBQSwitch ρ = .15; BF_{10} = 0.6; 3Q ρ = .23; BF_{10} = 0.8; BICQ_Dense ρ = .17; BF_{10} = 0.9). There was a negative correlation between picture-naming switching frequency and the LMS (ρ = −.22; BF_{10} = 0.7) and BICQ_Dual (ρ = −.45; BF_{10} = 3.4). Similar patterns were observed for the story-telling task, with at best small positive correlations (BSWQ_OS ρ = .22; BF_{10} = 1.3; LSBQSwitch ρ = .05; BF_{10} = 0.6; 3Q ρ = .16; BF_{10} = 0.6; LMS ρ = .19; BF_{10} = 0.9; BICQ_Dense ρ = .20; BF_{10} = 0.8) and a negative correlation with the BICQ_Dual (ρ = −.37; BF_{10} = 1.1).

Analysis 3: Correlations with the conversation task (validity analysis)

Within the conversation task, mean switching frequency was 1.8% of words (*SD* = 1.5, min = 0.2%, max = 6.3%). One participant showed a relatively high switching frequency (6.3%) that deviated from the other participants showing low switching frequencies.

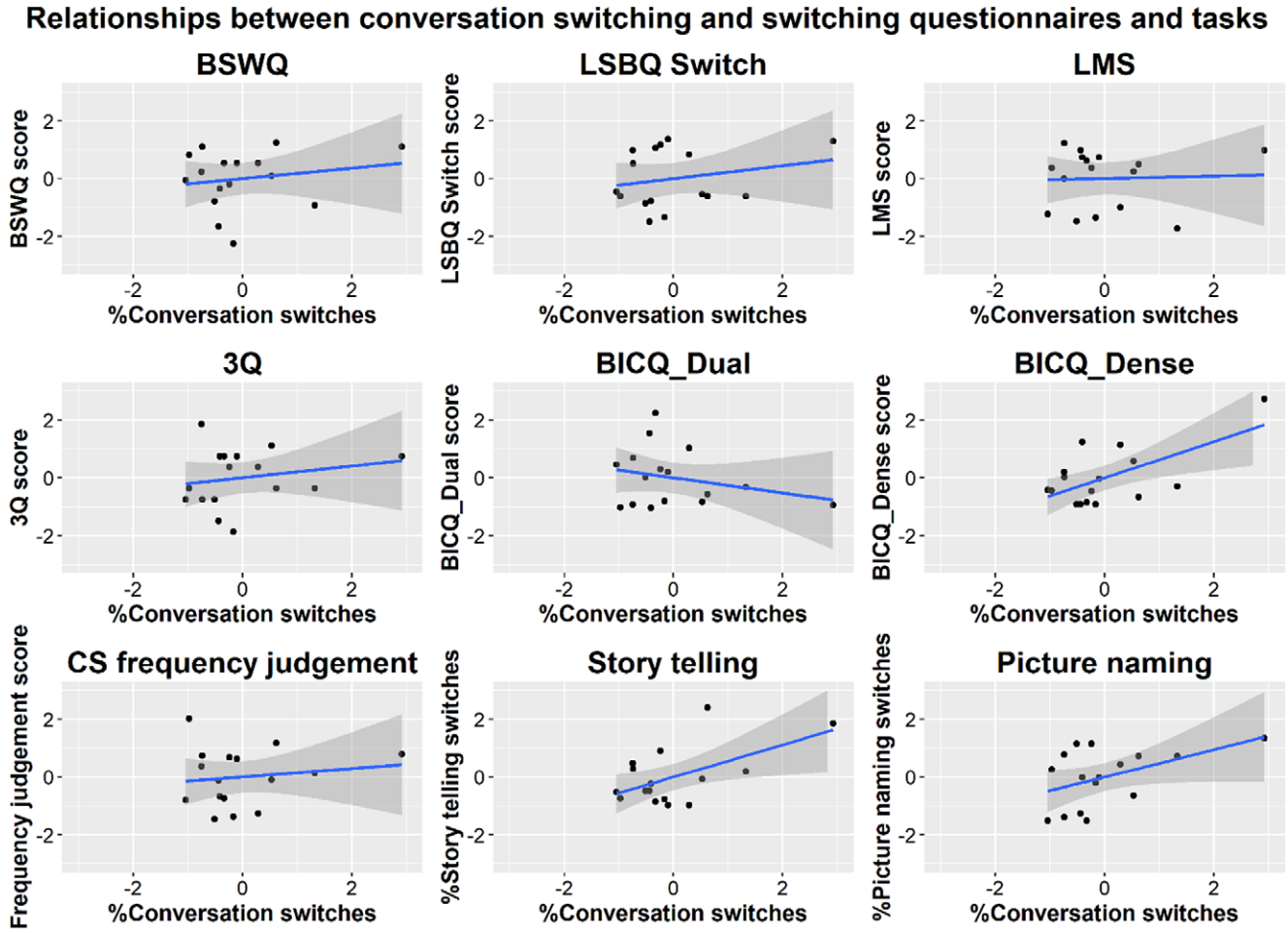

Figure 3. Scatterplots showing the correlations between the switching frequency in the conversation tasks (always shown on the x-axis) and the questionnaires (top two rows) and tasks (bottom row).

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

That participant was included in the analysis as their switching frequency is most in line with means observed in previous corpus data with Spanish-English bilinguals (e.g., Fricke & Kootstra, 2016). In the story-telling task, the mean was 2.5% (*SD* = 2.3, min = 0.2%, max = 8.0%). In the picture-naming task, as expected, switching frequency was higher (*M* = 25.4%, *SD* = 16.7, min = 0%, max = 47.9%). Most participants used Spanish as their base/matrix language in the conversations (11 participants; five participants used English or did not have one matrix language; *M* Spanish = 67.4%, *SD* = 36.3, min = 0.4%, max = 99.1%). In the story-telling and picture-naming tasks, participants varied more (story: *M* Spanish = 50.6%, *SD* = 33.1, min = 1.5%, max = 98.4%; picture naming: *M* Spanish = 48.5%, *SD* = 32.0, min = 0%, max = 100%). Percentage Spanish/English use correlated moderately to strongly across tasks (conversations/picture naming: ρ = .42; conversations/story telling: ρ = .70; picture naming/story telling: ρ = .47).

