Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol. 13 No.1 (2025) 1 - 21
ISSN 2295-3159

Language-dependent knowledge acquisition: Mechanisms underlying language-switching costs in arithmetic fact learning

Christian G. K. Hahn[1], Henrik Saalbach[1], Clemens Brunner[2] & Roland H. Grabner[2]

[1] Institute of Educational Sciences, Leipzig University,Germany
[2] Institute of Psychology, University of Graz, Austria

Article received 13 January 2023 / Article revised 30 May 2024 / Accepted 10 December 2024 / Available online 29 January 2025

Abstract

Within the research on bilingual learning, first studies have revealed that content learned in one language is retrieved more slowly when participants have to switch language from instruction to testing (i.e., language-switching costs, LSC). These costs are attributed to language-dependent knowledge representations. However, the cognitive mechanisms underlying LSC are still largely unknown. We investigated these mechanisms by using strategy as well as translation self-reports and by analysing oscillatory parameters in the electroencephalogram (EEG). Thirty-six university students learned arithmetic facts of three different operations over four days either in English or in German. Afterwards, they were tested in both languages with concurrent assessments of self-reports and electrophysiological activity. As expected, LSC in response latencies were observed in all arithmetic tasks. More importantly, analyses of self-reports and EEG revealed that both translation processes and calculation procedures contribute to LSC, with translation processes being the main cognitive mechanism underlying LSC. These results corroborate previous findings of language-dependent knowledge representations in arithmetic fact learning and shed new light on the cognitive mechanisms underlying LSC and possible educational consequences.

Keywords: bilingual learning; language-switching; arithmetic; self-reports; electroencephalograph

Corresponding author: Christian G. K. Hahn, Institute of Educational Sciences, University of Leipzig, Marschnerstraße 31, 04109 Leipzig, Germany, christian.hahn@uni-leipzig.de *DOI: https://doi.org/10.14786/flr.v13i1.1225*

1. Introduction

Speaking a second language is advantageous for various reasons (Baker, 2011). One common approach to foster second language learning is Content and Language Integrated Learning (CLIL). In CLIL, “a language other than the students’ mother tongue is used as a medium of instruction” (Dalton-Puffer, 2007, p. 1). Nowadays, almost all European countries offer programs with non-language classes being taught in a foreign language (EACEA, Eurydice & Eurostat, 2012). Within the German school landscape, for example, CLIL tracks are often introduced in grades six or seven, in which one or two school subjects (e.g., such as geography) are taught in a foreign language (Wolff, 2011). In this vein, educators hope to kill two birds with one stone: learning the subject content as well as a foreign language simultaneously. It is far from surprising that this concept of teaching is gaining more and more popularity, especially in a time when foreign language competencies are essential in the job market.

It is an unresolved question, however, whether CLIL programs may negatively affect the learning of the subject content (Baker, 2011; Pérez-Cañado, 2012). Negative effects of CLIL may arise when the acquired knowledge is stored in the language of instruction (the second language) and, therefore, is not (as) easily accessible in another language (the mother tongue). In fact, there is substantial evidence suggesting that some types of knowledge are stored in a language-dependent way and that language-switching from instruction to retrieval produces performance impairments. These performance impairments are referred to as language-switching costs (LSC) and are typically reflected in longer response latencies or lower accuracy. LSC can be evaluated in experimental training studies in which participants first had to learn new information in one language (training phase), and afterwards, were required to retrieve or apply this knowledge in both the language of instruction and another language (test phase). The comparison of test performance in both languages reveals whether LSC emerge for certain types of knowledge.

LSC have been found in different domains (for *autobiographic knowledge* see Marian & Neisser (2000); for *non-numerical knowledge* see Marian & Fausey, 2006) but have been most intensively studied in the field of arithmetic learning (Spelke & Tsivkin, 2001; Dehaene, Molko, Cohen & Wilson, 2004; Venkatraman, Siong, Chee & Ansari, 2006; Grabner, Saalbach & Eckstein, 2012; Saalbach, Eckstein, Andri, Hobi & Grabner, 2013; Hahn, Saalbach & Grabner, 2017; Volmer, Grabner & Saalbach, 2018). Spelke and Tsivkin (2001), for example, examined LSC in a Russian-English bilingual sample for exact (e.g., “What is the sum of fifty-four and forty-eight?”) and approximate (e.g., “Estimate the approximate cube root of twenty-nine!”) calculations. Participants were trained on arithmetic equations with written number words either in Russian or in English and were then tested with a verification task in both languages. While no LSC were found for approximate arithmetic, suggesting that this type of knowledge is language-independent, response latencies were significantly longer when the language of testing differed from the language of instruction in exact arithmetic. Thus, this study provided strong evidence that numerical fact knowledge, which is relevant for exact calculation, is language-dependent. Further studies corroborated this finding by showing LSC for arithmetic fact knowledge in different operations (multiplication and subtraction: Grabner et al., 2012; exact base-7 addition: Venkatraman et al., 2006; artificial facts: Hahn et al., 2017) and for different language combinations (Russian-English: Spelke & Tsivkin, 2001; Italian-German: Grabner et al., 2012; German-French: Saalbach et al., 2012; German-English: Hahn et al., 2017). In addition, it has been shown that these LSC emerge also in auditory stimuli (instead of written number words; Hahn et al., 2017) and affect the application of fact knowledge in new and more complex task contexts (Volmer et al., 2018).

To date, research on LSC in bilingual learning settings has mainly focused on the appearance of LSC but not on the underlying cognitive mechanisms. Understanding the mechanisms behind LSC is not only of interest to cognitive theories of language-dependent information processing and memory (e.g., Gentner & Goldin-Meadow, 2003; Malt & Wolff, 2010), but also of practical relevance, since it might help to prevent LSC within CLIL. In the domain of arithmetic, there are at least two general possibilities about the underlying mechanisms of LSC. On the one hand, LSC may emerge due to the translation of the knowledge stored in the language of instruction into the language of retrieval or application. For instance, when the arithmetic fact "13 x 8 = 104" is stored in English but needs to be applied in German, the fact could be first retrieved in English and then translated to German. On the other hand, they may result from additional calculation processes in the test language. In the example above, the performance impairment could result from the need to calculate (parts of) the arithmetic problem in German.

Since both general possibilities are compatible with the observed performance impairments during language switching, analyses of response latencies and solution rates are not informative regarding the underlying cognitive mechanisms. One approach to gain further insights into them is to use neurophysiological data as has been done in two functional magnetic resonance imaging (fMRI) studies. Venkatraman et al. (2006) trained 20 English-Chinese bilinguals on base-7 additions (exact number task, e.g., "one-four add three-six") and percentage value estimations (approximate number task, e.g., "forty-four percent of seventy") over a period of five days. Half of the participants were trained in Chinese, half in English. During the fMRI test session, participants had to perform the trained tasks in both languages. In contrast to Spelke and Tsivkin (2001), LSC in response latencies were found for both types of tasks. At the neurophysiological level, LSC were associated with stronger activation in task-dependent networks of brain regions. In the exact number task, additional activation occurred in language-related networks, suggesting that either the equation, the solution or both needed to be translated in order to retrieve the answer from memory. In the approximate number task, stronger activation was found in brain regions associated with magnitude processing and calculation, suggesting a greater effort for participants to solve problems in the untrained language. In the second fMRI study on this topic, Grabner et al. (2012) administered only an exact number task requiring the acquisition of arithmetic fact knowledge. Twenty-nine German-Italian bilinguals underwent a four-day training session of complex multiplication and subtraction problems. Behavioural results again revealed LSC for response latencies. In contrast to Venkatraman et al. (2006), the neurophysiological analyses showed increased activation during language switching in the brain regions associated with magnitude processing and calculation. Therefore, it was argued that LSC might be due to additional numerical processing rather than language translation. In sum, both fMRI studies found increased activation in the language-switching condition but were inconsistent regarding the involved brain networks. Therefore, they do not draw a conclusive picture on the mechanisms behind LSC in arithmetic. Furthermore, both studies used visual stimuli in the form of written number words, which do not represent an ecologically valid learning material (Hahn et al., 2017).

