Bilingualism: Language and Cognition

cambridge.org/bil

Assessing bilingual language switching behavior with Ecological Momentary Assessment

Jussi Jylkkä[1], Anna Soveri[1], Matti Laine[1,2] and Minna Lehtonen[1,3,4]

[1]Department of Psychology, Abo Akademi University, Finland; [2]Turku Brain and Mind Center, University of Turku, Finland; [3]Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Finland and [4]MultiLing Center for Multilingualism in Society across the Lifespan, Department of Linguistics and Scandinavian studies, University of Oslo, Oslo, Norway

Research Article

Cite this article: Jylkkä J, Soveri A, Laine M, Lehtonen M (2020). Assessing bilingual language switching behavior with Ecological Momentary Assessment. *Bilingualism: Language and Cognition* **23**, 309–322. https://doi.org/10.1017/S1366728918001190

Received: 13 March 2018
Revised: 26 November 2018
Accepted: 10 December 2018
First published online: 1 April 2019

Keywords:
bilingual executive advantage; language switching frequency; executive functions; Ecological Momentary Assessment; linear mixed effects model; multilevel modeling

Address for correspondence: Jussi Jylkkä, E-mail: jjylkka@abo.fi

Abstract

The putative bilingual executive advantage has been argued to stem from lifelong experience with executively demanding language behaviors, such as switching between the two languages. However, studies testing for possible associations between language switching frequency and EF in bilinguals have yielded inconsistent results. One reason for this could lie in the methods used that have evaluated the frequency and type of language switches with retrospective self-reports, as well as in problems in reliability and convergent validity of the executive tasks. By using Ecological Momentary Assessment (EMA) as a reference point for self-reports of language switches, we examined the validity of general retrospective self-reports of language switching. Additionally, we examined associations between language switching and EF using multilevel models. Our results indicated that the commonly used retrospective self-reports of language switching may lack convergent validity. However, we found tentative evidence that contextual language switches, assessed with EMA, may be associated with better inhibitory control, set shifting, and working memory.

1. Introduction

It has been suggested that bilinguals outperform monolinguals in executive functions (EF; e.g., Bialystok, 2009), although this putative advantage has been questioned in recent studies (de Bruin, Treccani & Della Sala, 2015; Lehtonen, Soveri, Laine, Järvenpää, de Bruin & Antfolk, 2018; Paap, Johnson & Sawi, 2016). This advantage is assumed to stem from bilinguals' lifelong language use, which is thought to engage and train domain-general EF through inhibition of the non-target language, switching between the two languages, and monitoring of the activation levels of the two languages (e.g., Linck, Schwieter & Sunderman, 2012; Rodriguez-Fornells, De Diego Balaguer & Münte, 2006). This Bilingual Training hypothesis also implies that bilinguals who switch more should exhibit better EF. The results from previous studies investigating the associations between language switching frequency and EF, however, have been inconsistent, as some studies have found associations between higher rates of language switching and better performance on some aspect of EF (Hartanto & Yang, 2016; Prior & Gollan, 2011; Soveri, Rodriguez-Fornells & Laine, 2011; Verreyt, Woumans, Vandelanotte, Szmalec & Duyck, 2016), while others have failed to find such associations (Johnson, Sawi & Paap, 2015; Jylkkä, Soveri, Wahlström, Lehtonen, Rodríguez-Fornells & Laine, 2017; Paap, Myuz, Anders, Bockelman, Mikulinsky & Sawi, 2017; Yim & Bialystok, 2012). This disagreement between previous studies may stem from various sources, including problems in the measurement of language switching (e.g., failure to give an accurate account retrospectively), problems in the convergent validity or reliability of the tests of executive functions (henceforth simply "executive tests"; Paap & Sawi, 2014), or simply non-existence of an association between language switching frequency and EF. In the present study, we mainly focused on the validity of self-reports of language switching. Additionally, we examined the convergent validity and reliability of executive tests.

Most previous studies investigating the possible association between language switching and EF have employed a single question to assess switching frequency – e.g., "How often are you in a situation in which you switch between languages?" (Verreyt et al., 2016); see also Johnson et al., 2015; Prior & Gollan, 2011) – while two studies (Jylkkä et al., 2017; Soveri et al., 2011) have utilized the Bilingual Switching Questionnaire (BSWQ) by Rodriguez-Fornells, Krämer, Lorenzo-Seva, Festman, and Münte (2012). The BSWQ is a 12-item questionnaire designed to assess language switching in bilinguals, and it has evidenced good psychometric properties when tested in a large sample of 556 Spanish–Catalan bilinguals (ibid.). However, neither the BSWQ nor the single questions have been tested for their ecological validity: that is, how well they correspond to the actual switching behavior of the

© Cambridge University Press 2019

participants. Ecological validity testing is important especially since the previous switching measures rely on retrospection, which may be vulnerable to errors and biases (e.g., Podsakoff, MacKenzie, Lee & Podsakoff, 2003). The present study is the first one to address this issue.

Our primary goal in the present study was to examine the convergent validity of both single switching questions and the BSWQ, by comparing the results from those instruments with language switching frequency reports gathered using Ecological Momentary Assessment (EMA; e.g., Shiffman, Stone & Hufford, 2008), where participants reported their language switching frequency every two hours for two weeks using a smartphone application. These data should provide a more accurate evaluation of everyday language switching than the previously used questions that are typically administered only once and require subjects to estimate their language switching during unspecified periods in the past. We also examined the test-retest reliability of the previously employed language switching questions.

Our secondary goal was to assess the associations between language switching (assessed with EMA) and EF (measured with four online executive tasks; Flanker and Simon tasks for inhibition or selective attention, number-letter task for set shifting, and an n-back task for working memory). Finally, we also examined the convergent validity and split-half reliability of the executive tasks, because previous research indicates problems in both (Paap & Sawi, 2014; Soveri, Lehtonen, Karlsson, Lukasik, Antfolk & Laine, 2016).

The executive cost effects and their associations with language switching were analyzed with multilevel methods (lmer in R). The choice of analysis was based on two main considerations. First, multilevel models are more powerful, because they enable the analysis of executive cost effects using raw trial-level data instead of means. Second, multilevel methods enable taking into account possible learning effects within the executive tasks by including trial number as a random effect.

It is important to note that the EMA approach does not yield objective data of language switching frequency either, as it is based on self-reports covering the past two hours. In contrast, Yim and Bialystok (2012) recorded the number of language switches during a structured interview that lasted on average 90 seconds. They found medium-sized associations between the frequency of language switching and performance in a verbal category fluency task where language switching was also required, but no significant associations with non-verbal (general executive) switching tasks. However, the sample of switching behavior that they collected was short and not necessarily representative of the participants' general language switching behavior, which could underlie their null findings. The EMA approach has the advantage of providing a longer and more representative sample of language switching behavior.

2. Method

2.1. Participants

The participants ($N = 30$) were neurologically healthy early balanced Finnish–Swedish bilinguals recruited through e-mail lists at the Åbo Akademi University and University of Helsinki. Sample size was determined based on effect sizes found in earlier studies that have assessed the validity of general retrospective questionnaires against EMA. In these studies, correlations have ranged between .61 and .89 – e.g., recall vs. diary assessment of the frequency of six different incontinence symptoms, mean $r = .81$ (Homma, Ando, Yoshida, Kageyama, Takei, Kimoto, Ishizuka, Gotoh & Hashimoto, 2002); positive and negative affect in a sample of ten adolescents (aged 15 to 18 years), mean Spearman $\rho = .89$ (Shrier, Shih & Beardslee, 2005); negative affect and eating disorder behaviors in anorexic, bulimic, and obese samples, mean Spearman $\rho = .61$ for negative affect and .61 for eating disorder behaviors (Wonderlich, Lavender, Wonderlich, Peterson, Crow, Engel, Le Grange, Mitchell & Crosby, 2015). With roughly 30 participants we had a statistical power of .80 to discover moderate validity of $r = .50$, which could be reasonably expected based on previous EMA research, albeit it stems from other research domains.

Key participant characteristics are summarized in Table 1. All participants had acquired both Finnish (L1) and Swedish (L2) and had used both languages at home before the age of seven. Their self-reported proficiency in Finnish vs. Swedish did not differ in other modalities ($|Z|$'s < 1.52, p's $> .1$) than in writing, which was higher for Swedish ($|Z| = 2.99$, $p = .003$). The participants received two movie tickets as compensation after they had performed all the parts of the study.

2.2. Procedure

The study was approved by the Joint Ethics Review Board of the Departments of Psychology and Logopedics at the Åbo Akademi University. The study began with a meeting (if possible) or a telephone call with the participant, to make sure they knew the requirements of the study, and to increase their commitment and thus minimize dropout. This also gave the participants an opportunity to ask questions. After this, the participants gave their written informed consent and filled in the background questionnaires that probed, among other things, their language background and possible psychiatric or neurological conditions. No participants had any conditions that would have led to their exclusion. All participants also filled in the language switching questions described below. After this, the participants received a link to a website where they could perform the EF tasks at a time of their choice. The EF tasks took approximately 1–1.5 hours to complete. The participants also received an invitation to install the EMA application on their smartphone. The EMA data collection of language switching behavior took two weeks (described in more detail below). After the two-week EMA period, the subjects filled in the language switching questions for the second time. Moreover, they estimated on a 1–5 scale how typical their language use was during the EMA period (1: highly exceptional; 5: highly typical).