Figure 3 shows the relationships between switching frequency in the conversations and the questionnaires and the relationships between the conversation switching frequency and the tasks. In general, correlations between the recorded daily-life conversations and the questionnaires were low (ρs < .3, see Figure 2). Bayes Factors for the BSWQ_OS, LSBQ_Switch, BICQ_Dual, 3Q and LMS ranged from BF_{10} 0.5 to 0.7, thus all providing evidence in the direction of the null hypothesis (no correlation with daily-life switching).[1] The BICQ_Dense correlation was larger (see Figures 2 and 3) and supported by a Bayes Factor of 6.4, suggesting higher validity of this questionnaire. The frequency-judgement task's correlation with the conversation switches was low (ρ = .14; BF_{10} = 0.6) and comparable to that of most questionnaires.

The story- and picture-production tasks showed slightly higher correlations with the conversation task (ρ = .22; BF_{10} = 3.4 and ρ = .36; BF_{10} = 1.9, respectively; see Figure 3). For consistency with the questionnaire analyses, Spearman's rho values are reported, even though a few tasks/measures used continuous rather than ordinal DVs. For those tasks, Pearson's *r* correlations with the conversation task were as follows: picture-naming task .47, story-telling task .56 and frequency-judgement task .15. For the questionnaires in Figures 2 and 3 using continuous measures, the Pearson's *r* values were −.26 for BICQ_Dual and .63 for BICQ_Dense.

Finally, focusing on the two sentence-production recordings (daily-life conversations and story-telling tasks), participants usually switched from Spanish to English, mostly producing English insertions. Indeed, when looking at the English switching frequency in the conversations (%switches to English relative to number of English words used), mean switching frequency was much higher (31%) than the overall mean of 1.8%. Focusing on English switching frequency only, the conversations and story-telling task correlated (*r* = .72, ρ = .57; BF_{10} = 18.1).[2]

Test–retest reliability

Finally, we assessed the test–retest reliability for the switching items included in Study 1 (switching questions based on the LSBQ and the LMS). As shown in Supplementary Table 5, most items showed moderate test–retest reliability.