An alternative way to examine the underlying mechanisms of LSC are self-reports. Self-reports have a long tradition in research on arithmetic and are typically used to assess the problem-solving strategy that is applied to solve a given arithmetic problem (e.g., LeFevre, Sadesky & Bisanz, 1996; Campbell & Xue, 2001; Imbo & Vandierendonck, 2007; Grabner & De Smedt, 2011; Vanbinst, Ghesquiere & De Smedt, 2012, cf. Kirk & Ashcraft, 2001; Smith-Chant & LeFevre, 2003). Strategy can be defined as a "procedure or set of procedures for achieving a higher-level goal or task" (Lemaire & Reder, 1999, p. 365). In general, arithmetic problems can be solved either by procedural strategies such as counting (e.g. 8 + 2 = 8 + 1 + 1 = 10) or transformation (e.g. 6 x 12 = 6 x 10 + 6 x 2 = 72), or by direct retrieval of the stored solution from memory (e.g., 6 x 7 = 42). Retrieval strategies are common in single-digit multiplications, which were often memorized by rote in school and for which an arithmetic fact network was built up over several years (e.g., Imbo & Vandierendonck, 2007; Grabner & De Smedt, 2011). But also, after repeated practice of problems typically solved through procedures (such as two-digit subtraction problems), solutions to these problems can be stored in declarative memory (e.g., Grabner & De Smedt, 2012; Hahn et al., 2019). Procedural strategies, in contrast, are used whenever the solution cannot be retrieved because of problem size (e.g., in two-digit multiplications) or operation (e.g., subtraction facts are typically not stored in a fact network; Ischebeck, Zamarian, Siedentopf, Koppelstätter, Benke, Felber & Delazer (2006). Thus, by means of trial-by-trial strategy self-reports it can be examined whether more procedural (calculation) processes take place when language-switching is required compared to when not. In addition to the problem-solving strategy, participants could report whether translation processes were involved in problem-solving. Interestingly, Venkatraman et al. (2006) reported that about two-thirds of the participants mentioned to have thought occasionally in the language of training while performing tasks in the language-switching condition. Unfortunately, there was no systematic acquisition of these comments. Translation self-reports could directly address the question of whether LSC are due to translation processes. However, to the best of our knowledge, such translation self-reports have not been used in previous research on LSC.

Finally, insights into the mechanisms underlying LSC can also be obtained by manipulating the fact learning task. A limitation of most previous studies on LSC in arithmetic lies in the requirement to learn new facts through practicing more complex arithmetic problems, such as two-digit times one-digit multiplications (e.g., Grabner et al., 2012; Volmer et al., 2018; Hahn et al., 2017). Even though these facts can be expected to be retrieved from memory after multiple days of training, LSC can still derive either from translation or from (additional) calculation processes. Therefore, Hahn et al. (2017) introduced a "pure" fact learning condition consisting of artificial arithmetic facts. Specifically, they required participants to learn artificial facts (e.g. 17 box 2 = 93) in addition to complex multiplication (e.g. 16 x 4 = 64) and subtraction (e.g. 52 – 9 = 43) facts. Since the solutions to the artificial problems cannot be calculated but have to be memorized by rote, LSC can be traced back only to translation processes. Thus, comparing the size of LSC across pure and typical fact learning, the impact of translation processes can be estimated. In Hahn et al. (2017), LSC were found in all three operations, and there was no difference in their extent between artificial problems and multiplication problems, suggesting a similar mechanism in these two operations.

1.1 The present study

The main aim of the present study is to provide further insights into the mechanisms underlying LSC in arithmetic fact learning. To this end, we administered an experimental training design with artificial, multiplication, and subtraction problems using auditory stimuli, similar to Hahn et. al. (2017). Extending previous studies, participants had to provide two kinds of trial-by-trial self-reports, one on the problem-solving strategy and one on the use of translation processes.

Moreover, we included electroencephalography (EEG) to complement the information from the strategy self-reports. Here we focus on oscillatory EEG activity in the theta band (event-related synchronization, ERS), which has turned out to be sensitive in distinguishing arithmetic procedures and fact retrieval (e.g., Grabner & De Smedt, 2011, 2012; Tschentscher & Hauk, 2016). In particular, the application of procedures has turned out to be accompanied by lower theta ERS than the application of fact retrieval. Thus, if calculation procedures contribute to LSC, this should be reflected in a difference in theta EEG activity between the switching and the no-switching condition.

We predicted to find longer response latencies for fact learning when the language of training differs from the language of testing, independently of the arithmetic task, while no LSC were expected for accuracy rates (Hypothesis 1). With respect to the cognitive mechanism underlying LSC, two hypotheses were tested: In case that LSC are caused by additional translation processes, a higher frequency of self-reported translated trials for problems in the switching condition compared the no-switching condition should emerge across all three arithmetic tasks (Hypothesis 2a). Alternatively, if LSC are caused by additional calculation procedures, in multiplication and subtraction problems a higher frequency of self-reported procedural strategy use in the switching compared to the no-switching condition can be expected (Hypothesis 2b). Accordingly, if calculation procedures underlie LSC in multiplication and subtraction, lower theta ERS can be expected in the switching compared to the no-switching condition (Hypothesis 3). Finally, we explored correlations between individual differences in LSC and the control variables we assessed, i.e., L2 vocabulary knowledge, intelligence, and arithmetic competencies.

2. Methods

2.1 Participants

The study included 47 right-handed adult students. Eleven participants had to be excluded from analysis: four participants due to missing one training session, three due to technical incidents during the test session, and four due to strong EEG artifacts throughout the test session. The final sample consisted of 36 participants, aged between 20 and 28 years ($M = 23.0$, $SD = 2.1$). Participants were randomly assigned to either a German (L1) or English (L2) training group. All participants studied English Linguistics, had German as mother-tongue, and received their previous math education in German. They gave written informed consent and were paid for their participation. The study was approved by the local ethics committee.

2.2 Experimental Stimuli

The study included 18 problems: 6 artificial, 6 multiplication, and 6 subtraction problems. Artificial problems were two-digit, and one-digit numbers connected via an arbitrary symbol ("box") and a two-digit solution (00 box 0 = 00). These solutions were different from results of any existing arithmetic operation and needed to be memorized by rote. Multiplication problems were two-digit times one-digit problems with two-digit solutions (00 x 0 = 00). Subtraction problems were two-digit minus two-digit problems with two-digit solutions (00 – 00 = 00). In all sessions, the problems were presented auditorily to the participants via a loudspeaker, designed with the text-to-speech software of Voice Reader Studio 15 (Linguatec, 2015). All stimuli (i.e., the whole equation) had the same length (i.e., 1850 milliseconds).

2.3 Additional Measures

2.3.1 English vocabulary knowledge

Participants' English (L2) vocabulary knowledge was assessed by administering the online version of the LexTALE. The LexTALE has been developed to account for the increasing need in experimental studies to assess vocabulary knowledge of English as a second language within a short time scale (Lemhöfer & Broersma, 2012). In this test, participants have to indicate whether presented words are existing English words or not. Such Yes/No tests have been found to be valid measures of L2 vocabulary knowledge (Mochida & Harrington, 2006). Lemhöfer and Broersma (2012) were further able to show that the LexTALE was a better predictor than commonly used self-ratings for vocabulary knowledge, and LexTALE scores have a substantial correlation with common measures of general English proficiency (i.e., Quick Placement Test (QPT; Syndicate, U.C.L.E. (2001)) and Test of English for International Communication (TOEIC; Schmitt, 2005)). The LexTALE consists of 60 items (40 words, 20 non-words). Non-words are orthographically correct and pronounceable but represent strings without meaning. Furthermore, we added a second short test for vocabulary knowledge, the Dialang (Huhta, Luoma, Oscarson, Sajavaara, Takala & Teasdale, 2002). Similarly, the Dialang placement test includes 75 words that need to be marked as existing or non-existing in the English language. In contrast to the LexTALE, answers can be corrected once marked, because all words appear on the same screen. Scores for both tests were averaged to create the final score for L2 vocabulary knowledge. The two tests were strongly correlated ($r = .80$; $p < .001$).