The Bilingual Switching questionnaire

The BSWQ (Rodriguez-Fornells et al., 2012) measures four factors with three questions per factor: (1) Switches into L1 (BSWQ-L1S; e.g., "When I cannot recall a word in L2, I tend to immediately produce it in L1"), (2) Switches into L2 (BSWQ-L2S; same questions as in L1S but with languages inverted), (3) Contextual Switches (BSWQ-Contextual Switches; e.g., "There are situations in which I always switch between the two languages"), and (4) Unintended Switches (BSWQ-Unintended Switches; e.g., "It is difficult for me to control the language switches I introduce during a conversation"). All responses were given on a 5-point scale from "completely disagree" to "completely agree" and coded so that a higher score indicated higher switching frequency. The score on each factor was the sum of

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

Table 1. Key participant characteristics

	L1		L2	
	Mean (SD)	Range	Mean (SD)	Range
Age	25.1 (5.02)	20–38	–	–
Age of acquisition	0.84 (1.08)	0–4	1.34 (1.65)	0–5
Self-ratings of ability (1–7)				
Reading	6.75 (.70)	4–7	6.93 (.27)	6–7
Writing	6.21 (.92)	4–7	6.82 (.39)	6–7
Speaking	6.68 (.61)	5–7	6.89 (.31)	6–7
Listening	6.89 (.31)	6–7	6.93 (.26)	6–7
Overall	6.79 (.41)	6–7	6.93 (.26)	6–7

the three individual questions tapping on that factor; the range for each factor was thus 3–15. In the present study, we followed the procedure employed by Soveri et al. (2011), and combined the L1S and L2S factors into a general language switching factor BSWQ-Language Switches, representing the mean of L1S and L2S. BSWQ does not specify the time period from which language use is evaluated (Rodriguez-Fornells et al., 2012).

The single language switching questions

The single language switching questions assessed language switching over an undefined time span in retrospect without separating different types of switch. Our aim was that the questions would be similar to those used in earlier studies (Johnson et al., 2015; Prior & Gollan, 2011; Verreyt et al., 2016). We utilized two questions concerning switching on average: "On average I switch between languages during a day [frequency range]" (Single Question-Average Switching Frequency), and "On average I make many brief language switches during a day" [agree-disagree] (Single Question-Many Brief Switches). For the first question, responses were given on a five-point scale classifying the frequency of switches as follows: "0–2", "3–5", "6–10", "11–20", and "over 20 times". For the second question, a five-point scale from "completely disagree" to "completely agree" was employed. Both questions were recoded into scales from one to five for analysis, where one represents the frequency range "0–2" or "completely disagree".

The Ecological Momentary Assessment

Three language switching questions were used in the EMA application: frequency of intended switches (EMA-Intended Switches), frequency of unintended switches (EMA-Unintended Switches), and percentage of contextual switches (EMA-Contextual Switches; see Table 2). These were based on our interpretation of what the BSWQ factors intend to measure. Additionally, we included a question about switches between writing and speech, but it was omitted from analyses because only spoken language switches have been assessed in earlier studies.[1] The questions were introduced to the participants in the first meeting. They were given written instructions and examples on how to answer them, as well as what was counted as a language switch. We defined a language switch as borrowing a word from another language during a conversation, switching the conversation language completely, or engaging in two consecutive discussions in different languages with a maximum interval of roughly 5 minutes. We instructed the participants not to count as language switches expressions from the other language that were highly commonplace and could be considered as part of the dialect.

After the first meeting the participants installed the EMA application (MetricWire™) on their iOS or Android smartphone via a link sent to them in an e-mail. The application asked the participants to answer four questions about their language use six times a day between 9am and 9pm. The questions appeared on roughly two-hour intervals. The application sent the responses directly to a server, where they were available for download by the experimenter. The EMA period started after the participants had installed the application on their mobile phone. After 14 whole days, the participants were notified that the application could now be uninstalled. The participants gave on average 4.75 responses a day ($SD = .96$, range = 1.40–5.87). They judged that their language use during the EMA period was highly typical compared to their language use in general ($M = 4.41$ on a 1–5 scale, $SD = .57$, range = 3–5). Typicality ratings were missing for three participants. We summarize the average language switching ratings from the EMA period, the BSWQ, and the individual questions, in Table 3.

In the second meeting that took place after the EMA period, we interviewed the subjects on how easy they considered the questions were to answer. These interviews were not transcribed or coded, but overall the participants considered it easy to answer the questions.

The executive tasks

The executive tasks were presented in randomized order. The Simon task can be considered as a measure of inhibition or conflict resolution (Simon & Rudell, 1967). In the task, the participant has to categorize the color (blue or red) of the stimulus by pressing either a left (for blue) or right (for red) button press. The stimulus can appear on either side of the screen. On congruent trials, the response button is on the same side as the stimulus (e.g., red stimulus on the right side of the screen), whereas on incongruent trials the stimulus appears on the side opposite to the correct response button (e.g., red stimulus on the left side). It is hypothesized that on incongruent trials, the participant needs to inhibit the irrelevant information about the spatial location of the stimulus. The Simon effect is the difference in reaction time (RT) or accuracy between congruent and incongruent trials; a larger difference is assumed to reflect worse inhibitory skills. The present version of the test consisted of 100 trials, half congruent and half incongruent. Each trial began with a fixation cross (800 ms) followed by a 250 ms blank interval. After that, the stimulus appeared and remained on the screen for 1 000 ms unless a response was given. Finally, the screen was blank for 500 ms. The stimuli were presented in four blocks with 5-second intervals in-between.

The Flanker task (adapted from Eriksen & Eriksen, 1974) can be considered as a measure of inhibition or selective attention. In

[1]The variable addressing switches between writing and speech was originally included to attain a more comprehensive estimate of language switching behavior. We did not intend to use this variable to validate the BSWQ or single questions, which only concern with spoken switches, but instead planned to include it in the regression analyses on the associations between language switching and EF. However, in our sample of 30, including all the four EMA-predictors would have resulted in an overfitted model. Thus, we omitted the writing and speech variable, because it is the only one that has not been used in earlier studies.

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

Table 2. The language switching questions in the EMA period

Question Name	Abbreviation	Question	Scale	Re-coding
EMA-Intended Switches	EMA-IS	"After your last answer (or during the last max two hours), how many times did you intentionally switch between languages in your speech?"	0, 1–3, 4–6, 7 or more often	0 = 1; 1–3 = 2; 4–6 = 3; 7 or more often = 4
EMA-Unintended Switches	EMA-US	"After your last answer (or during the last max two hours), how many times did you unintentionally switch between languages in your speech?"	0, 1–3, 4–6, 7 or more often	0 = 1; 1–3 = 2; 4–6 = 3; 7 or more often = 4
EMA-Contextual Switches	EMA-CS	"How large proportion of the aforementioned language switches took place in situations in which you typically always switch between languages?"	0–25%, 25–50%, 50–75%, 75–100%	0–25% = .125; 25–50% = .375; 50–75% = .625; 70–100% = .85

Table 3. Language switching ratings from the EMA period, and answers to BSWQ and the single questions before and after the EMA period. The EMA ratings are two-hour averages for the the whole period.

EMA ratings	M (SD)	range
IS	1.71 (.33)	1.22–2.45
US	1.40 (.43)	1.01–2.73
CS	.46 (.17)	.19–.85
Pre-EMA BSWQ		
LS	8.17 (1.64)	5–11
US	5.62 (2.09)	3–11
CS	9.45 (1.90)	5–13
Post-EMA BSWQ		
LS	8.45 (1.39)	5.5–11
US	6.27 (2.21)	3–12
CS	10.37 (2.06)	7–14
Pre-EMA simple questions		
ASF	3.10 (.98)	2–5
MBS	4.10 (.94)	2–5
Post-EMA simple questions		
ASF	2.93 (.83)	2–5
MBS	4.03 (1.13)	1–5

this task, the participant is presented with an array of five arrows, and is instructed to categorize the direction of the central arrow. On congruent trials, all arrows point in the same direction (e.g., >>>>>), whereas on incongruent trials the central arrow points in the opposite direction than the "flankers" (e.g., >><>>). The flanker effect is the difference between congruent and incongruent trials. A larger difference reflects worse performance. In the present variant of the task, there were 100 trials, of which half were congruent and half incongruent. Each trial began with a fixation cross (800 ms), followed by the stimulus that remained on the screen for 800 ms unless a response was given, and finally a blank screen (500 ms). The stimuli were presented in four blocks with 5-second intervals in-between.

It can be argued that the simple congruency effect in the Simon and Flanker tasks does not necessarily reflect executive processes, and that sequential congruency effects (e.g., the Gratton effect; Gratton, Coles & Donchin, 1992) are more central measures in this respect (Botvinick, Braver, Barch, Carter & Cohen, 2001). Hence, in these tasks we also analyzed the Gratton effect, which states that congruency effects may be larger after congruent trials than after incongruent trials. Hypothetically, this is because conflict monitoring has been activated during incongruent trials, diminishing possible congruency effects during the following trial (see also Kerns, 2006; Siemann, Herrmann & Galashan, 2018 with regard to serial congruency effects in the Simon and Flanker tasks, respectively).

The number-letter task (adapted from Rogers & Monsell, 1995) is assumed to tap on non-verbal task switching (or flexibility, or mental set shifting). We used this task instead of a language switching task because we wanted to examine whether everyday language switching is associated with general (non-lexical) task switching. In the number-letter task, the participant is presented with a number-letter pair (e.g., A3) in one of two vertically aligned boxes. The participant is instructed to categorize the number as odd (response button one) or even (response button two) if the pair appears in the upper box, and the letter as vowel (response button one) or consonant (response button two) if the pair appears in the lower box. The present version of the task consisted of two single task blocks that included 32 trials each (letter categorization only or number categorization only), and one mixed block (80 trials). In the mixed block, there were 32 switch trials and 47 repetition trials (the first trial is neither). On switch trials, the task (location of the stimulus) switches from the previous trial, whereas on repetition trials, the task (location) stays the same. Each trial began with a blank interval (150 ms), followed by a fixation cross (300 ms). After that, the stimulus appeared and remained on the screen for 3.000 ms unless a response was given. There was a short break between each block. The task yields two measures: a switch cost, which is the difference in RT or accuracy between switch and repetition trials, and a mixing cost, which is the difference between repetition trials and single block trials, generally assumed to tap on monitoring. In both measures, a higher difference indicates worse performance.