3.4. Discussion

Study 2 examined the validity of language-switching questionnaires and tasks, relative to daily-life conversations. This was based on a small data set, which should be taken as an initial investigation that requires further larger scale studies for more nuanced evaluation.

The overall emerging pattern is that questionnaires generally correlated well with each other. Test–retest reliability of the switching questions included in Study 1 was also generally moderate. However, correlations with switching frequency in the recorded conversations were low.

Correlations between the switching questionnaires suggested good validity relative to each other. The positive correlations between most questionnaires and the BICQ_Dense (asking about intra-sentential switching) together with the negative correlations with the BICQ_Dual (asking about between-utterance/-people switching) suggest that bilinguals perhaps think most strongly about intra-sentential switching when considering their general language-switching behaviours in questionnaires. This aligns with Cox et al.'s (2020) results showing correlations between questionnaires and an autobiographical memory task for intra- but not for inter-sentential switching.

Contrary to the picture-naming/story-telling tasks, the code-switching frequency-judgement task correlated well with some questionnaires. This suggests that even though the frequency-judgement task asks people to report general code-switching exposure rather than one's own behaviour, participants might still respond to these ratings in a similar manner as when rating their own switching behaviours.

In terms of validity relative to daily-life switching, the current (small-scale) study did not provide evidence for (strong) correlations between questionnaires and daily-life conversations. Although very preliminary and influenced by a high-switching participant, the BICQ_Dense score (asking specifically about intra-sentential switching) might be an exception, potentially also because the participants predominantly switched intra-sententially in their conversations. However, the BICQ might also benefit from asking people to reflect on different types of switching, thereby offering participants a more concrete (reference) framework and more precise definition of switching. Finally, the BICQ adjusts for time spent in different contexts (e.g., home, work), which might further increase the self-ratings' accuracy.

The two production tasks showed higher validity than most questionnaires, which could support the use of short lab-based tasks to capture language-switching behaviours. The switching frequency by-language (switches to English) analysis furthermore suggested that this can be done most accurately by focusing on switches to participants' non-matrix language. However, given the small sample size, further research is required to better understand the validity of these tasks versus questionnaires.

The current study was only a small, first step towards assessing tasks and questionnaires capturing code-switching frequencies. The sample size was small and likely unable to reliably capture small relationships between questionnaires and real-life switching. With larger sample sizes, significant relationships might emerge. However, power analyses based on the current data suggest that at least for some questionnaires (e.g., the frequently used BSWQ), sample sizes over 500 participants might be needed to reach over 80% power to detect a significant correlation with daily-life recordings (G*Power, using the observed ρ = .12, not correcting for

[1] As a check, we also examined the correlation between conversation switches and the self-reported switching with that person. This was close to 0 (ρ = 0.02). In terms of language use (frequency of Spanish use), the correlation between self-reported Spanish use with this conversation partner and Spanish used in the recording was ρ = .24.

[2] We did not conduct these English-switch analyses for the picture-naming task as people were less consistent in the use of one matrix language and not all participants switched to English in that task.

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press

multiple comparisons). If correlations are indeed low, this raises further concerns around the practical usability of these questions to capture daily-life language experiences at an individual level.

Another limitation is that most participants switched very little in their conversations. Only one participant switched more frequently. That participant appeared as an outlier in the current data set (and had some influence on the current results) but was in fact most in line with previous corpus research looking at switching frequency in Spanish-English bilinguals (Fricke & Kootstra, 2016). Future research will not just require much larger sample sizes but also a larger range in individual differences in switching behaviours, ideally also including participants using different languages and from different language backgrounds.

We furthermore asked participants to only record conversations with one other person. It is likely that participants' switching behaviours vary depending on the context and people present (e.g., bilinguals are unlikely to switch with monolinguals). In questionnaires asking participants to reflect on their switching, they are likely to consider their estimated switching frequency across multiple conversation partners, potentially also considering how often they are in a scenario (with bilinguals) where they can switch. An analysis including recordings of many daily-life conversations, with different people and environments, might show higher questionnaire validity. However, we also included a question asking participants to self-report their switching frequency with the chosen conversation partner. The correlation between that specific self-report and the recorded switches with that person was close to zero (footnote 1). This strongly suggests that estimating your language switching frequency is difficult even when thinking about one specific person. Such estimates might be hindered by various factors, including people not always being aware of their switching and self-ratings being influenced by attitudes or stigmas.