2.3.1 Arithmetic fluency

Since the present study was conducted in the field of arithmetic, all participants were tested on their arithmetic fluency using the French Kit (French, Ekstrom & Price, 1963). In this paper-and-pencil test, participants have to solve as many arithmetic problems as possible within a given time period. For each page, the time limit was two minutes. All subtests consist of two pages. The first subtest contains 60 three-term addition problems with multi-digit addends (e.g., 50 + 42 + 15 = ...), the second subtest 60 multi-digit division problems per page (e.g., 56 : 8 = ...), the third subtest six alternating rows of 10 multi-digit subtraction and multiplication problems per page (e.g. 42 – 17 = ..., and 62 x 6 = ...), and the fourth subtest 60 multi-digit addition and subtraction problems with a suggested answer (e.g., 22 + 29 = 41) that had to be verified. The final score for arithmetic fluency is calculated as the total number of correctly solved problems.

2.3.2 General Intelligence

Participants' intelligence profiles were assessed by using the short version of the Berlin Intelligence Structure Test (BIS-4; Jäger, Süß & Beauducel, 1997). This test includes 15 tasks drawing on three content components of intelligence (numerical, figural, and verbal) and four operational abilities (processing speed, memory, reasoning, and creativity). The overall duration of the test is 45 minutes. The raw scores of the individual tests are aggregated to an IQ score for general intelligence.

2.4 Procedure

The study consisted of five sessions on consecutive days: four training sessions and one test session. All sessions took place at the Department of Psychology of the University of Göttingen, Germany. Training session 1 and the test session were administered in an EEG lab, while training sessions 2, 3, and 4 took place in a computer lab. During the four-day training, participants had to learn the 18 arithmetic problems either in German (L1) or in English (L2). In *Training Session 1* as well as the test session, participants' brain activity was recorded by means of EEG, and the applied strategies were assessed with self-reports as described below.

2.4.1 Training session 1

Training session 1 started with the instruction of the training program as well as an introduction to EEG recording. For later artifact removal (see below), we recorded the EEG during three minutes of eye movements, in which participants were instructed (via visual cues on the display) to roll their eyes, blink, move them up or down, or just keep their eyes open or closed.

Then the experimental task was presented in three blocks. Within each block, there was only one type of task (i.e., MUL, SUB, ART), with each of the 6 problems presented six times (not in succession). The order of the blocks was counterbalanced over the sample and all four training sessions. As depicted in Figure 1, each trial started with a fixation point for two seconds. Then, the problem was presented auditorily via loudspeakers either in English or in German, depending on the training group. Participants had to orally give the answer to the problem as fast as possible in the instructed language. The response time was collected by a voice key. Timeout was set to 8.15 seconds after stimulus presentation (i.e. 10 seconds minus 1.85 seconds stimulus length). The examiner – seated outside the EEG cabin – typed in the given answer, after which the participants received visual feedback on the screen (i.e., a red screen for an incorrect answer and a green screen for a correct answer), followed by the correct answer presented again via the loudspeaker. The next slide asked for the strategy the participant had used to answer the problem (*strategy report*). Using a button response box, participants indicated whether they used (a) fact retrieval (e.g., knowing the answer from memory without any type of calculation), (b) a procedural strategy (e.g., calculating the answer), or (c) any other strategy (e.g., guessing the answer). These strategy reports have been used and validated to assess strategy use in arithmetic in several studies before (Campbell & Xue, 2011; Grabner & De Smedt, 2011; Lefevre et al., 1996). The timeout for the report was set to five seconds. The next trial started after an inter-trial interval of two seconds. Notably, since the participants could not know the solutions to the artificial problems in the first training session, all artificial problems were presented together with the solution twice. Thereafter, participants had to solve these problems on their own, similar to multiplication and subtraction problems. The first session took between 30 and 40 minutes, depending on the individual speed of each participant.

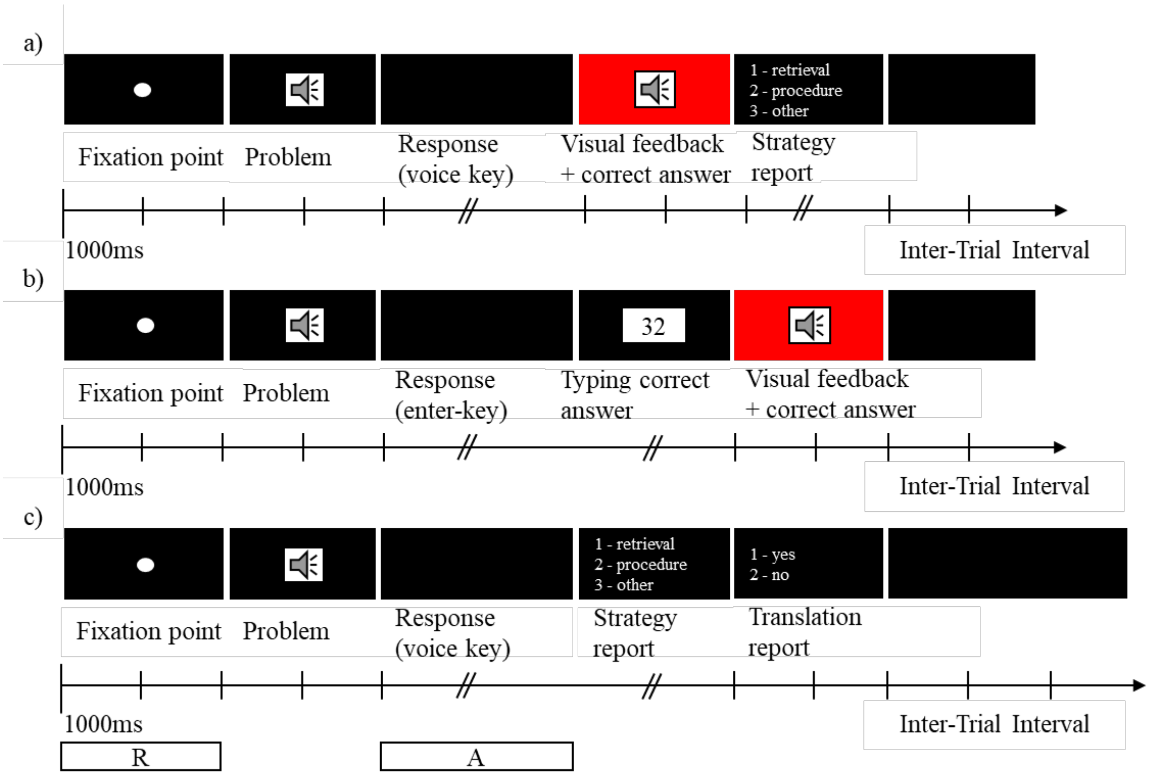

Figure 1. Schematic time course of a) training session 1, b) training session 2, 3, and 4, and c) test session. R = reference interval; A = activation interval.

2.4.2 Training sessions 2, 3, and 4

Over the next three consecutive days, there were three additional training sessions to learn the 18 problems. Each session had a duration between 25 and 35 minutes. In these sessions, no EEG was recorded. Before the training, participants were again given instructions on how to proceed during the session. As with *Training session 1*, the three task blocks were counterbalanced across the participants. The fixation point lasted for two seconds, and the problems were presented via headphones. Furthermore, participants were instructed to press the ENTER key as soon as they had the answer in mind. This was used as an alternative measure of response time to the voice-key in sessions 1 and 5. Afterwards, they were asked to enter the solution using a numerical keypad. Then, participants received corrective visual feedback (correct or incorrect) followed by the correct solution presented auditorily. In these training sessions, no strategy reports were collected. After the four training sessions each problem had been repeated 24 times. This number of trials is in line with previous studies to make sure that participants had sufficiently learned the answer to each problem (Hahn et al., 2017; Grabner & De Smedt, 2012).

2.4.2 EEG test session

In the test session on day 5, the problems were presented in both languages, requiring language-switching or not. After completing the eye-movement EEG as described before (*Training session 1*), all differences to *Training session 1* were explained to the participants before starting with the test session. First, participants did not receive feedback to their responses. Further, participants completed six blocks, including both English and German problems. Within each block, the three operations and the two languages were randomly mixed. Similar to *Training session 1*, participants had to indicate immediately after giving the answer which strategy they used to answer the problem (*strategy report*). The timeout was five seconds. Participants were then asked whether they translated any numbers during problem solving (i.e., by pressing either button 1 or 2 on a response box). We refer to this as *translation report*. The timeout was again set to five seconds.