The spatial n-back task (Carlson, Martinkauppi, Rämä, Salli, Korvenoja & Aronen, 1998) is used as a measure of working memory updating and monitoring. In this task, a box appears in one of eight possible locations and the participant has to judge whether the location is the same as on the previous trial (1-back sequence) or two trials back (2-back sequence), depending on instructions presented on the screen before the sequence. There were two blocks of 80 trials, consisting of four sequences of 20 trials. There was a 15-second break between the blocks. Of the four sequences, two were 1-back and two were 2-back sequences. Each sequence consisted of six targets (stimulus location same as n trials ago) and 14 non-targets (stimulus location different than n trials ago). At the beginning of each sequence,

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

the number "1" or "2" appeared on the screen, indicating whether the sequence was a 1- or 2-back sequence. The square appeared on the screen in one of eight locations for 100 ms, followed by a blank interval of 3 000 ms irrespective of whether a response was given. The task yields the n-back effect, which is the difference in RT or accuracy between 1-back and 2-back trials. A larger difference indicates worse performance.[2]

3. Results

Background information was missing from one participant, including the pre- and post BSWQs and single switching questions. This participant was thus omitted from the test-retest reliability and validity analyses, but was included in the analyses on the associations between language switching and EF.

3.1. Reliability of the BSWQ and the single switching questions

The test-retest reliability of the BSWQ and the individual language switching questions was assessed using Pearson (r) and intraclass (ICC) correlations, as well as the Smallest Real Difference (SRD). Bayes Factors (BF) were calculated as an estimate for strength of evidence for r.[3] Compared to Pearson's r, which tests for a linear association between the test and retest session (irrespectively of absolute differences), the ICC takes into account both changes in a participant's performance and systematic changes in group means (e.g., Vaz, Falkmer, Passmore, Parsons & Andreou, 2013). The SRD can be considered as a confidence interval for a difference between two testing sessions: it specifies the value under which the mean difference lies with 95% probability (e.g., Lexell & Downham, 2005). Here we report only Pearson's r, but all three reliability measures are presented in Table 4.

The results revealed that the test-retest reliabilities were high for BSWQ-Unintended Switches ($r = .80$, 95% CI [.62, .90], $p < .001$), marginal for BSWQ-Language Switches ($r = .62$, 95% CI [.32, .80], $p < .001$), and low for BSWQ-Contextual Switches ($r = .27$, 95% CI [−.11, .58], $p > .1$). The test-retest reliabilities were low for Single Question-Average Switching Frequency ($r = .40$, 95% CI [.040, .67], $p < .05$) and Single Question-Many Brief Switches ($r = .53$, 95% CI [.20, .75], $p < .01$).[4] In terms of the Bayes Factors, evidence was very strong for the observed correlation in the case of BSWQ-Unintended Switches and Language Switches ($BF_{10} = 11 \times 10^4$ and $BF_{10} = 95$, respectively), strong for Single Question-Many Brief Switches ($BF_{10} = 14$), but anecdotal for the other questions (BF_{10}'s < 2.2).

3.2. Convergent validity of the BSWQ and single questions

The convergent validity of the BSWQ and the single questions administered before the EMA period, against mean ratings in EMA, were assessed using Pearson correlations (see Table 5). Pre-EMA data were used to rule out possible effects of the EMA period on the ratings. It is possible that the subjects would have been more accurate in their retrospective assessments after having paid close attention to their language switching for two weeks.

We expected that the EMA-Intended Switches, EMA-Unintended Switches, and EMA-Contextual Switches would correlate with the BSWQ-Language Switches, BSWQ-Unintended Switches, and BSWQ-Contextual Switches, respectively. We did not have specific hypotheses about what the single questions would preferentially correlate with, and thus we correlated them against all the EMA questions.

Of the BSWQ factors, BSWQ-Unintended Switches showed a high correlation with EMA-Unintended Switches ($r = .63$, 95% CI [.35, .81], $p < .001$), while the other correlations were low (see Table 5). Of the single switching questions, Single Question-Average Switching Frequency correlated moderately with EMA-Intended Switches ($r = .42$, 95% CI [.059, .68], $p < .05$) and EMA-Contextual Switches ($r = .39$, 95% CI [.030, .66], $p < .05$). The correlations between the Single Question-Many Brief Switches and the EMA questions were very low (see Table 5). In terms of the Bayes Factors, there was extremely strong evidence for the observed correlation between BSWQ-Unintended Switches and EMA-Unintended Switches ($BF_{10} = 145$), but at most anecdotal evidence for the correlations between other factors or questions (BF_{10}'s < 2.6).[5]

3.3. Reliability and convergent validity of the executive tasks

In all EF tasks, a subject's performance was excluded if the overall accuracy was below chance level (at alpha = .05 overall accuracy .58). Based on this criterion, three subjects in the n-back task, and one subject in the number-letter task were excluded from the analyses. No subjects were excluded in the Simon and Flanker tasks.

The executive cost effects were statistically significant and in the expected direction: the Simon effect was 33.10 ms ($t = 6.00$, $p < .001$); the Flanker effect was 58.40 ms ($t = 13.38$, $p < .001$); the N-back effect was 77.56 ms ($t = 2.97$, $p = .006$); and the number-letter switch cost was 306.48 ms ($t = 10.98$, $p < .001$) and mixing cost 165.55 ms ($t = 7.03$, $p < .001$).

We examined the split-half (odd or even trials) reliability of the executive tasks with Pearson's r and ICC. The split-half reliabilities are summarized in Table 6, which indicates that the split-half reliability was low (r's < .60; cf. Strauss, Sherman & Spreen, 2006) for all the tasks, except for the number-letter task switch and mixing cost, both of which having an adequate reliability (r's > .70). In terms of Bayes Factors, evidence for these last mentioned correlations was decisive (BF_{10}'s > 100), whereas evidence for the other correlations was moderate at best BF_{10}'s < 3.4).

Convergent validity was only examined between the Simon and Flanker tasks, which were both assumed to tap on inhibition

[2]The executive tasks were identical to those used in Jylkkä et al. (2017), where we also briefly review differences between mono- and bilinguals on these tasks.

[3]The Bayes factor represents evidence for the alternative hypothesis relative to evidence for the null hypothesis. A BF below 1 is evidence for the null hypothesis; 17–3 is typically considered as anecdotal; 3–10 moderate; 10–30 strong; 30–100 very strong; and > 100 extreme evidence for the alternative hypothesis (Jeffreys, 1961).

[4]We use the conservative terminology of Strauss, Sherman, and Spreen (2006) in describing the strength of the reliability correlations: low reliability $r < .60$; marginal $r = .60 - .69$; adequate $r = .70 - .79$; high $r = .80 - .89$, very high $r \geq .90$. With respect to validities, which can be expected to be lower, we use the more liberal classification of Cohen (1988): small $r = .10 - .30$; medium $r = .30 - .50$; large $r > .50$.

[5]Some of the participants responded comparatively seldom in the EMA period. To rule out the possibility that this could affect the correlations, we re-ran the analysis excluding those participants with less than three answers per day on average. This resulted in excluding two participants. The results stayed the same overall, with the only noteworthy ($p < .05$ or $BF_{10} > 3$) associations being between EMA-US and BSWQ-US ($r = .69$, $p < .001$, $BF_{10} = 567$), EMA-IS and Single Question-Average Switching Frequency (ASF; $r = .48$, $p = .012$, $BF_{10} = 4.96$), and EMA-CS and ASF ($r = .48$, $p = .011$, $BF_{10} = 5.11$).

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

Table 4. Reliability values for the BSWQ factors and the single questions [a]

Factor/question	Scale	*r*	95% CI	BF_{10}	Pre minus post (SD)	±SRD	$ICC_{2,1}$	95% CI
LS	3–15	.62	[.32, .80]	95	−.26 (1.35)	2,65	.61	[.32, .79]
US	3–15	.80	[.62, .90]	11×10^4	−.66 (1.37)	2,68	.77	[.53, .89]
CS	3–15	.27	[−.11, .58]	.58	−.90 (2.43)	4,75	.25	[−.093, .55]
ASF	1–5	.40	[.040, .67]	2.1	.17 (1.00)	1.96	.40	[.044, .66]
MBS	1–5	.53	[.20, .75]	14	.069 (1.03)	2.02	.52	[.20, .75]

[a] LS = Language Switches; US = Unintended Switches; CS = Contextual Switches; ASF = Single Question-Average Switching Frequency; MBS = Single Question-Many Brief Switches

Table 5. Pearson correlations between the EMA questions, and BSWQ-factors and Single Questions administered before the EMA period [a]

Factor/question	EMA-IS [95% CI]	BF_{10}	EMA-US [95% CI]	BF_{10}	EMA-CS [95% CI]	BF_{10}
The pre-EMA period BSWQ factors						
BSWQ-LS	−.14 [−.48, .24]	.30	.15 [−.23, .49]	.31	−.33 [−.62, .046]	.95
BSWQ-US	−.042 [−.40, .33]	.24	.63*** [.35, .81]	145	−.032 [−.39, .34]	.23
BSWQ-CS	−.16 [−.50, .22]	.32	.25 [−.13, .57]	.53	−.29 [−.60, .081]	.73
The Simple Questions						
ASF	.42* [.059, .68]	2.56	−.018 [−.38, .35]	.23	.39* [.030, .66]	1.90
MBS	.014 [−.35, .38]	.23	.035 [−.34, .40]	.23	−.020 [−.38, .35]	.23

[a] IS = Intended Switches; LS = Language Switches; US = Unintended Switches; CS = Contextual Switches; ASF = Single Question-Average Switching Frequency; MBS = Single Question-Many Brief Switches

Table 6. Split-half reliabilities for the executive cost effects

	r	p	BF_{10}	95% CI	$ICC_{2,1}$	p	95% CI
Simon effect	.18	.33	.34	[−.18, .50]	.17	.17	[−.16, .47]
Flanker effect	.28	.13	.67	[−.081, .57]	.27	.061	[−.073, .56]
Number-letter switch cost	.72	<.001	6405	[.50, .85]	.70	<.001	[.48, .84]
Number-letter mixing cost	.74	<.001	13270	[.52, .86]	.69	<.001	[.46, .83]
N-back effect	.44	.018	3.35	[.084, .70]	.43	.012	[.067, .69]

(Nee, Wager & Jonides, 2007). The tasks did not correlate ($r = .24$, $p = .22$, $BF_{10} = .49$, 95% CI = [−.15, .56]).