4. Conclusion

Recent research suggested self-ratings of language use might be less influenced by participants' background (e.g., their language solidarity) than, for example, self-rated proficiency (Hernández-Rivera et al., 2024). Our findings provide further support for the use of self-reported language use by showing that language-use questionnaires (in particular the LSBQ, Anderson et al., 2018) might be reliable across time and form valid reflections of daily-life language use. The validity of language switching questionnaires (as evaluated in pilot-form in Study 2) might be lower, with language switching ratings potentially more strongly influenced by personal attitudes, questions being interpreted in different ways and bilinguals not always being aware of their own behaviours. Future work is therefore necessary to understand and improve the validity of currently used switching questionnaires and to further examine the use of short lab-based production tasks to potentially complement, and mitigate issues around, self-reported switching behaviours.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728926100996.

Acknowledgements. I would like to thank Sarah Gouveia, Marina Martin Maroto and Nerea Ramos Aguilera for their help with this project.

Funding statement. This project was funded by a *Language Learning* Early Career Research Grant.

Data availability statement. The data are available on https://osf.io/6bvk5/

Competing interests. The author declares none.

References

Anderson, J. A., **Mak, L.**, **Keyvani Chahi, A.**, & **Bialystok, E.** (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, **50**, 250–263.

Arndt, H. L., **Granfeldt, J.**, & **Gullberg, M.** (2023). Reviewing the potential of the experience sampling method (ESM) for capturing second language exposure and use. *Second Language Research*, **39**(1), 39–58.

Bonfieni, M., **Branigan, H. P.**, **Pickering, M. J.**, & **Sorace, A.** (2019). Language experience modulates bilingual language control: The effect of proficiency, age of acquisition, and exposure on language switching. *Acta Psychologica*, **193**, 160–170.

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition*, **16**(1), 32–48.

Coumel, M., **Liu, C.**, **Trenkic, D.**, & **de Bruin, A.** (in press). How do changes in language environment modulate bilingual language switching in production and comprehension? *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Cox, J. G., **LaBoda, A.**, & **Mendes, N.** (2020). I'm gonna Spanglish it on you": Self-reported vs. oral production of Spanish–English codeswitching. *Bilingualism: Language and Cognition*, **23**(2), 446–458.

de Bruin, A., & **Martin, C. D.** (2022). Perro or txakur? Bilingual language choice during production is influenced by personal preferences and external primes. *Cognition*, **222**, 104995.

de Bruin, A., **Samuel, A. G.**, & **Duñabeitia, J. A.** (2018). Voluntary language switching: When and why do bilinguals switch between their languages? *Journal of Memory and Language*, **103**, 28–43.

de Bruin, A., & **Xu, T.** (2023). Language switching in different contexts and modalities: Response-stimulus interval influences cued-naming but not voluntary-naming or comprehension language-switching costs. *Bilingualism: Language and Cognition*, **26**(2), 402–415.

Dewaele, J. M., & **Wei, L.** (2014). Attitudes towards code-switching among adult mono-and multilingual language users. *Journal of Multilingual and Multicultural Development*, **35**(3), 235–251.

Fricke, M., & **Kootstra, G. J.** (2016). Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, **91**, 181–201.

Furnham, A., & **Henderson, M.** (1982). The good, the bad and the mad: Response bias in self-report measures. *Personality and Individual Differences*, **3**(3), 311–320.

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**(3), 640–665.

Gollan, T. H., **Kleinman, D.**, & **Wierenga, C. E.** (2014). What's easier: Doing what you want, or being told what to do? Cued versus voluntary language and task switching. *Journal of Experimental Psychology: General*, **143**(6), 2167–2195.

Green, D. W., & **Abutalebi, J.** (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, **25**(5), 515–530.

Gullifer, J. W., & **Titone, D.** (2021). Engaging proactive control: Influences of diverse language experiences using insights from machine learning. *Journal of Experimental Psychology: General*, **150**(3), 414–430.