2.5 Data Analysis

2.5.1 Behavioural Data Acquisition and Analysis

Accuracies and response latencies for correctly solved trials were analysed with ANOVAs. Trials with voice-key errors in *Training session 1* and the *Test session* were excluded from analyses. Before the main analyses, we tested whether the two training groups (training in L1 vs. L2) differed in L2 vocabulary knowledge, intelligence, and arithmetic fluency, using *t*-tests for independent samples. For the training data, the ANOVA included the two within-subject factors Arithmetic Task (artificial vs. multiplication vs. subtraction) and Training Day (day1 vs. day2 vs. day3 vs. day4). The testing data ANOVA comprised the two within-subject factors Arithmetic Task and Language Switching (no-switching vs. switching). A potential impact of the training group (English vs. German) was analysed by means of t-tests for independent samples on the observed LSC. In case of violation of the sphericity assumption (Mauchly's test), degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. All post-hoc tests were conducted using Bonferroni adjusted alpha levels. For the analyses of strategy and translation reports, we conducted mixed repeated-measure ANOVAs, including the within-subject factors Arithmetic Task and Language Switching. For these analyses, the distributions of strategy and translation reports were calculated for correctly solved trials, i.e. frequencies for the three strategies (retrieval vs. procedure vs. other) and the two options for the translation report (no vs. yes). Effect sizes are presented as Cohen's *d* or partial eta-squared ($\eta_p{}^2$).

2.5.2 EEG Data Acquisition and Analysis

EEG was recorded from 64 scalp electrodes with a BioSemi ActiveTwo system (BioSemi, Amsterdam, The Netherlands). Three additional electrodes recorded ocular activity (electrooculogram, EOG); two placed horizontally at the outer canthi of both eyes, and the third above the nasion between the inner canthi of both eyes. Both EEG and EOG signals were sampled at 256 Hz. EEG data analysis focused on oscillatory brain activity and was conducted using MNE Python (Gramfort, Luessi, Larson, Engemann, Strohmeier, Brodbeck, Goj, Jas, Brooks, Parkkonen & Hämäläinen, 2013) as well as custom Python scripts.

After manually marking artifact segments (segments with excessive muscle activity) and removing bad channels (channels with excessive amount of noise or channels with flat power spectral density), we re-referenced the data to the average of all remaining channels. Next, we removed ocular activity using a regression-based approach with coefficients calculated from the eye movement EEG session (Gratton, Coles & Donchin, 1983). Using the clean data, we computed band power in the theta band (4–7 Hz) for each epoch (that is, we filtered the continuous data with a FIR filter with suitable filter characteristics and squared the resulting values). Similar to previous studies (e.g., De Smedt , Grabner & Studer, 2009; Grabner & De Smedt, 2011), we quantified task-related changes in theta EEG activity by computing event-related synchronization (ERS), i.e., the percentage increase in theta power during task processing (an activation period) compared to a pre-stimulus reference period. Specifically, within each epoch, we computed the median theta band power within the reference interval (R) between –2.75 seconds to –0.25 seconds before stimulus onset and the median within the activation interval (A) between 1.85 seconds (end of stimulus presentation) until voice onset. Then, based on the median across epochs for both R and A, we computed theta ERS using the formula: ERS (%) = $(A / R - 1) \cdot 100\%$ (Pfurtscheller & Lopes da Silva, 1999).

For statistical analyses, we averaged all channels per hemisphere and computed the logarithm to make the data more normal. We performed a repeated-measures ANOVA with factors Arithmetic Task, Language Switching, and Hemisphere (left vs. right).

3. Results

Table 1 summarizes the individual characteristics of the participants, separately for the two training groups. There were no significant differences between the German and the English training group in vocabulary knowledge of L2, general intelligence, or arithmetic fluency.

Table 1

Mean scores (standard errors) for the German and English training group (N=18 for each group)

Measure	German Training (L1)	English Training (L2)	*p*
Vocabulary Knowledge L2 (%)	80.4 (3.1)	85.9 (2.2)	.16
General Intelligence (IQ)	94.8 (2.0)	97.1 (1.6)	.37
Arithmetic Fluency (raw score)	128.0 (9.5)	128.8 (5.5)	.94

3.1 Training Data

Training data for response latencies and accuracies are displayed in Figure 2. In both measures, performance improved significantly over the training. For RT, there was a strong main effect of Training Day (F(2.14, 74.84) = 97.68, $p < .001$, $\eta_p{}^2 = .74$), with significant decreases for each consecutive day (all *p*s < .001). In addition, there was a main effect of Arithmetic Task (F(2, 70) = 31.35, $p < .001$, $\eta_p{}^2 = .47$). Artificial problems were solved faster than multiplication problems (1713 ms vs. 2065 ms; t(35) = -4.20, $p < .001$, $d = 0.48$), but more slowly than subtraction problems (1713 ms vs. 1357 ms; t(35) = 3.87, $p < .001$, $d = 0.61$), and multiplication more slowly than subtraction problems (2065 ms vs. 1357 ms; t(35) = 7.68, $p < .001$, $d = 1.08$). There was an interaction between Training Day and Arithmetic Task (F(2.92, 102.16) = 22.19, $p < .001$, $\eta_p{}^2 = .39$), revealing strong training effects for multiplication problems in the first two trainings sessions, in contrast to substantially smaller training effects for artificial and subtraction problems.

For accuracies, there was a main effect of Training Day (F(1.72, 60.12) = 119.49, $p < .001$, $\eta_p{}^2 = .77$), with significant increases for each consecutive day (all *p*s < .001). Further, there was a main effect of Arithmetic Task (F(1.34, 46.92) = 37.28, $p < .001$, $\eta_p{}^2 = .52$). Artificial problems were solved less accurately than multiplications (80.1% vs. 90.1%; t(35) = -5.19, $p < .001$, $d = 1.03$) as well as subtractions (80.1% vs. 95.0%; t(35) = -7.12, $p < .001$, $d = 1.56$), while multiplications were solved less accurately than subtractions (90.1% vs. 95.0; t(35) = -4.50, $p < .001$, $d = 0.96$). Further, there was a significant interaction between Training Day and Arithmetic Task. (F(2.60, 91.03) = 32.16, $p < .001$, $\eta_p{}^2 = .48$), attributable to the strong training effects for artificial problems, in contrast to substantially smaller training effects for subtractions and multiplications, having already high accuracy on day 1.