Due to low reliability of the executive tasks, we suspected that learning effects may have occurred. Thus, we analyzed the basic cost effects in the tasks and their associations with language switching using multilevel models (linear mixed effects modelling), which can take into account variation between trials (or learning effects). Additionally, multilevel models are more powerful than mean-based methods since they take all variance in the test performance into account and can be used to analyze serial congruency effects as well.

3.4. Linear mixed effects models on the basic executive cost effects and their associations with language switching

All multilevel analyses were performed with the package lme4 in R, using simple coding for fixed factors. In simple coding, one of the factor levels is chosen as the baseline, against which the other factor levels are compared. Changing the baseline does not affect model fit statistics; it only enables examining different model estimates. In analyses where we were interested in main effects irrespectively of factor baseline, a Type III ANOVA was performed on the multilevel model using Satterthwaite's method for approximating degrees of freedom.

Analyses on the Simon task

Typically congruency effects, such as the Simon effect, are calculated by subtracting the mean RT of congruent trials from the mean RT of incongruent trials, but in linear mixed effects modelling this cannot be done because the method does not operate on means. Thus, the Simon effect was analyzed as the contrast between congruent and incongruent trials. We used a linear mixed effects model with RT as dependent variable, Congruency as fixed factor, and Participant and Trial as random effects. Incongruent trials were significantly slower than congruent trials ($E = 32.89$, $SE = 4.04$, $t = 8.15$, $p < .001$). Next we examined the effect of Trial by adding it as a continuous predictor instead of a random effect. In this model, the congruency effect was still significant although weaker than in the model without Trial as predictor ($E = 19.49$, $SE = 8.26$, $t = 2.36$, $p = .018$). There was a near-significant interaction between Congruency and Trial ($E = .27$, $SE = .14$, $t = 1.92$, $p = .055$), with the congruency

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

effect becoming weaker with higher Trial number. This was due to steeper decline in RT in the incongruent than congruent condition.

Next we examined the Gratton effect, assumed to represent a behavioral correlate of conflict monitoring, where the congruency effect is supposed to weaken after incongruent as opposed to congruent trials (Botvinick et al., 2001). This was examined in a model with Previous Congruency and Congruency as fixed factors, and Participant and Trial as random effects. In this model, the basic congruency effect was significant ($F = 66.05$, $p < .001$), but there was also a significant interaction between Previous Congruency and Congruency, i.e., a significant Gratton effect ($E = -68.90$, $SE = 8.02$, $t = -8.59$, $p < .001$). In other words, the congruency effect weakened when a trial was preceded by an incongruent trial.

The associations with the basic Simon congruency effect and the EMA variables were examined in a model with Congruency and the EMA variables as predictors of RT and Participant and Trial as random effects. None of the EMA variables interacted with Congruency (p's > .39), except for a near-significant interaction between Congruency and EMA-CS ($E = -63.17$, $SE = 36.10$, $t = -1.75$, $p = .080$; see Figure 1): the congruency effect was weaker the more contextual switches the participant made.

Next we examined the associations between the Gratton effect and the EMA variables in a model with Congruency, Previous Congruency, and the EMA variables as predictors, and with the same random effects. The three-way interactions between the EMA variables, Congruency, and Previous Congruency were not significant (p's > .097)

Analyses on the Flanker task

Similar models were used to examine basic Flanker congruency effects as in the Simon task. The congruency effect was strong ($E = 56.45$, $SE = 2.55$, $t = 22.13$, $p < .001$). In the model with Trial as predictor, the congruency effect was somewhat larger than in the basic model without Trial as predictor ($E = 64.66$, $SE = 5.17$, $t = 12.50$, $p < .001$), and there was a near-significant interaction between Trial and Congruency ($E = -.16$, $SE = .089$, $t = -1.75$, $p = .080$), with the congruency effect being smaller the higher the Trial number was. As in the case of the Simon task, this was due to steeper decline in RT in the incongruent than congruent condition.

Gratton effect in the Flanker task was examined similarly as in the Simon task. In the model with both Previous Congruency and Congruency as predictors, the basic Congruency effect was significant ($F = 490.09$, $p < .001$), but there was no Gratton effect ($E = -2.66$, $SE = 5.10$, $t = -.52$, $p = .60$).

Next we examined the associations between the Flanker congruency effect and the EMA variables. None of the associations were significant (p's > .16), except for the interaction between Congruency and EMA-CS ($E = -55.10$, $SE = 22.64$, $t = -2.43$, $p = .015$): the Flanker effect became smaller the more the participant reported Contextual Switches (Figure 2).

Analyses on the Number-letter task

Basic Number-letter switch and mixing costs were examined with a model with RT as dependent variable, Category (Single, Repetition, and Switch) as fixed factor, and Participant and Trial as random effects. Repetition trials were used as the baseline. There was a significant switch cost ($E = 270.68$, $SE = 16.19$, $t = 16.72$, $p < .001$) and mixing cost ($E = -159.07$, $SE = 13.58$, $t = -11.71$, $p < .001$), both in the expected direction.[6]

To examine possible learning effects, we used a model with Category as fixed factor, Trial as continuous predictor, and Participant as random effect. In this model, both switch cost ($p < .001$) and mixing cost ($p = .0035$) were significant and there was also an interaction between mixing cost and Trial ($E = -1.37$, $SE = .42$, $t = -3.29$, $p = .001$): the mixing cost became larger as Trial increased (Figure 3). This was mainly because responses in the Single block became faster with higher Trial ($p < .001$) but not in the Repetition trials ($p = .27$).

Next we examined the associations between the Switch and Mixing costs and the EMA language switching variables in a model with Category as fixed factor, the EMA variables as continuous predictors, and Trial and Participant as random effects. Switch cost was associated with both EMA-IS ($E = 119.75$, $SE = 46.60$, $t = 2.57$, $p = .010$) and EMA-CS ($E = -337.28$, $SE = 89.17$, $t = -3.78$, $p < .001$), but in opposite directions: higher IS predicted a larger switch cost and higher CS a smaller switch cost. The mixing cost was associated with EMA-US ($E = 81.24$, $SE = 21.66$, $t = 3.75$, $p < .001$): the more a participant reported US, the smaller their mixing cost was (Figure 4).

Because there was a large difference in the learning effect between the single and mixed blocks in the number-letter task, the mixing cost might be problematic as a theoretical construct. Thus, as a post hoc test we examined the associations between the EMA variables and performance over Block (Mixed or Single). We hypothesized that mixed block performance in the number-letter task, which is mainly a bottom-up driven shifting task (shifting of task is determined by the location of the stimulus), would be positively associated with US and CS, as both can be taken to reflect bottom-up driven language switching (accidental or context-driven switches), and negatively associated with IS which can be taken to reflect top-down driven language switching (a person shifts language at will). Higher frequency of EMA-IS was in fact associated with worse performance in the mixed block than in the single block ($E = -109.53$, $SE = 34.14$, $t = -3.21$, $p = .0013$), whereas both EMA-US ($E = 65.50$, $SE = 19.16$, $t = 3.42$, $p < .001$) and EMA-CS ($E = 202.77$, $SE = 65.41$, $t = 3.10$, $p = 0.0020$) were associated with better performance in the mixed compared to the single block (Figure 5).

Analyses on the N-back task

The n-back effect was examined in a model with RT as dependent variable, Condition (1-back vs. 2-back) as fixed factor, and Participant and Trial as random effects. The data was subsetted to include only match trials. Responses were significantly slower in the 2-back condition compared to 1-back ($E = 80.67$, $SE = 14.13$, $t = 5.71$, $p < .001$): that is, the n-back effect was significant. Next we included Trial as a continuous predictor in the model to examine learning effects. In this model, the n-back effect was significant ($E = 79.79$, $SE = 27.86$, $t = 2.86$, $p = .0043$), and there was no main effect of Trial ($p = .21$) or interaction between Condition and Trial ($p = .94$).

Associations between the n-back effect and the EMA language switching variables were examined similarly as in the other EF tasks. All the interactions were roughly around alpha level: higher

[6]Because in the model repetition trials were used as baseline, estimate for switch trials indicates how much slower responses were in the Switch condition, and estimate for Single indicates how much faster responses were in the Single block.

