Language-Switching Costs in Bilingual Mathematics Learning

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm Biologie
der Georg-August University School of Science (GAUSS)

vorgelegt von

Christian G. K. Hahn

aus Neuss

Göttingen, 2019

Thesis Committee

Prof. Roland H. Grabner

Institute of Psychology, University of Graz, Austria

Prof. Henrik Saalbach

Educational Psychology, University of Leipzig, Germany

Members of the Examination Board

Reviewer: Prof. Roland H. Grabner

Institute of Psychology, University of Graz, Austria

Second Reviewer: Prof. Sascha Schroeder

Educational Psychology, Georg-August-University of Göttingen, Germany

Further members of the Examination Board

Prof. Henrik Saalbach

Educational Psychology, University of Leipzig, Germany

Prof. Lars Penke

Biological Personality Psychology, Georg-August-University of Göttingen, Germany

Prof. Annekathrin Schacht

Affective Neuroscience and Psychophysiology, Georg-August-University of Göttingen, Germany

Prof. Nivedita Mani

Psychology of Language, Georg-August-University of Göttingen, Germany

Date of oral examination: 19th, of June 2019

Acknowledgements

At the moment of writing, I am overwhelmed with feelings. It is neither easy nor possible to summarize the past six years in a few lines. Several times, I have considered the idea of not finishing the work due to variety of circumstances. A dissertation is hard work, it is exhausting at times and sometimes I missed the answer to the question: why am I doing this? A dissertation is challenging, it shows you all the facets of motivation and reveals your true intellectual strengths and weaknesses. It teaches you what discipline really means. I do not want to miss anything, I experienced during these last six years. My thanks ultimately go to all the people I have encountered since I started studying in 2008. You all influence the way I think and work. All of you shaped me in a certain way and contributed to the fact that this work came to existence. In the following, people are explicitly mentioned who I consider to be particularly influential for this work.

I would like to thank my supervisor Roland Grabner. It is still a mystery to me how patient and with what trust you supervised me during these six years. You hired me, even though I certainly could not have been the first choice on paper. You read between the lines and saw my potential. Despite many years of long-distance supervision, I always had the impression that you were there for my concerns, that I could ask a question at any time and that responding to me was at your priority. I appreciate this quality very much and your way of supervision has also been vital for me to be able to finish this project. My thanks also go to my second supervisor Henrik Saalbach. I thank you for your supportive words and critical negotiations concerning my ideas in research. This has repeatedly led me to reconsider details, putting my ego aside in order not to act too deadlocked. Thank you for seeing my potential as well and hiring me. Also without you this work would never have been finished. One of my main motivations in recent months has been to finish the work for both of you, as my own drive was hard to find.

I would like to thank Maria, Frieder, Stephen, Alexander, Matthias, Tobias, Ruben, Anna, Lars and Christina. You are all part of my time at the University of Göttingen. You have all influenced the project. Be it through your support in the organization, be it through your help with analyses, be it because you were simply there when I needed to talk. I thank Sascha Schroeder, helping me in the last course of the dissertation, taking over the role of supervision. I did not take that for granted.

Thanks go to my department in Leipzig. You all had to listen to me again and again when I was at a loss. You also helped me to complete this work: Anika, Ben, Berit, Catherine, Cathrin, Conny, Franziska, Franziska, Gerlind, Monique, Robert, Susanne.

I thank my parents and their partners Andreas and Erika. I would like to thank all my friends who supported me in my private life. You have given me the necessary balance.

I would like to thank Ingrid Quintana, the woman I am about to marry. I cannot put into words how much strength you have given me over the last two years.

All of you and many more have been part of my journey. I thank you with all my heart.