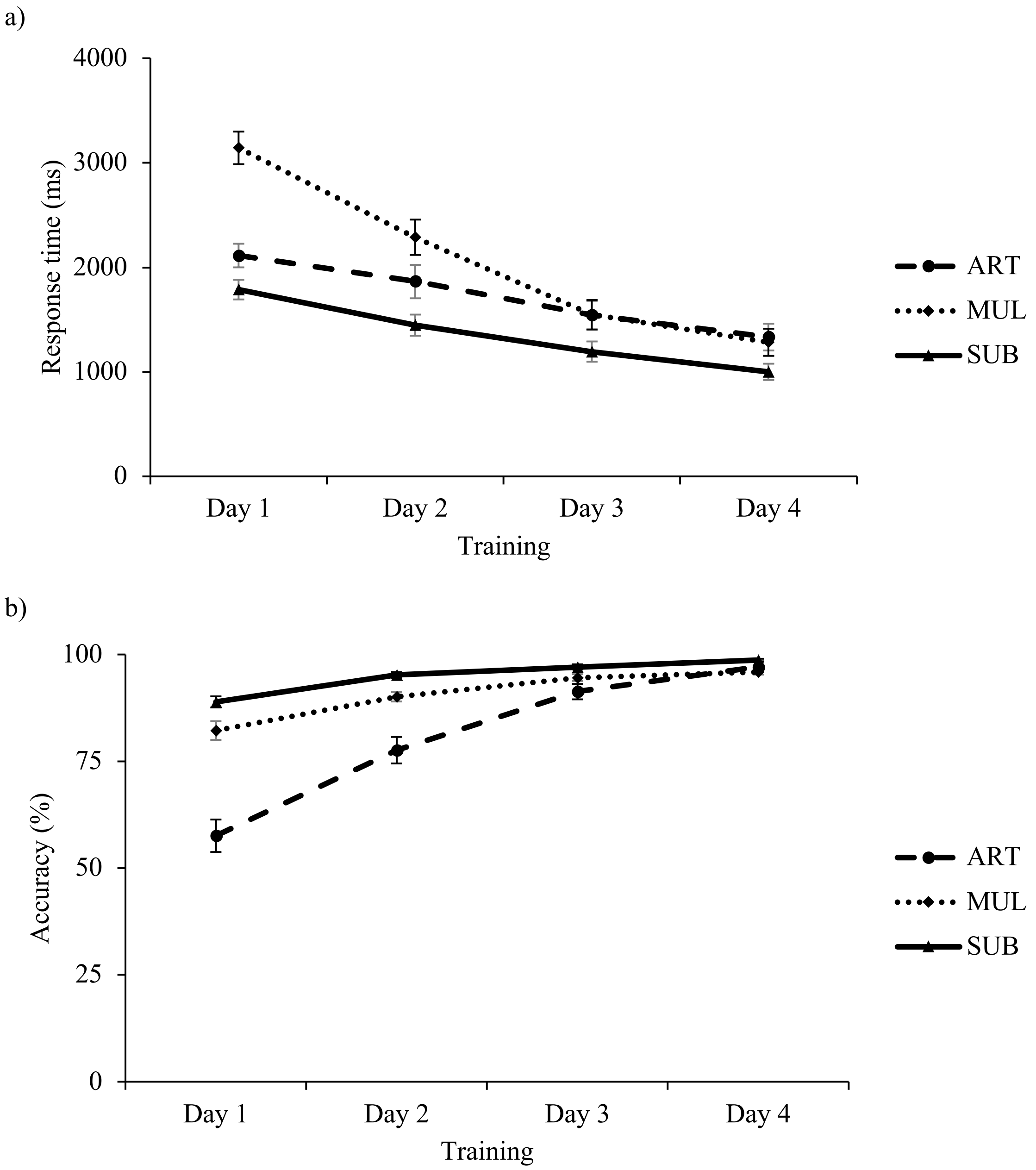

Figure 2. Training data for reaction time (a) and accuracy (b). Error bars indicate the standard error (SE). Separate lines represent the three different tasks. ART = artificial problems, MUL = multiplication problems, SUB = subtraction problems

3.2 Test Session: Performance Data

3.2.1 Language Switching Costs

Descriptive statistics of accuracies and response latencies in the three arithmetic tasks and two switching conditions are shown in Table 2.

Table 2

Mean response latency (for correctly solved trials) in milliseconds (top rows) and accuracy rates in percentage correct (bottom rows) as a function of arithmetic task and switching condition. Standard errors are given in parentheses. LSC were only observed for response latencies.

	Artificial	Multiplication	Subtraction
Response latency in milliseconds			
No language switching	1492 (82)	1493 (100)	1203 (80)
Language switching	1613 (84)	1638 (97)	1352 (97)
Difference	121	145	149
Accuracy in percentage correct			
No language switching	94.0 (1.8)	93.2 (0.9)	96.3 (0.8)
Language switching	91.9 (2.1)	93.3 (0.9)	95.5 (0.9)
Difference	-2,1	-0,1	0,8

Hypothesis 1: We predicted to find longer response latencies for fact learning when the language of training differs from the language of testing, independently of the arithmetic task, while no LSC were expected for accuracy rates.

In line with hypothesis 1, there was a strong main effect of Language Switching across trials at test on response latencies ($F(1, 35) = 22.93$, $p < .001$, $\eta_p^2 = .40$), showing that problems in the no-switching condition were solved faster (1396 ms) than problems in the switching condition (1534 ms). In addition, there was a significant main effect of Arithmetic Task ($F(1.59, 55.62) = 8.19$, $p = .002$, $\eta_p^2 = .19$). Post-hoc analyses revealed that subtraction problems (1278 ms) were solved faster than artificial (1552 ms) and multiplication problems (1565 ms; *p*s < .006). All other effects were not significant (all *p*s > .85). An additional t-test revealed that the two training groups (English vs. German) differed in the LSC in response latencies ($t(34) = .6.07$, $p = .019$; means of 127 vs. 151 ms, respectively). For accuracy, there was no main effect of Language Switching ($F(1, 35) = 2.64$, $p = .11$, $\eta_p^2 = .07$), and none of the other effects was significant (all *ps* > .17).

3.3 Test Session: Strategy and translation reports

Figure 3 displays the distribution of self-reported (a) translation use and (b) procedural strategy use across operations. Since the frequency of trials within the strategy category “other” was very low (< 2.5%), these trials were excluded from further analyses.

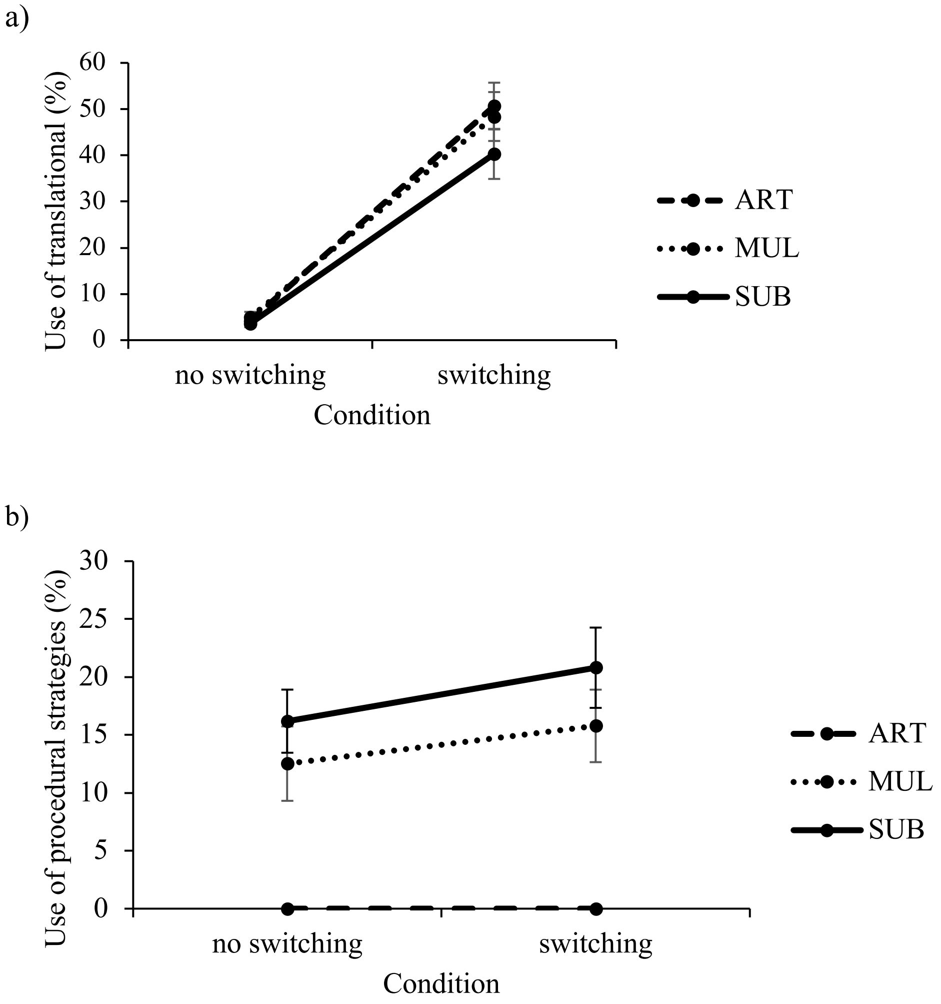

Figure 3. Distribution of self-reports during the test session for a) translation processes and b) procedural strategies. Error bars indicate the standard error (SE).

Hypothesis 2a: In case that LSC are caused by additional translation processes, a higher frequency of self-reported translated trials for problems in the switching condition compared the no-switching condition should emerge across all three arithmetic tasks.

In line with the hypothesis 2a, the repeated measures ANOVA on translation reports showed a main effect of Language Switching ($F(1, 35) = 68.52, p < .001, \eta_p^2 = .66$), indicating that the frequency of translation use was higher in the switching condition (46.45%) compared to the no-switching condition (4.24%). Further, there was a main effect of Arithmetic Task ($F(2, 70) = 7.04$, $p = .002$, $\eta_p^2 = .17$). Post-hoc pairwise comparisons revealed that the frequency of translation use was higher for artificial (27.43%) compared to subtraction problems (21.96%), as well as higher for multiplication (26.66%) compared to subtraction problems (*p*s < .02). Finally, there was an interaction of Arithmetic Task and Language Switching ($F(2, 70) = 5.19$, $p = .008$, $\eta_p^2 = .13$). Post-hoc *t*-tests showed that the frequency of using translation during switching was lower for subtraction (40.30%) compared to artificial (50.66%; $p = .005$, $d = .33$) and multiplication problems (48.39%; $p = .002$, $d = .25$). As validation of the translation reports, we conducted an additional analysis of the response latencies.