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

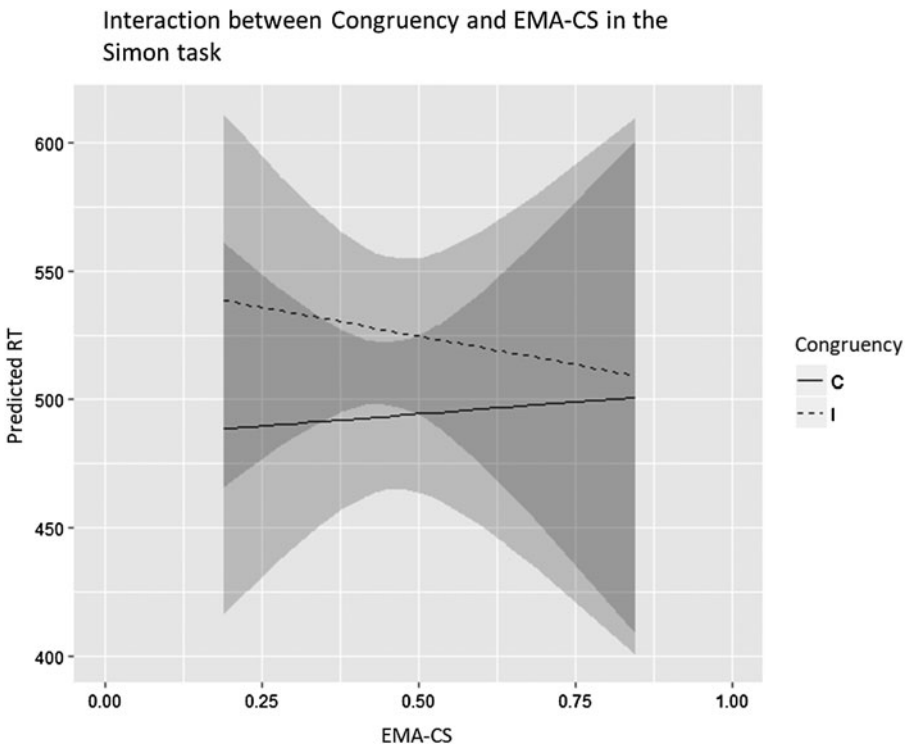


Fig. 1. Interaction between EMA-Contextual Switches and the Simon congruency effect. The Simon effect was smaller the more CS the participant reported.

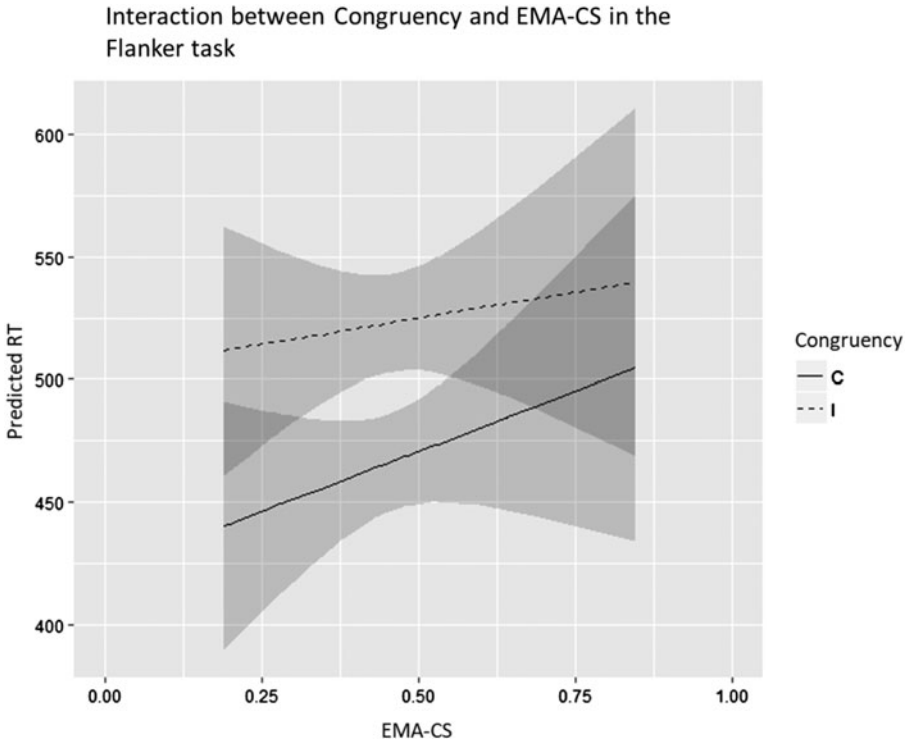


Fig. 2. Interaction between Congruency and EMA-Contextual Switches in the Flanker task. The Flanker effect became smaller when the participant reported more CS.

frequency of EMA-IS ($E = -117.75$, $SE = 64.45$, $t = -1.83$, $p = .068$) and EMA-CS effect (E = −212.76, SE = 119.82, t = −1.78, p = .076) were near-significantly associated with smaller n-back effect, whereas EMA-US was associated with larger n-back effect ($E = 70.76$, $SE = 35.80$, $t = 1.98$, $p = .048$; see Figure 6).

4. Discussion

It has been suggested that bilinguals outperform monolinguals on tasks measuring executive functions, assumedly because aspects of bilingual experience train executive functions. One proposed mechanism for this Bilingual Training has been language switching (Linck et al., 2012; Rodriguez-Fornells et al., 2006). Earlier

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

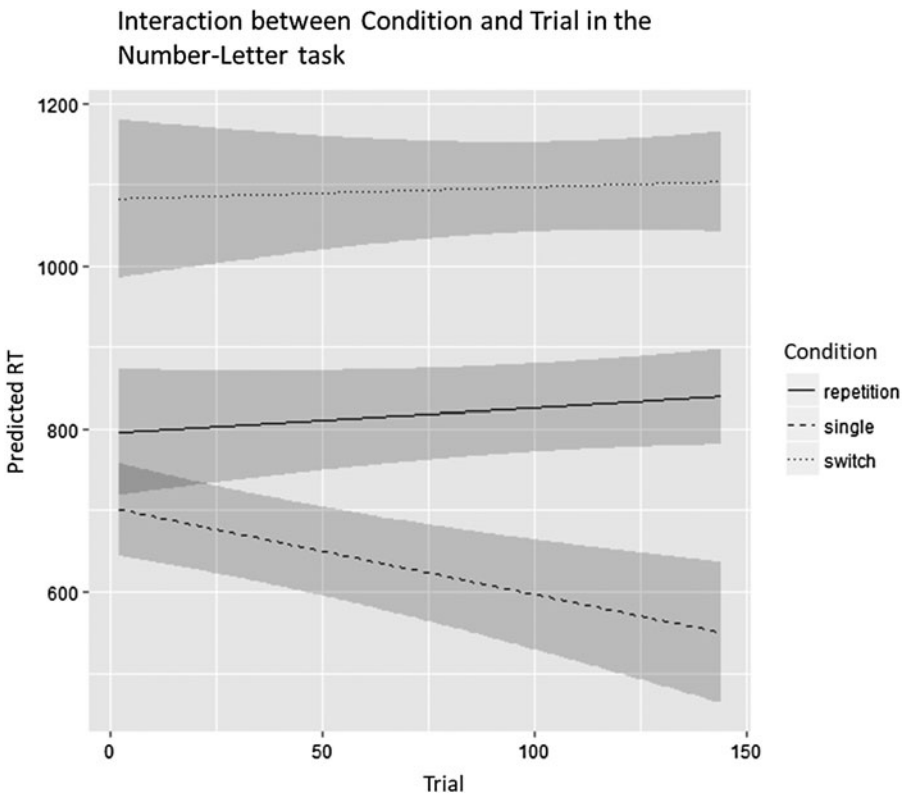


Fig. 3. Interaction between Category (repetition, switch, or single) and Trial Number in the Number-letter task. There was a strong learning effect in the single block but not in the other conditions.

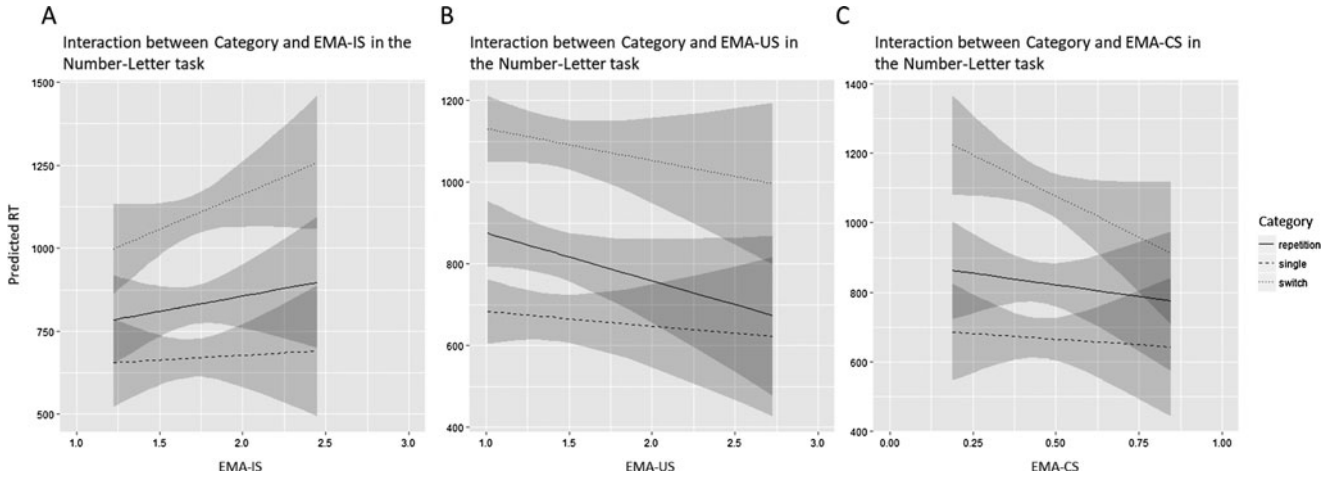


Fig. 4. Associations between the EMA-language switching variables and the Number-letter task, separated by Category (single, repetition, or switch). More frequent intended switching predicted larger number-letter switch cost (panel A); more unintended switching predicted smaller mixing cost (B); higher rate of contextual switches predicted smaller switch cost (C).