Hartanto, A., & **Yang, H.** (2016). Disparate bilingual experiences modulate task-switching advantages: A diffusion-model analysis of the effects of interactional context on switch costs. *Cognition*, **150**, 10–19.

Hartanto, A., & **Yang, H.** (2020). The role of bilingual interactional contexts in predicting interindividual variability in executive functions: A latent variable analysis. *Journal of Experimental Psychology: General*, **149**(4), 609–633.

Hernández-Rivera, E., **Kalogeris, A.**, **Tiv, M.**, & **Titone, D.** (2024). Self-evaluations and the language of the beholder: Objective performance and language solidarity predict L2 and L1 self-evaluations in bilingual adults. *Cognitive Research: Principles and Implications*, **9**(1), 75.

Hofweber, J., **Marinis, T.**, & **Treffers-Daller, J.** (2016). Effects of dense code-switching on executive control. *Linguistic Approaches to Bilingualism*, **6**(5), 648–668.

Hofweber, J., **Marinis, T.**, & **Treffers-Daller, J.** (2018). Predicting executive functions in bilinguals using ecologically valid measures of code-switching

behavior. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language. The state of the science across its subfields. Studies in bilingualism (54)* (pp. 181–205). John Benjamins.

JASP Team (2024). JASP (Version 0.19.3)[Computer software].

Jylkkä, J., **Soveri, A.**, **Laine, M.**, & **Lehtonen, M.** (2020). Assessing bilingual language switching behavior with ecological momentary assessment. *Bilingualism: Language and Cognition*, **23**(2), 309–322.

Kałamała, P., **Szewczyk, J.**, **Chuderski, A.**, **Senderecka, M.**, & **Wodniecka, Z.** (2020). Patterns of bilingual language use and response inhibition: A test of the adaptive control hypothesis. *Cognition*, **204**, 104373.

Lai, G., & **O'Brien, B. A.** (2020). Examining language switching and cognitive control through the adaptive control hypothesis. *Frontiers in Psychology*, **11**, 1171.

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.

Li, P., **Sepanski, S.**, & **Zhao, X.** (2006). Language history questionnaire: A web-based interface for bilingual research. *Behavior Research Methods*, **38**(2), 202–210.

Li, P., **Zhang, F.**, **Yu, A.**, & **Zhao, X.** (2020). Language history questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, **23**(5), 938–944.

MacIntyre, P. D., **Noels, K. A.**, & **Clément, R.** (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, **47**(2), 265–287.

Mann, A., & **de Bruin, A.** (2022). Bilingual language use is context dependent: Using the language and social background questionnaire to assess language experiences and test-rest reliability. *International Journal of Bilingual Education and Bilingualism*, **25**(8), 2886–2901.

Marian, V., **Blumenfeld, H. K.**, & **Kaushanskaya, M.** (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, **50**(4), 940–967.

Okamoto, K., **Ohsuka, K.**, **Shiraishi, T.**, **Hukazawa, E.**, **Wakasugi, S.**, & **Furuta, K.** (2002). Comparability of epidemiological information between self-and interviewer-administered questionnaires. *Journal of Clinical Epidemiology*, **55**(5), 505–511.

Rodriguez-Fornells, A., **Krämer, U. M.**, **Lorenzo-Seva, U.**, **Festman, J.**, & **Münte, T. F.** (2012). Self-assessment of individual differences in language switching. *Frontiers in Psychology*, **2**, 388.

Shiffman, S., **Stone, A. A.**, & **Hufford, M. R.** (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology.*, **4**(1), 1–32.

Surrain, S., & **Luk, G.** (2019). Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism: Language and Cognition*, **22**(2), 401–415.

Tomoschuk, B., **Ferreira, V. S.**, & **Gollan, T. H.** (2019). When a seven is not a seven: Self- ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition*, **22**(3), 516–536.

Uchihara, T., & **Clenton, J.** (2023). The role of spoken vocabulary knowledge in second language speaking proficiency. *The Language Learning Journal*, **51**(3), 376–393.

https://doi.org/10.1017/S1366728926100996 Published online by Cambridge University Press