Preliminary Note

Throughout the dissertation I will use the pronoun ‘we’ instead of ‘I’. The work here is my own in terms of hypotheses, analyses and conclusions, but it is effectively the product of close collaboration and constructive debate with my supervisors Roland H. Grabner and Henrik Saalbach, as well as colleagues of the Georg-Elias-Müller-Institute of Göttingen, and the Department of Education in Leipzig. Three empirical studies are presented. The first study has been published in a peer-reviewed journal and the following text appears unchanged. However, the Figures of the original publications were adjusted to be consistent with the present format. Studies 2 and 3 are manuscripts in preparation. Their introductions and discussions were shortened and adapted for this thesis to avoid redundancy and highlight similarities and contrasts between the studies.

Original publication

Hahn, C. G., Saalbach, H., & Grabner, R. H. (2017). Language-dependent knowledge acquisition: investigating bilingual arithmetic learning. Bilingualism: Language and Cognition, 1-11.

Table of Content

List of Figures

List of Tables

List of Abbreviations

CLIL	Content-and-Language-Integrated-Learning
NFK	Numerical fact knowledge (i.e., stored information on numerical problems that are determined by mathematical law, such as 9 being the results of 3 x 3), often referred to as exact number task or exact calculation task.
LSC	Language-switching costs
ACC	Accuracy (in %)
RT	Reaction time (in milliseconds)
ART	Artificial problems, representing pure fact learning
MUL	Multiplication problems
SUB	Subtraction problems
OLDADD	Trained addition problems in base-7
NEWADD	Untrained addition problems in base-7
NEWSUB	Untrained subtraction problems in base-7
no switching	Test situation in which the language of application is the same than the language of instruction
no-switching	Same as no switching (note: since Study 1 has already been published, it may use different abbreviations.)
switching	Test situation in which the language of application is different from the language of instruction
trained problems	Set of arithmetic problems that is trained during the training sessions
untrained problems	Set of arithmetic problems that is new during the test session
e.g.,	“for example,” (abbreviation for exempli gratia)
i.e.,	“that is,” (abbreviation for id est; used to give specific clarification)
cf.	“compare” or “consult” (used to provide contrasting or opposing information)
vs.	“versus”

General Introduction

Speaking a second language is advantageous for several reasons. In his book *The Foundations of Bilingual Education*, Baker (2011), for instance, highlights ideological (e.g., assimilation), international (e.g., trade) and individual (e.g., cultural awareness) benefits. One common approach trying to foster second language learning in school 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). In 2012, Denmark, Greece, Iceland and Turkey were the only European countries that did not offer any kind of program with non-language classes being taught in a foreign language (EACEA, Eurydice, & Eurostat, 2012). Within the German school context, CLIL is often introduced in grades six or seven, with pupils having the choice to switch into a CLIL track. This CLIL track is most commonly linked to having one or two school subjects (e.g., such as geography) taught in a foreign language (Wolff, 2011). Thus, by using a foreign language to teach content subjects, 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 where language abilities seem to be helpful on the job market. However and most critical, empirical research does not provide a convincing picture if CLIL even improves second language competencies (Johnson & Swain, 1997; Nold, Hartig, Hinz, & Rossa, 2008; Cheng, Kirby, Quiang, & Wade-Wolley, 2010; Köller, Leucht, & Pant 2012; Hüttner & Smit, 2013; Lo & Lo, 2014; cf. Bruton, 2013; Roquet and Pérez-Vidal, 2015; Pladevall-Ballester & Vallbonab, 2016). Furthermore, it is an unresolved question whether CLIL programs may negatively affect the learning of the subject content when the mastery of subject content is dependent on the mastery of the language (Baker, 2011; Perez-Canado, 2012). However, learning the subject content should be the major interest.