Overall, response latencies in trials without reported translation were significantly shorter (1557 ms) than in trials with reported translation (2029 ms; *t(31)* = --6.66, *p* < .001, *d* = .78)[1].

Hypothesis 2b: Alternatively, if LSC are caused by additional calculation procedures, in multiplication and subtraction problems a higher frequency of self-reported procedural strategy use in the switching compared to the no-switching condition can be expected.

The repeated measures ANOVA on strategy reports revealed a main effect of Language Switching (*F*(1, 35) = 14.38, *p* < .001, η_p^2 = .29), indicating that the frequency for procedural strategy use was higher in the switching condition (12.19%) compared to the no switching condition (9.57%). Further, there was a main effect of Arithmetic Task (*F*(2, 70) = 19.52, *p* < .001, η_p^2 = .36). Post-hoc pairwise comparison revealed a higher frequency of procedural strategy use for multiplication (14.15%) and subtraction (18.49%) compared to artificial problems (0%; *p*s < .001). No other effects were significant (*p*s > .10). As validation of the strategy reports, we conducted another additional analysis of response latencies. This revealed that trials in which retrieval strategies were reported were solved significantly faster compared to procedural strategies (1445 ms vs. 2308 ms; *t(22)* = -5.52, *p* < .001, *d* = 1.24)[2].

3.3.1 Relative importance of translation and procedural strategies for LSC

Since switching effects were found to be associated with both self-reported translation and procedural strategy use, we conducted an additional analysis to evaluate their relative importance for LSC. Specifically, we conducted a multiple regression analysis in which we used individual differences in LSC as dependent variable and switching scores of translation reports (percentage of translation: switching – no-switching) and strategy reports (percentage of procedures: switching – no-switching) as independent variables. For response latencies, the regression model explained 22.4% of the variance in LSC (R^2 = .22, *F*(2, 33) = 4.77, *p* = .02). The translation report score was a significant predictor (*ß* = .48, *p* = .005), whereas the strategy report score was unrelated to LSC (*ß* = -.01, *p* = .93). Hence, the more participants used translation processes in the switching (compared to the no-switching) condition, the higher were the LSC. In contrast, despite the fact that participants used significantly more procedural strategies during the switching condition and procedural strategies had significantly longer response latencies, this factor did not predict LSC regarding response latencies. The same analysis was conducted for accuracies. The regression model showed no explanatory value for the prediction of LSC (R^2 = .10, *F*(2, 33) = 1.90, *p* = .17).

3.3.2 Patterns of strategy and translation reports

To provide a more fine-grained picture of the use of strategy and translation processes for LSC, Figure 4 displays a descriptive overview of the different self-report combinations for the three operations.

[1] 4 participants have been excluded from analyses, solving >10 trials per condition

[2] 13 participants have been excluded from analyses, solving >10 trials per condition

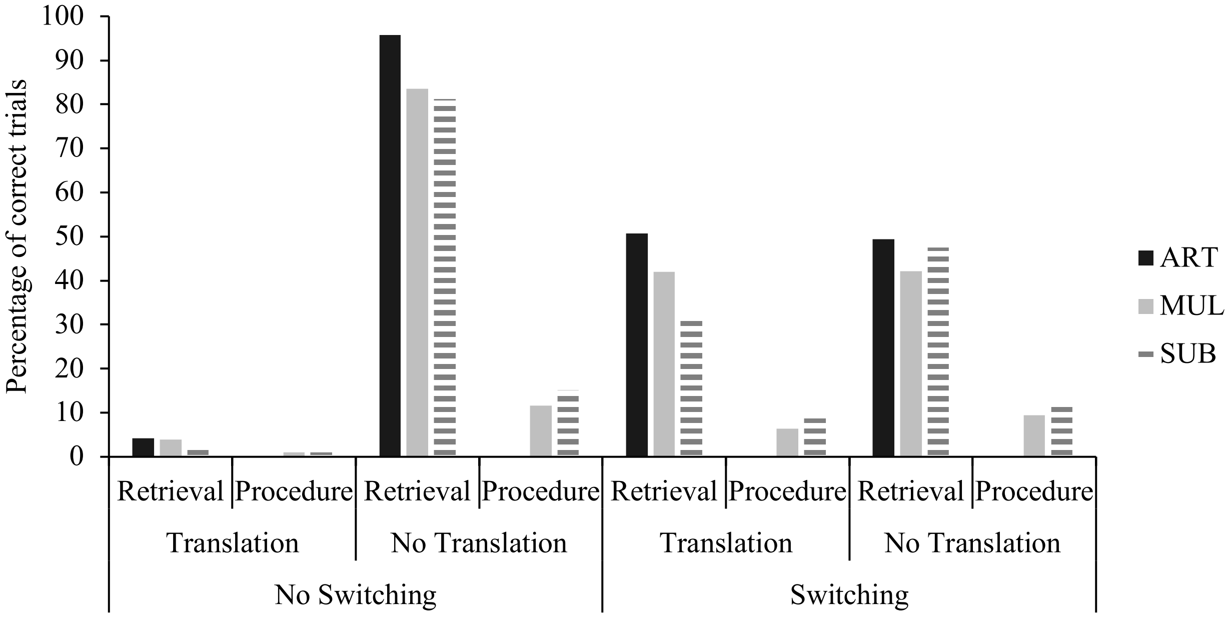

Figure 4. Descriptive overview of the different self-report combinations for the three operations.

In the no-switching condition on the left, there is a high rate of self-reported retrieval, summing up to 100 % for artificial problems and to around 85-90 % for multiplication and subtraction problems. Virtually all problems were reported to be solved without any translation.

In the switching condition, the rate of retrieval remains the same for artificial problems and is slightly decreased for multiplication and subtraction problems (see also Figure 3). In the latter problems, procedures occur both with and without self-reported translations. The first case may reflect that the problem itself is translated into the trained language, calculated there, and then the solution is translated back into the untrained language. The second case may indicate that a procedure is applied without any translation, i.e., in the untrained language. Within the self-reported retrieval and translated trials, it is likely that the solution has been retrieved in the trained language and then translated into the untrained language. Here, a slightly higher percentage was observed for multiplication compared to subtraction problems.

Unfortunately, no inference statistics for performance data can be calculated for the different self-report combinations as there are too few participants (< 10) with at least 10 correctly solved trials for each strategy combination.

3.4 Exploratory analyses of individual differences

Finally, we examined relations between LSC in response latencies and individual differences in the assessed control variables. The results are shown in Table 3. In none of the control variables, a significant association with LSC was observed.

Table 3

Descriptive Statistics and intercorrelations between Language Switching Costs (LSC) and Vocabulary Knowledge in English (L2), General intelligence (IQ) as well as French Kit (math fluency)

Variable	*n*	*M*	*SD*	*r*
1. LSC	36	110	197	-
2. Vocabulary Knowledge	36	83	11	-.20
3. General Intelligence (IQ)	35	97	12	-.16
4. French Kit	36	128	32	-.04

3.5 Test Session: EEG data

Hypothesis 3: If calculation procedures underlie LSC in multiplication and subtraction, lower EEG theta ERS can be expected in the switching compared to the no-switching condition.

Table 4 lists mean theta ERS values for all combinations of conditions and arithmetic tasks. The repeated-measures ANOVA revealed a significant main effect of Language Switching ($F(1, 35) = 6.86$, $p < .01$, $\eta_p{}^2 = .16$). As expected, the switching condition was associated with a significantly lower theta ERS than the no-switching condition (18.9% vs. 16.1%; $d = 0.10$). Furthermore, the interaction between Language Switching and Hemisphere was significant ($F = 4.43$, $p = .043$, $\eta_p{}^2 = .11$. Whereas ERS was about the same for the two hemispheres in the no-switching condition (left: 18.7%, right: 19.2%), the left hemisphere showed lower ERS (15.0%) compared to the right hemisphere (17.2%) in the switching condition. However, no interaction between Arithmetic Task and Language Switching emerged ($F = 1.13$, $p = .32$, $\eta_p{}^2 = .03$). Table 4 reveals that the switching effect (in terms of Cohen's *d*) is descriptively largest in the subtraction, followed by the multiplication and the artificial condition, the latter with a close to zero effect size.