studies examining the relationship between language switching frequency and EF have, however, yielded inconsistent results (Hartanto & Yang, 2016; Johnson et al., 2015; Jylkkä et al., 2017; Paap et al., 2017; Prior & Gollan, 2011; Soveri et al., 2011; Verreyt et al., 2016). This discrepancy between previous studies suggests that the methods for assessing language switching and EF may be problematic, or that language switching does not train EF. To investigate this issue, we employed EMA of language switching to the single language switching questions and BSWQ, which have been employed in earlier studies. With its comprehensive coverage of language switching behavior in the participant's natural, everyday environment, EMA provides a reference point against which the previously employed methods could be compared. It is also worth pointing out here that the present participants estimated that their language use during the two-week EMA period was highly typical. To test the Bilingual Training hypothesis, we also examined whether language switching behavior, as measured with EMA, was associated with performance on EF tasks. Finally, we also examined the reliability and convergent validity of the executive tasks.

Overall, the results indicated that the convergent validity and test-retest reliability of the previously employed language

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

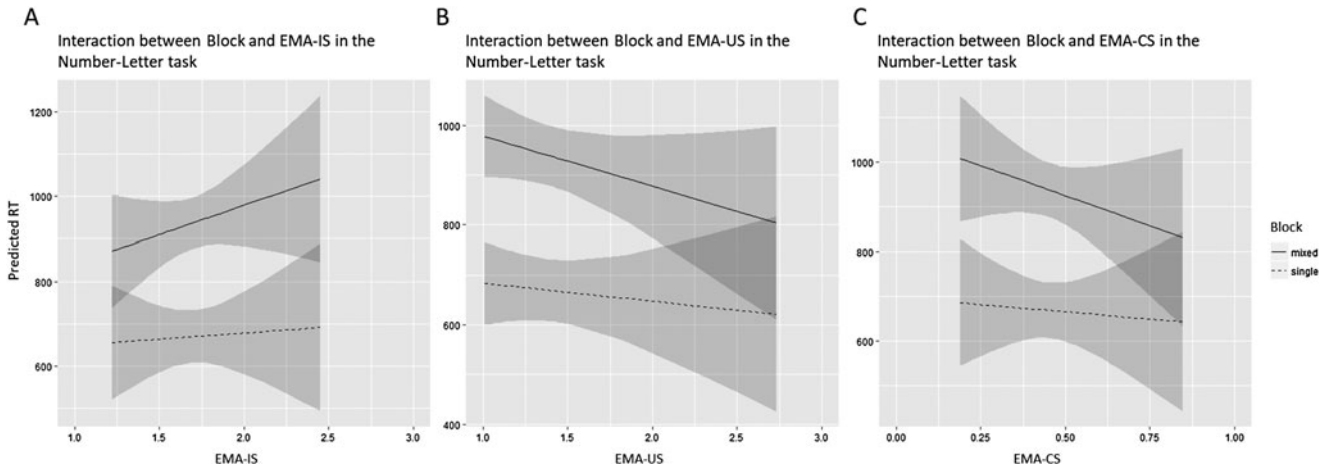


Fig. 5. Associations between the EMA-language switching variables and the Number-letter task, separated by Block (single or mixed). Higher rate of intended switches predicted larger RTs in the mixed block compared to the single block (panel A), whereas higher rate of unintended or contextual switching predicted smaller RTs in the mixed block compared to the single block (B and C).

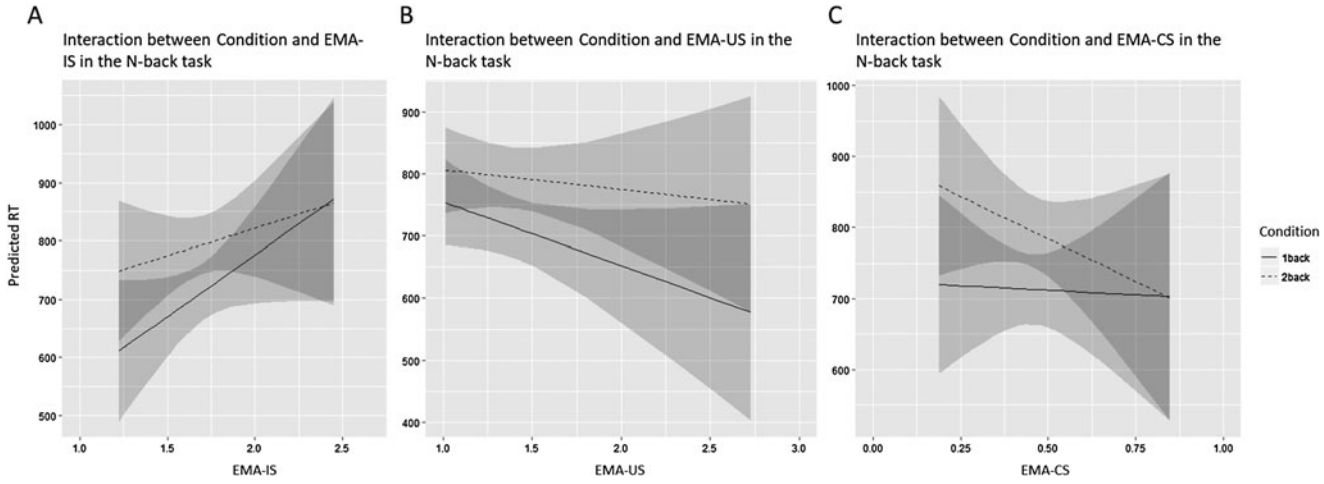


Fig. 6. Associations between the EMA-language switching variables and the n-back task. Higher rate of Intended and Contextual Switching predicted smaller n-back effect (A and C) whereas higher rate of Unintended Switching predicted larger n-back effect (B).

switching instruments were poor. As to the associations between the executive cost effects and language switching, most consistent findings concerned contextual language switching, which predicted better performance across all of the tasks.

4.1. Test-retest reliability of the BSWQ and the single questions

We examined the test-retest reliability of the switching questions employed in previous research with Pearson's *r*, ICC, and SRD. The results showed high test-retest reliability for BSWQ-Unintended Switches, marginal reliability for BSWQ-Language Switches, and low reliability for BSWQ-Contextual Switches and both single questions. Earlier studies on retrospective self-assessments of other behaviors than language switching report test-retest correlations between .67 and .89.[7] The present test-retest correlations were comparable to these studies only concerning BSWQ-Unintended Switches and BSWQ-Language Switches, indicating that the BSWQ-Contextual Switches factor and the single questions have lower test-retest reliability than other types of self-reports. The low reliability could be due to the questions being difficult to grasp, the complexity of the behavior itself, or simply memory failures. It is also possible that language switching shows more variation over time (for example depending on context) than other types of behavior.

4.2. Convergent validity of the BSWQ and the single questions

We assessed the convergent validity of the language switching questions employed in previous studies (BSWQ and single questions) by comparing them to the EMA variables. Regarding the BSWQ, the results showed high convergent validity only for the BSWQ-Unintended Switches. Of the single questions, Average

[7]E.g., eating disorder behaviors: test-retest *r*'s = .81 – .94 (mean *r* = .89) on the Eating Disorder Examination Questionnaire (Luce & Crowther, 1999); alcohol use: *r*'s = .64 – .92 (mean *r* = 78) in the Alcohol Use Disorders Identification Test (Reinert & Allen, 2002); physical activity: *r*'s = .34 – .89 (mean *r* = .67) on four different questionnaires (Sallis & Saelens, 2000); in all studies the interval was two weeks.

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

Switching Frequency showed moderate, but Many Brief Switches very low convergent validity.

In previous studies where the validity of self-reports of behaviors other than language switching has been examined using EMA, the reported correlations with related EMA-measures have ranged between .61 and .89 (see section 2.1). In the present study, the correlations between the BSWQ and the EMA questions ranged between .23 and .63, with 14 of 15 correlations being below .3. The correlations between the single questions and EMA ranged between .01 and .43. The convergent validity for the language switching questions were, thus, on the whole lower than for self-report measures of other behaviors. It is possible that this is due to language switching being a relatively frequent and automatic behavior with low affective salience, making it harder to recall in a reliable way (cf. Dolcos & Cabeza, 2002). However, unintended language switches could have somewhat higher emotional salience and thus better recall because in those instances the speaker fails to follow their intention to use a specific language. Moreover, considering that the reliability of many of the retrospective questions was quite low, one cannot expect that their convergent validity with the EMA questions could be very high either.

It cannot be ruled out that the EMA questions and the BSWQ questions tapped on slightly different forms of language switching. For instance, the questions on the LS factor in BSWQ do not mention "intendedness", but instead can be considered to tap on all types of language switches irrespective of type. However, because the BSWQ also includes the Unintended Switches factor, we considered it as redundant to have questions that overlap (overall switches include unintended switches), and thus included the specification of intendedness in the first EMA question. Another reason for the low validity could be that the BSWQ has three questions for each factor, whereas EMA relied only on one question per factor. This choice was motivated by the fact that including multiple questions per factor in EMA would have been too laborious for the participants, given that they had to respond to the questions six times a day.