Within the present research project, we intent to take a closer look at one of the most fundamental aspects of human learning that all CLIL approaches need to deal with: the possibility that the acquisition of knowledge is connected to the language of instruction and therefore qualifies as a determinant for later retrieval of that knowledge. Hence, the following main research questions enclose the present work: **is the recall of information affected when the language of knowledge application differs from the language of knowledge acquisition? And if yes, why?** So far, a number of research studies issued this question within different domains (Marian & Neisser, 2000; Spelke & Tsivkin, 2001; Venkatraman, Siong, Chee, & Ansari, 2006; Marian & Fausey, 2006; Grabner, Saalbach, & Eckstein, 2012; Saalbach, Eckstein, Andri, Hobi, & Grabner, 2013), finding performance impairments when acquired knowledge had been encoded in one language and retrieved in another language (i.e., language-switching is required). **Throughout this project, such performance changes will be called language-switching costs (LSC). LSC reflect either longer response time (RT), lower accuracy (ACC), or both.** Beforehand to mention, the experimental studies conducted within this project will take a look at the field of mathematics. Numerous studies have shown that skills in mathematics early in childhood serve as a key predictor for subsequent learning in several areas (e.g., Dowker, 2005; Duncan et al. 2006; Claessens, Duncan, & Engel, 2009; Geary, 2013; Watts, Duncan, Siegler, & Davis-Kean, 2014). Therefore, we are especially eager to further our knowledge on the language dependency of learning in the context of mathematics. Study 1 and 2 will focus on declarative knowledge (from now on called numerical fact knowledge (NFK)), whereas Study 3 will put its emphasis on procedural knowledge. Due to a limited scope of this dissertation and the importance of basic arithmetic knowledge, the field of conceptual knowledge can not be considered here.

State of the art

Previous studies on LSC will be outlined in detail in the upcoming section in order to recognize the need for further research as well as understand specific methodological challenges this project tried to overcome. This project was built upon these studies with the purpose of extending the current evidence.

Language-dependent memory and the self

In 2001, Marian and Neisser were investigating whether recall of autobiographic memories is language-dependent (see also preliminary work of Otoya, 1987; Schrauf & Rubia, 1998). The research was based on the *encoding-specificity principle* by Tulving and Thomas (1973) stating that the quality of memory retrieval is higher when the environment of learning matches the one of testing. Within their study, language was viewed as the environmental factor. In one of their experiments, Marian and Neisser therefore tested twenty university students being Russian-US immigrants. Participants were interviewed in either Russian (L1) or English (L2) asked to tell brief stories in response to word prompts (e.g., summer). The authors found that significantly more memory was recalled from the time living in Russia, when the interview was held in Russian. In contrast, more recent memories from the time being in the United States were recalled when the interview was held in English. Thus, the language of testing influenced the kind of memory accessed by the interviewee. In the same vein, Marian and Kaushanaskayat (2004) were able to illustrate that language can influence self-construal. In their study, forty-seven Russian-English bilinguals were interviewed in Russian (L1) as well as in English (L2). As in the study by Marian and Neisser (outlined above), participants were asked to response to a number of prompts in the appropriate language. Among others, the authors found that speaking in English lead to more memories expressed in an individualistic way (measured by the number and kind of personal and group pronounce used), compared to statements made in Russian (which is seen as a more collectivist culture). These two examples indicate that language can shape thinking and hence resulting output.