Table 4

Mean ± standard error of theta ERS/ERD (in %) as well as Cohen's d (the effect size of the difference between switching and no switching) for all combinations of conditions and operations

	Artificial	Multiplication	Subtraction
Switching	17.8% ± 1.3%	16.5% ± 1.2%	14.0% ± 1.2%
No switching	18.5% ± 1.2%	19.2% ± 1.2%	19.1% ± 1.3%
d	0.56	2.25	4.25

4. Discussion

The aim of the present study was to provide further insights into the mechanisms underlying LSC in arithmetic fact learning by using trial-by-trial self-reports that were complemented by EEG data. Bilingual adult students were trained on four consecutive days to learn 18 problems of three different operations (artificial problems, multiplications, subtractions) in either German (L1) or English (L2). On the fifth day, all participants were tested on the arithmetic problems in both languages.

We found clear-cut LSC across all three operations for response latencies, thus confirming our first hypothesis. Specifically, participants required more time to solve the problems of all three operations in the language-switching condition than in the no-switching condition. This finding replicates the results of previous studies on arithmetic fact learning in a more "natural" task context as auditory stimuli presentation was combined with a voice key for oral responses. Previous research either collected data using visual stimuli and keyboard responses (e.g., Grabner et al. 2012; Saalbach et al., 2013) or auditory stimuli and keyboard responses (Hahn et al., 2017). No LSC were observed for accuracy rates. This is in line with previous findings revealing that LSC do not emerge in accuracy when participants are given sufficient time to respond (Hahn et al., 2017). In the present study, participants had an even more generous time frame to answer in each trial (i.e., 13 seconds with an average response latency < 2 seconds), which may have led to a ceiling effect in accuracy.

The present study was also novel in that it is the first in which self-reports were used to uncover the cognitive mechanisms underlying LSC. In line with our expectations, participants not only indicated to use more translation processes in the language-switching (compared to the no-switching) condition (hypothesis 2a), but also reported to having applied more procedural strategies (hypothesis 2b). Thus, both hypotheses were confirmed. Even though the latter finding suggests that LSC can be explained by additional numerical processing, in particular calculation, as suggested by Grabner et al. (2012), it needs to be emphasized that only about 12% of the trials in the language-switching condition had been reported to be solved through procedural strategies. In addition, LSC were found for artificial problems, which can only be retrieved from memory to the same extent as for multiplication and subtraction. Therefore, it is unlikely that procedural strategies alone can account for the overall LSC found in our sample. Rather, our findings suggest that translation processes play a major role in LSC (Venkatraman et al., 2006). Approximately 46% of the trials in the language-switching condition were reported as translation trials. These trials also showed significantly longer response latencies than its counterpart (i.e., no translation). Further, the multiple regression analysis including both types of self-reports revealed that only the amount of translation trials is a significant predictor for overall LSC in response latencies.

LSC for artificial problems were assumed to be only due to translation processes. However, the analyses of the translation reports revealed that about 50% of the artificial trials in the language-switching condition were indicated not to include translation processes. There are at least three explanations for this finding. First, when considering that all problems were presented six times in the test session, participants might have had a training in the switching condition during the test session itself. In other words, at some point during the test session (e.g., after solving an arithmetic problem two or three times in the switching condition) participants have acquired the answer to a problem in the previously untrained language and did not require translation any longer. However, in our sample, the use of translation strategies solving items in the untrained language was mixed from the beginning. We analysed the percentage of translation reports across blocks and indeed observed a decrease (block 1+2: 59 %, block 3+4: 53 %, block 5+6: 41 %). Still, if the above-mentioned explanation was true, the translation percentage in blocks 1+2 should be much higher. Second, it is certainly possible that there are participants who trained problem equations in their native language after the end of a training session or before a new training session the next day, regardless of the fact that the actual training language was English. In these cases, a bond between problem equations and training language would be distorted. This possibility seems likely, since a large number of participants articulated their ambition and gave the impression of being upset about having a low solution rate in the first training sessions. Finally, the validity of the translation reports may be limited. In contrast to the problem-solving strategy reports in arithmetic, which are already well-established and validated (Campbell & Xue, 2011; Grabner & De Smedt, 2011; Lefevre et al., 1996), the present study is the first in which trial-by-trial translation reports were required in the domain of arithmetic. Even though translation trials were associated with longer response latencies, it remains elusive to what extent participants accurately reported the occurrence of these processes. In the test session, in which participants had to constantly switch between languages and the three operations, it appears likely that some participants might have had a hard time reliably indicating for each trial what exactly had taken place. In spite of these potential pitfalls, the self-report data provides evidence for translation processes playing a key role in the appearance of LSC.

In addition to self-reports, we collected EEG data to test the link between LSC and calculation processes at the neurophysiological level. Based on the sensitivity of EEG theta activity to arithmetic problem-solving strategies (e.g., Grabner & De Smedt, 2011, 2012; Tschentscher & Hauk, 2016), we hypothesized to find lower theta ERS in the switching (compared to the no-switching) condition because we assume switching to be accompanied by the stronger application of calculation procedures. In line with this assumption, we observed an effect of language switching consisting of lower theta ERS when switching was required. This finding corroborates the results from the strategy reports suggesting that LSC are partly due to additional calculation processes. A closer look at the switching effect sizes, however, revealed that the effects were generally small and only slightly differed between the operations. The largest (but still small) effect of d = 0.18 was observed for subtraction. As expected, the effect size for the artificial numerical facts was practically zero (d = 0.03).

Since this field of research is still in its infancy, it is currently difficult to derive clear implications for practice. The present study can only reflect the actual classroom situation to a very limited extent, since it is a laboratory study focusing on only a fraction of actual school content (i.e., arithmetic fact knowledge). Up to this point, the majority of studies have focused on mathematical knowledge and, likewise within those studies, primarily on simple learning demands (i.e., factual knowledge). To better match real-life learning context, a next step towards procedural and conceptual knowledge would be highly desirable. Moreover, there has been insufficient discussion about the extent to which LSC are of temporal persistence. Thus, we might ask what happens when the instructional language changes during learning phases. The very few studies that have been conducted to evaluate content knowledge acquisition in CLIL instruction suggest that CLIL students perform more poorly (Lo & Lo, 2014; Piesche et al., 2016) or need to spend more time to meet the learning gains of non-CLIL students (Dallinger, Jonkmann, Hollm & Fiege, 2016). In the study of Piesche et al. (2016), for instance, it was shown in six-graders that monolingually educated groups outperformed bilingually educated groups regarding learning gains directly after an intervention (i.e. five 90min-lessons on “Floating and Sinking”) as well as at follow-up six weeks later (small effect-sizes). Such evidence raises the question whether basic concepts (e.g. “Floating and Sinking”) or basic arithmetic shall be learned in the language in which the knowledge will be applied. We might not kill two birds with one stone but even perhaps create little performance gaps we do not see yet, when time efficiency stays the primary concern, with quality of content falling by the wayside. This concern might be especially true considering elementary knowledge which builds the foundation for later study. Learning content and foreign language together may put unnecessary load on the working memory (Sweller, Ayres & Kalyuga, 2011). Yet, a recent study evaluating a long-term immersion program failed to find LSC when language of instruction and language of testing differed (Fleckenstein, Gebauer & Möller, 2019). However, this may also be the result of selection effects having more intelligent students in immersion classes compared to conventional classes. Last, there has not been ample research focusing on individual characteristics. Overall, the research in the area of knowledge acquisition and the understanding on how subject matter content and language of acquisition interact remains important and at its beginning. Thus, there remains a great need for more experimental and ecologically valid research on the topic.

To conclude, in the present study LSC were observed for multiplication, subtraction, and a pure fact learning task using auditory stimuli and an oral response task. By analysing self-reports (i.e., strategy and translation reports), we were able to shed new light on the question of why LSC in arithmetic learning appear. The evidence suggests that translation processes play a key role for LSC in fact knowledge and that this type of knowledge indeed is strongly tied to the language of acquisition. In addition to translation processes, LSC may at least partly be due to the stronger use of calculation procedures, which was observed in both strategy reports and EEG data. Thus, self-reports appear to be a promising way to further elucidate the cognitive mechanisms underlying performance decrements in educational settings in which instruction is provided in a different language than the mother tongue.