It could be argued that the participants in our sample were comparatively low-frequency switchers, reporting on average 0 to 3 switches during every two hours (see Table 3 and the recoding that we used in Table 2). This could affect the associations between language switching and the EF tasks (assuming that more switching would result in higher EF), but it should not affect the convergent validity measures for the retrospective questions. However, even if we suppose that the participants were low-switchers, this did not result in floor effects: there was variance in the data even in the low range due to the fact that the EMA switching measures were averages over the two-week period (see Table 3). In addition, we do not know what should be considered as a typical rate of switches for bilinguals over a two-hour period.

Overall, our results bring into doubt the possibility of assessing language switching frequency in retrospect during an unspecified period in the past, and question the reliability of previous studies using such methods.

4.3. Reliability and convergent validity of the executive tasks

We examined the convergent validity and split-half reliability of the executive tasks because previous research indicates that they may be problematic (Paap & Sawi, 2014; Soveri et al., 2016). To investigate the reliability of the executive tasks and possible learning effects, we examined split-half reliabilities (odd versus even trials). The number-letter switch and mixing costs showed adequate reliabilities, but reliabilities of the other executive cost effects were low.[8] This could be due to learning effects, in particular different learning curves in different conditions of the task (similar learning effects in all conditions of a task would not affect Pearson's correlations, which are not sensitive to absolute differences). This motivates the use of linear mixed effects models in the analyses on the associations between EF and language switching, because they take into account possible changes in performance between trials.

Convergent validity was only assessed for the Simon and Flanker tasks, which are both assumed to tap on inhibition, even though they typically show no correlation (Jylkkä, Lehtonen, Lindholm, Kuusakoski & Laine, 2018; Paap & Greenberg, 2013; Paap & Sawi, 2014). In the present study, these two congruency effects were not correlated either ($r = .24$, $p = .22$), indicating that the two tasks either measure distinct types of processes, or involve lots (and different kinds) of residual variance that mask their possible common variance. Possible reasons for the lack of correlation between Simon and Flanker will be discussed in the next section, where the tasks were analyzed at the trial level.

4.4. Trial-level analysis of the executive cost effects, learning effects, and serial congruency effects

In the multilevel analyses, where trial and participant were included as a random effects, the basic executive cost effects were statistically significant and in the expected direction. The sizes of the cost effects were comparable to the cost effects defined as raw subtractions, but the estimates were more accurate in terms of t-values.

When we included trial as a predictor in the models, we observed that in the Simon and Flanker tasks the congruency effect became smaller the higher the trial number, because there was more improvement on the incongruent than on the congruent ones. The effect was, however, only close to significance and small in size. Moreover, in the number-letter task there was a strong learning effect in the single block but not in the mixed block. This could be argued to render the mixing cost as unreliable: the mixing cost apparently becomes larger when the task proceeds, but this is mainly due to improvement in the single block, not worsening of performance on repetition trials. These findings prompt caution when drawing conclusions of executive performance simply based on overall means in each condition, as there could be a trial-level interaction behind the means.

In the Simon and Flanker tasks, we also examined how congruency of the previous trial affects the congruency effect on the next trial. Earlier studies have found that for both tasks, the congruency effect is lower after incongruent trials than after congruent trials (Botvinick et al., 2001; Kerns, 2006; Siemann et al., 2018). This is called the Gratton effect and is assumed to reflect conflict monitoring. In the present study, a strong Gratton effect was found in the Simon task but not in the Flanker task. It thus

[8]Compared to a recent test-retest study by Soveri et al. (2016), the split-half reliabilities were lower except for the number-letter task. In the Soveri et al. study, the Pearson correlation for the Simon effect was .37 (in our study .18), for the number-letter switch cost .68 (our study .72), for the number-letter mixing cost .65 (our study .74), and for the visuospatial n-back effect .72 (our study .44); the Flanker task was not included in Soveri et al.). One should note, however, that the results from these two studies are not directly comparable, because Soveri et al. examined reliability between two test sessions with a three- or six-week interval, whereas we used split-half reliability.

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

appears that the Simon task taps on conflict monitoring to a higher extent than the Flanker task, and it is possible that the Flanker task is indeed more of a selective attention task. It is possible that the participants ignore the information from the flankers when they focus on the central arrow in the Flanker task, and this could result in minimal conflict. In the Simon task, on the other hand, one cannot selectively attend to just the color of the stimulus, ignoring the location, as the location changes constantly. This could explain the lack of correlation between the two tasks.

4.5. Associations between executive performance and language switching

Associations between EF and language switching (assessed with EMA) were examined in multilevel models. In the Simon and Flanker tasks, more contextual language switching was significantly or near-significantly associated with better performance (smaller congruency effect). In the n-back task, higher rate of unintended language switching significantly predicted larger n-back effect (worse WM), and more frequent intended and contextual switching near-significantly predicted smaller n-back effect (better WM). These findings can be interpreted in line with the Bilingual Training hypothesis: participants who make more intended or contextual switches get more training in general inhibitory control processes, selective attention, and working memory, and subsequently perform better.

Because the Simon task showed a significant Gratton effect, we also analyzed associations between this effect and language switching. There was a near-significant interaction between frequency of intended switches and the Gratton effect. It appeared that after congruent trials, IS predicted a larger congruency effect than after incongruent trials. However, the main three-way interaction was only near-significant, and the additional two-way interactions that were performed to disentangle where the effect stemmed from were not even near-significant. Thus, it would be hazardous to make any interpretations based on these results, and these associations should be examined in a larger sample.

In the number-letter task, the findings were less clear. The switch cost was larger the more the participant reported intended language switches, whereas the mixing cost was smaller the more the participant reported unintended or contextual language switches. Initially, this appears to run against the hypothesis that more language switching would be associated with better executive functions. We hypothesized that this pattern of results could stem from the fact that switches in the number-letter task were bottom-up driven (by the location cue), whereas the reported language switches could be both bottom-up (contextual or unintended switches) or top-down driven (intended switches). Thus, we hypothesized that higher frequency of CS and US might be associated with better performance in the number-letter mixed block compared to the single block, whereas IS would be more negatively associated with mixed block performance compared to the single block. These hypotheses were supported in the post hoc analyses.

Overall, the most consistent interactions between the EMA language switching variables and the executive cost effects were found with respect to the Contextual Switches variable, which predicted better performance (smaller cost effects) in all of the tasks. These findings could be interpreted in line with the ADAPTIVE CONTROL (AC) HYPOTHESIS (Green & Abutalebi, 2013), which differentiates between three types of communicative contexts that engage general EF to different extent. In a SINGLE LANGUAGE CONTEXT speakers utilize mainly one language, which puts minimal load on EF (the speaker mainly controls that the non-target language would not activate). In a dense code switching context, on the other hand, speakers can utilize any language they are proficient in, which is likewise minimally demanding for general EF: the speakers have no need to control which language they use. Finally, in a DUAL LANGUAGE CONTEXT speakers have to utilize a specific language with specific speakers, which hypothetically loads most heavily on EF. If we suppose that the contextual switches variable in the EMA application reflected dual language context switching, this would explain why it was most strongly associated with EF. This is a question that should be examined more closely in future studies.

Importantly, the results do not imply causality and are compatible not only with the Bilingual Training hypothesis, but also with an individual differences account. It could thus be that participants with better general EF abilities simply switch between languages more because they can efficiently do so. The possible causal effect of bilingual behavior should be ideally investigated in longitudinal setups (cf. Laine & Lehtonen, 2018). In any case, these results from the present study are in line with the hypothesis that bilingual language switching and general executive tasks rely on the same cognitive functions, which is a prerequisite of the Bilingual Training hypothesis (for a more detailed discussion, see Jylkkä, 2017).

4.6. Limitations

The main limitation of the present study was its modest sample size, which was mainly due to the setup that was relatively demanding for the participants. On the other hand, for the purposes of examining the convergent validity of the retrospective questions, our sample of 30 was arguably sufficient: a sample of this size yields statistical power of roughly .80 to discover a correlation of .50 for convergent validity, which can be expected based on similar earlier studies. Moreover, the limited sample size is also counteracted by the use of multilevel models, which take all variance in the data into account and increase statistical power.

5. Conclusions

Bilinguals have been proposed to outperform monolinguals on a range of executive functions. This enhancement has been suggested to stem from bilingual language use, such as frequent language switching, assumedly training EF. The Bilingual Training hypothesis can thus be taken to imply that higher frequency of language switching in bilinguals is associated with better EF. Earlier studies investigating this have provided inconsistent results. In previous research, language switching has been assessed with general questions or questionnaires that probe switching behavior over an undefined period in the past. The results of the present study, using Ecological Momentary Assessment to measure language switching behavior, suggest that such general language switching questions may lack test-retest reliability and convergent validity. This could explain why the results have been inconsistent. We also found problematic issues with the executive tasks, the two most central ones being as follows: (1) in the number-letter task, learning occurred at different rates across task conditions, which may render the cost effect measures unreliable, and (2) we observed the Gratton effect only in the Simon, not in the Flanker task, suggesting that the two tasks

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

may tap on different types of processes, even though both are often considered as inhibitory tasks. Finally, we found tentative evidence that higher rate of contextual language switches, assessed with Ecological Momentary Assessment, may be associated with better executive control, set shifting, and working memory performance.

Data used in the analysis is available online in Open Science Framework at https://osf.io/387fp/

Acknowledgements. The present study was supported by grants from the Academy of Finland (project #288880), the Emil Aaltonen Foundation project grant, and the University of Helsinki 3-year Funds to Minna Lehtonen. Matti Laine was supported by grants from the Academy of Finland (project #260276) and the Abo Akademi University Endowment (the BrainTrain project). We are grateful to Juhani Virta for his help in testing the participants. We thank all members of the BrainTrain research group at the Abo Akademi University for helpful discussions.