In 2006, Marian and Fausey investigated the topic in the field of academic-like information. In contrast to previous studies, where participants where tested on already existing memories, twenty-four Spanish-English bilinguals had to learn new information within the fields of history, biology, chemistry, and mythology. For each participant learning took place in both languages. Participants’ language-proficiency was collected via self-reports. Analysis revealed on average a higher reading, understanding, and speaking proficiency in Spanish (L1) compared to English (L2). Since individual differences were present, the sample was grouped into balanced-bilinguals (i.e., comparable proficiency in Spanish and English), and unbalanced-bilinguals (i.e., more dominant in Spanish than in English). Participants listened alternately to blocks of stories in Spanish and English, with distracting blocks in-between (i.e., puzzle task). Afterwards, they were tested in both languages. Language of instruction and testing were counterbalanced (e.g., one group of participants heard English stories first, followed by Spanish stories, and were tested in English, followed by Spanish, and so on). Data on RT (i.e., time between the end of the question and the onset of a participant´s answer) and ACC were collected. LSC for ACC were found only for the group of balanced bilinguals. This means, balanced bilinguals answered more questions correctly if the language of testing matched the language of learning (i.e., no switching) in contrast to the non-match of languages (i.e., switching), no matter in which direction (i.e., L1 to L2; L2 to L1). The authors argued that no LSC in ACC for the unbalanced-bilinguals were found, because these participants may have already encoded the incoming information in their dominant language, even though the learning language was the non-dominant. Therefore, it was less likely that the incoming information was connected to the language of learning. Thus, within the testing situation, it was unlikely to make more mistakes when being tested in English, compared to Spanish. In case of balanced-bilinguals, however, the language of instruction becomes the language that is strongly tied to the incoming information. Regarding RT, there was no difference between the two groups. Both groups where faster when language of instruction was Spanish. LSC were only found, when training took place in Spanish and testing in English (i.e., L1 to L2), not vice versa (i.e., L2 to L1). The authors argued, that this was due to the specific sample tested. Since participants were living in the United States, they were rather used to encode information in English and using it in Spanish, than vice versa. Overall, the study indicates that a match or mismatch of language of instruction and language of application negatively affects performance.

Language-dependent memory: the case of arithmetic

Most intensively, LSC were investigated in the field of arithmetic. The study of Spelke and Tsivkin (2001) marks the groundwork study for all upcoming research concerned with LSC in arithmetic. Within three single studies, the authors examined language-dependent memory in a sample of eight Russian-English balanced-bilingual adults. A language comprehension test for both languages decided whether participants were qualified to take part in the study. Within the first experiment, participants were trained for two days in two different NFK tasks as well as two different approximation tasks. The NFK tasks included large addition problems (e.g., “What is the sum of fifty-four and forty-eight?”), and small addition problems in the base-6 and base-8 number system (e.g., “What is the sum of five and three in base-6?”; see Supplementary Material on page 109, explaining calculation in a different number system). Approximation was presented by approximation of cube roots (e.g., “Estimate the approximate cube root of twenty-nine!”), as well as approximation of logs base-2 (e.g., “What is the base-2 logarithm of 45?”). During the training and test sessions, different sets of problems were presented in written number form in either Russian (L1) or English (L2). On a third day, participants were tested for the exact same sets in both languages. Within each test block, only one language was present (i.e., block-wise language switching). Testing also included new problems in both languages. No LSC were found the approximation tasks, implying that this type of knowledge (i.e., procedural knowledge, namely, how to estimate a cube roots or logarithms) is stored in a language-independent way. For both NFK tasks (i.e., normal addition and addition in different base systems), LSC were found for RT. The effects were independent of the direction of switching (i.e., L1 to L2 or vice versa). Regarding new problems, it was found that for NFK, trained problems were consistently solved quicker than untrained problems, independent if the task involved language-switching or not (i.e., participants were faster solving trained problems in the switching condition than untrained problems in the no switching condition). Interestingly, participants solved untrained NFK problems faster, when the problems shared features with trained problems (e.g., same first addend) than if they did not and only if the language matched. These two latter findings strengthened the evidence for a language-specific learning of NFK. Whereas results regarding LSC for NFK were replicated in a second experiment, findings for untrained problems were not. In a third experiment, LSC were even found when numerical facts were put in a context of historical or geographic content, but not for non-numerical facts. Statements on LSC for ACC were rather difficult to make, since problems were solved to a high extent in both conditions (i.e., no switching vs. switching). Overall, the three experiments provided the first evidence that the internal representation of NFK is at least to some extent language-dependent. This is in line with studies in the field of numerical cognition showing that the retrieval of numerical facts is linked to brain circuits associated with language processing and storage of verbal information (e.g., Lee, 2000; Dehaene, Molko, Cohen, & Wilson 2004; Domahs & Delazer, 2005; Venkatraman, Siong, Chee, & Ansari, 2006; cf. Benn, Zheng, Wilkinson, Siegal, & Varley, 2012; Klessinger, Szczerbinski, & Varley, 2012). Based on the emphasize of a different knowledge content, the study by Marian and Fausey (2006) was already outlined in the previous section, but was the one that followed the study by Spelke and Tsivkin (2001).