Keypoints

- It explores learning and instruction in the context of bilingual learning that is of high current societal relevance
- It introduces the methodology of different self-reports into research on language-switching costs
- Behavioural and neurophysiological methods are applied to investigate cognitive mechanisms of a well-established but poorly understood phenomenon

References

Baker, C. (2011). *Foundations of bilingual education and bilingualism*. Bristol, UK: Multilingual matters.

BioSemi, Amsterdam, The Netherlands. https://www.biosemi.com/

Campbell, J. I., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General, 130*(2), 299-315. https://doi.org/10.1037/0096-3445.130.2.299

Dallinger, S., Jonkmann, K., Hollm, J., & Fiege, C. (2016). The effect of content and language integrated learning on students' English and history competences–Killing two birds with one stone?. *Learning and instruction*, *41*, 23-31. https://doi.org/10.1016/j.learninstruc.2015.09.003

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms* (Vol. 20). John Benjamins Publishing.

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology, 14*(2), 218-224. https://doi.org/10.1016/j.conb.2004.03.008

De Smedt, B., Grabner, R. H., & Studer, B. (2009). Oscillatory EEG correlates of arithmetic strategy use in addition and subtraction. *Experimental Brain Research*, *195*(4), 635-642. https://doi.org/10.1007/s00221-009-1839-9

EACEA, Eurydice, & Eurostat (2012). Key data on teaching languages at school in Europe. Brussels: Eurydice.

Fleckenstein, J., Gebauer, S. K., & Möller, J. (2019). Promoting mathematics achievement in one-way immersion: Performance development over four years of elementary school. *Contemporary Educational Psychology, 56*, 228-235. https://doi.org/10.1016/j.cedpsych.2019.01.010

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Manual for kit of reference tests for cognitive factors (revised 1963)*. Educational Testing Service Princeton NJ.

Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). *Language in mind: Advances in the study of language and thought*. MIT press. https://doi.org/10.7551/mitpress/4117.001.0001

Grabner, R. H., & De Smedt, B. (2011). Neurophysiological evidence for the validity of verbal strategy reports in mental arithmetic. *Biological psychology*, *87*(1), 128-136. https://doi.org/10.1016/j.biopsycho.2011.02.019

Grabner, R. H., & De Smedt, B. (2012). Oscillatory EEG correlates of arithmetic strategies: a training study. *Frontiers in psychology*, *3*, 428. https://doi.org/10.3389/fpsyg.2012.00428

Grabner, R. H., Saalbach, H., & Eckstein, D. (2012). Language-switching costs in bilingual mathematics learning. *Mind, Brain, and Education*, *6*(3), 147-155. https://psycnet.apa.org/doi/10.1111/j.1751-228X.2012.01150.x

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). *MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience, 7*, 267. https://doi.org/10.3389/fnins.2013.00267

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifacts. *Electroencephalography and Clinical Neurophysiology, 55*(4), 468–484. https://doi.org/10.1016/0013-4694(83)90135-9

Hahn, C. G., Saalbach, H., & Grabner, R. H. (2017). Language-dependent knowledge acquisition: investigating bilingual arithmetic learning. *Bilingualism: Language and Cognition*, *22*(1), 1-11. http://dx.doi.org/10.1017/S1366728917000530

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Teasdale, A. (2002). DIALANG: A diagnostic language assessment system for learners. Common European framework of reference for languages: Learning, teaching, assessment. Case studies, 130-145.

Imbo, I., & Vandierendonck, A. (2007). The development of strategy use in elementary school children: Working memory and individual differences. *Journal of experimental child psychology*, *96*(4), 284-309. https://psycnet.apa.org/doi/10.1016/j.jecp.2006.09.001

Ischebeck, A., Zamarian, L., Siedentopf, C., Koppelstätter, F., Benke, T., Felber, S., & Delazer, M. (2006). How specifically do we learn? Imaging the learning of multiplication and subtraction. *Neuroimage*, *30*(4), 1365-1375. https://doi.org/10.1016/j.neuroimage.2005.11.016

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). Berliner Intelligenzstruktur-Test: BIS-Test Form 4. Göttingen: Hogrefe.

Kirk, E.P., and M.H. Ashcraft. 2001. Telling stories: the perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology. Learning, Memory, and Cognition* *27*(1): 157–175. https://psycnet.apa.org/doi/10.1037/0278-7393.27.1.157

LeFevre, J. A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 216. http://dx.doi.org/10.1037/0278-7393.22.1.216

Lemaire, P., & Reder, L. (1999). What affects strategy selection in arithmetic? The example of parity and five effects on product verification. *Memory & Cognition, 27*(2), 364-382. https://doi.org/10.3758/BF03211420

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. https://psycnet.apa.org/doi/10.3758/s13428-011-0146-0

Linguatec (2015). Retrieved from http://www.linguatec.de/en/text-to-speech/voice-reader-studio-15/.

Lo, Y.Y., & Lo, E. S. C. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research, 84*(1), 47–73. https://doi.org/10.3102/0034654313499615

Malt, B., & Wolff, P. (Eds.). (2010). *Words and the mind: How words capture human experience*. New York: Oxford University Press.

Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology, 20*(8), 1025-1047. https://doi.org/10.1002/acp.1242

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General, 129*(3), 361. https://doi.org/10.1037/0096-3445.129.3.361

Mochida, K., & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing, 2*, 73–98. https://doi.org/10.1191/0265532206lt321oa

Pérez-Cañado, M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, *15*(3), 315-341. https://doi.org/10.1080/13670050.2011.630064

Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology, 110*(11), 1842–1857. https://doi.org/10.1016/S1388-2457(99)00141-8

Piesche, N., Jonkmann, K., Fiege, C., & Keßler, J. U. (2016). CLIL for all? A randomised controlled field experiment with sixth-grade students on the effects of content and language integrated science learning. *Learning and Instruction*, *44*, 108-116. https://doi.org/10.1016/j.learninstruc.2016.04.001

Saalbach, H., Eckstein, D., Andri, N., Hobi, R., & Grabner, R. H. (2013). When language of instruction and language of application differ: Cognitive costs of bilingual mathematics learning. *Learning and Instruction*, *26*, 36-44. https://doi.org/10.1016/j.learninstruc.2013.01.002

Smith-Chant, B. L., & LeFevre, J. A. (2003). Doing as they are told and telling it like it is: Self-reports in mental arithmetic. *Memory & Cognition*, *31*(4), 516-528. https://doi.org/10.3758/BF03196094

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45-88. https://doi.org/10.1016/S0010-0277(00)00108-6

Schmitt, D. (2005). Test of English for International Communication (TOEIC). *ESOL Tests and Testing,* 100-102.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive Load Theory* (pp. 71-85). Springer, New York, NY. https://doi.org/10.1007/978-1-4419-8126-4_6

Syndicate, U.C.L.E. (2001). *Oxford Quick Placement Test*. Oxford: Oxford University Press.

Tschentscher, N., & Hauk, O. (2016). Frontal and parietal cortices show different spatiotemporal dynamics across problem-solving stages. *Journal of Cognitive Neuroscience*, *28*(8), 1098-1110. https://doi.org/10.1162/jocn_a_00958

Vanbinst, K., Ghesquiere, P., & De Smedt, B. (2012). Numerical magnitude representations and individual differences in children's arithmetic strategy use. *Mind, Brain, and Education*, *6*(3), 129-136. https://doi.org/10.1111/j.1751-228X.2012.01148.x

Venkatraman, V., Siong, S. C., Chee, M. W., & Ansari, D. (2006). Effect of language switching on arithmetic: A bilingual fMRI study. *Journal of Cognitive Neuroscience*, *18*(1), 64-74. https://doi.org/10.1162/089892906775250067

Volmer, E., Grabner, R. H., & Saalbach, H. (2018). Language switching costs in bilingual mathematics learning: Transfer effects and individual differences. *Zeitschrift für Erziehungswissenschaft*, *21*(1), 71-96. https://doi.org/10.1007/s11618-017-0761-0

Wolff, D. (2011). Der bilinguale Sachfachunterricht (CLIL): Was dafür spricht, ihn als innovatives didaktisches Konzept zu bezeichnen. In *Forum Sprache, 6*, 74-83.