References

Bialystok E (2009) Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition* **12**(1), 3–11. https://doi.org/10.1017/S1366728908003477

Botvinick MM, Braver TS, Barch DM, Carter CS and Cohen JD (2001) Conflict monitoring and cognitive control. *Psychological Review* **108**(3), 624–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11488380

Carlson S, Martinkauppi S, Rämä P, Salli E, Korvenoja A and Aronen HJ (1998) Distribution of cortical activation during visuospatial n-back tasks as revealed by functional magnetic resonance imaging. *Cerebral Cortex* **8** (8), 743–752. https://doi.org/10.1093/cercor/8.8.743

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

de Bruin A, Treccani B and Della Sala S (2015) Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science* **26**(1), 99–107. https://doi.org/10.1177/0956797614557866

Dolcos F and Cabeza R (2002) Event-related potentials of emotional memory: Encoding pleasant, unpleasant, and neutral pictures. *Cognitive, Affective, & Behavioral Neuroscience* **2**(3), 252–263. https://doi.org/10.3758/CABN.2.3.252

Eriksen BA and Eriksen CW (1974) Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* **16**(1), 143–149. https://doi.org/10.3758/BF03203267

Gratton G, Coles MG and Donchin E (1992) Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology. General* **121**(4), 480–506. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1431740

Green DW and Abutalebi J (2013) Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology* **25**(5), 1–16. https://doi.org/10.1080/20445911.2013.796377

Hartanto A and 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. https://doi.org/10.1016/j.cognition.2016.01.016

Homma Y, Ando T, Yoshida M, Kageyama S, Takei M, Kimoto K, Ishizuka O, Gotoh M and Hashimoto T (2002) Voiding and incontinence frequencies: Variability of diary data and required diary length. *Neurourology and Urodynamics* **21**(3), 204–209. https://doi.org/10.1002/nau.10016

Jeffreys H (1961) *Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Johnson HA, Sawi O and Paap KR (2015) Language switching frequency in bilinguals is inconsistently linked to executive functioning. In *Annual Meeting of the Cognitive Neuroscience Society*.

Jylkkä J (2017) *Bilingual Language Switching and Executive Functions*. Åbo Akademi. Retrieved from http://www.doria.fi/bitstream/handle/10024/147587/jylkka_jussi.pdf?sequence=2

Jylkkä J, Lehtonen M, Lindholm F, Kuusakoski A and Laine M (2018) The relationship between general executive functions and bilingual switching and monitoring in language production. *Bilingualism: Language and Cognition* **21**(3). https://doi.org/10.1017/S1366728917000104

Jylkkä J, Soveri A, Wahlström J, Lehtonen M, Rodríguez-Fornells A and Laine M (2017) Relationship between language switching experience and executive functions in bilinguals: An Internet-based study. *Journal of Cognitive Psychology* **29**(4), 404–419. https://doi.org/10.1080/20445911.2017.1282489

Kerns JG (2006) Anterior cingulate and prefrontal cortex activity in an FMRI study of trial-to-trial adjustments on the Simon task. *NeuroImage* **33**(1), 399–405. https://doi.org/10.1016/j.neuroimage.2006.06.012

Laine M and Lehtonen M (2018) Cognitive consequences of bilingualism: where to go from here? *Language, Cognition and Neuroscience*, 1–8. https://doi.org/10.1080/23273798.2018.1462498

Lehtonen M, Soveri A, Laine A, Järvenpää J, de Bruin A and Antfolk J (2018) Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin* **144**(4), 394–425.

Lexell JE and Downham DY (2005) How to assess the reliability of measurements in rehabilitation. *American Journal of Physical Medicine & Rehabilitation* **84** (9), 719–723. https://doi.org/10.1097/01.phm.0000176452. 17771.20

Linck JA, Schwieter JW and Sunderman G (2012) Inhibitory control predicts language switching performance in trilingual speech production. *Bilingualism: Language and Cognition* **15**(03), 651–662. https://doi.org/doi:10.1017/S136672891100054X

Luce KH and Crowther JH (1999) The reliability of the Eating Disorder Examination-Self-Report Questionnaire version (EDE-Q). *The International Journal of Eating Disorders* **25**(3), 349–351. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10192002

Nee DE, Wager TD and Jonides J (2007) Interference resolution: insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, & Behavioral Neuroscience* **7**(1), 1–17. https://doi.org/10.3758/CABN.7.1.1

Paap KR and Greenberg ZI (2013) There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology* **66**(2), 232–258. https://doi.org/10.1016/j.cogpsych.2012.12.002

Paap KR, Johnson HA and Sawi O (2016) Should the search for bilingual advantages in executive functioning continue? *Cortex* **74**(February 2016), 305–314. https://doi.org/10.1016/j.cortex.2015.09.010

Paap KR, Myuz HA, Anders RT, Bockelman MF, Mikulinsky R and Sawi OM (2017) No compelling evidence for a bilingual advantage in switching or that frequent language switching reduces switch cost. *Journal of Cognitive Psychology* **29**(2), 89–112. https://doi.org/10.1080/20445911.2016.1248436

Paap KR and Sawi O (2014) Bilingual advantages in executive functioning: Problems in convergent validity, discriminant validity, and the identification of the theoretical constructs. *Frontiers in Psychology* **5**(962), 1–15. https://doi.org/10.3389/fpsyg.2014.00962

Podsakoff PM, MacKenzie SB, Lee J-Y and Podsakoff NP (2003) Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology* **88**(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Prior A and Gollan TH (2011) Good language-switchers are good task-switchers: Evidence from Spanish–English and Mandarin–English bilinguals. *Journal of the International Neuropsychological Society* **17**(4), 682–691. https://doi.org/10.1017/S1355617711000580

Reinert DF and Allen JP (2002) The Alcohol Use Disorders Identification Test (AUDIT): A review of recent research. *Alcoholism, Clinical and Experimental Research* **26**(2), 272–279. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11964568

Rodriguez-Fornells A, De Diego Balaguer R and Münte TF (2006) Executive control in bilingual language processing. *Language Learning* **56**(s1), 133–190. https://doi.org/10.1111/j.1467-9922.2006.00359.x

Rodriguez-Fornells A, Krämer UM, Lorenzo-Seva U, Festman J and Münte TF (2012) Self-assessment of individual differences in language switching. *Frontiers in Psychology* **3**(388), 1–15. https://doi.org/10.3389/fpsyg.2011.00388

Rogers RD and Monsell S (1995) Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General* **124**(2), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Sallis JF and Saelens BE (2000) Assessment of physical activity by self-report: Status, limitations, and future directions. *Research Quarterly for Exercise*

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press

and Sport **71**(2), 1–14. Retrieved from http://www.tandfonline.com/doi/pdf/10.1080/02701367.2000.11082780

Shiffman S, Stone AA and Hufford MR (2008) Ecological Momentary Assessment. *Annual Review of Clinical Psychology* **4**, 1–32. https://doi.org/10.1007/978-1-4419-1005-9_947

Shrier LA, Shih M.-C. and Beardslee WR (2005) Affect and sexual behavior in adolescents: A review of the literature and comparison of momentary sampling with diary and retrospective self-report methods of measurement. *Pediatrics* **115**(5), 573–581. https://doi.org/10.1542/peds.2004-2073

Siemann J, Herrmann M and Galashan D (2018) The effect of feature-based attention on flanker interference processing: An fMRI-constrained source analysis. *Scientific Reports* **8**(1), 1580. https://doi.org/10.1038/s41598-018-20049-1

Simon JR and Rudell AP (1967) Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *The Journal of Applied Psychology* **51**(3), 300–304. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6045637

Soveri A, Lehtonen M, Karlsson LC, Lukasik K, Antfolk J and Laine M (2016) Test–retest reliability of five frequently used executive tasks in healthy adults. *Applied Neuropsychology: Adult* **0**(0), 1–11. https://doi.org/10.1080/23279095.2016.1263795

Soveri A, Rodriguez-Fornells A and Laine M (2011) Is there a relationship between language switching and executive functions in bilingualism? Introducing a within-group analysis approach. *Frontiers in Psychology* **2** (183), 1–8. https://doi.org/10.3389/fpsyg.2011.00183

Strauss E, Sherman EMS and Spreen O (2006) *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. Oxford: Oxford UP.

Vaz S, Falkmer T, Passmore AE, Parsons R and Andreou P (2013) The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS ONE* **8**(9), e73990. https://doi.org/10.1371/journal.pone.0073990

Verreyt N, Woumans E, Vandelanotte D, Szmalec A and Duyck W (2016) The influence of language-switching experience on the bilingual executive control advantage. *Bilingualism: Language and Cognition* **19**(1), 1–10. https://doi.org/10.1017/S1366728914000352

Wonderlich JA, Lavender JM, Wonderlich SA, Peterson CB, Crow SJ, Engel SG, Le Grange D, Mitchell JE and Crosby RD (2015) Examining convergence of retrospective and ecological momentary assessment measures of negative affect and eating disorder behaviors. *The International Journal of Eating Disorders* **48**(3), 305–311. https://doi.org/10.1002/eat.22352

Yim O and Bialystok E (2012) Degree of conversational code-switching enhances verbal task switching in Cantonese–English bilinguals. *Bilingualism: Language and Cognition* **15**(04), 873–883. https://doi.org/10.1017/S1366728912000478

https://doi.org/10.1017/S1366728918001190 Published online by Cambridge University Press