Regarding the research design, it is to note that the study of Spelke and Tsivkin (2001) represents the first in the field of LSC using a training design. In so-called training studies, participants train specific tasks over a period of several days, before tested on a final test day in order to examine fact learning (e.g., Grabner, Ischebeck, Reishofer, Koschutnig, Delazer, Ebner, & Neuper, 2009). Within the context of LSC, participants first have to learn new information in one language (e.g., the second language; training session), before they are required to apply this knowledge in both the language of instruction and another language (test session). The comparison of test performance in both languages reveals whether LSC emerge for certain types of knowledge, regarding RT and/or ACC. All studies reviewed in the following paragraphs as well as the three studies of the current project follow this methodology.

Motivated by Spelke and Tsivkin, Venkatraman et al. (2006) conducted the second experimental training study on LSC in NFK. In order to investigate possible mechanisms underlying LSC, neurophysiological measurements were included (i.e., functional magnetic resonance imaging, fMRI). Over a period of five days, 20 English-Chinese bilingual adults were trained in a NFK and an approximation task. Training participants in addition problems in the base-6 number system (e.g., one-four add three-six") represented the NFK task. As approximation tasks, participants were trained to estimate percentages (i.e., "forty-four percent of seventy"). As in previous research, written number words were used as stimuli. Answers were given by choosing among two options presented on a computer screen. During the training, half the sample trained NFK in English (L1) and approximation in Chinese (L2), and the other half the other way around. There was no formal indication of language proficiency. Participants had at least 10 years of formal education in both languages, and were categorized as balanced-bilinguals. On day six, participants were tested on both tasks in both languages. The test design included four successive blocks of exact tasks and four successive blocks of approximate tasks. Each block either contained English or Chinese tasks, so that participants were not switching languages on a trial-by-trial basis. In contrast to findings by Spelke and Tsivkin (2001), LSC were found for both tasks regarding RT. No LSC were found for ACC. Authors interpreted the finding of LSC for the approximation task by guessing that the task was more difficult compared to the task used in the study by Spelke and Tsivkin. They did not disagree with the assumption that approximation is rather stored in a language independent manner, since their neurophysiological results indicated greater activation in visuospatial circuits, compared to higher demands in language-specific areas within the switching condition of the NFK. Therefore, first evidence was provided for possible underlying mechanisms of LSC. LSC may appear on account of additional linguistic processes that might suggest that subjects need to translate information to arrive at the solution.

In 2012, Grabner et al. published additional data on LSC, likewise using fMRI to gain more insight into possible underlying mechanisms of LSC. According to the authors, the study of Venkatraman and colleagues (2006) had statistical weak spots, which made it difficult to proper interpret the data to conclude that LSC are due to additional linguistic processing. 29 Italian-German bilingual adults were trained for four days in NFK (i.e., multiplication and subtraction problems). Half the sample trained in Italian, the other half in German. Language proficiency in German (L2) was stated between upper intermediate (B2) and highly competent (C2). On day five, participants were tested in both languages, for the trained problems as well as for untrained problems of the same arithmetic operations. As in previous studies, stimuli were provided in written number format. During training, participants had to choose among three possible answers, whereas during testing, participants only verified a predetermined number. The test-design was a random switching between tasks and languages within each block. In total, the test session included sixteen blocks of fifteen trials each, with each single problem presented three times. Overall, trained problems were solved faster than untrained problems in both conditions (no switching vs. switching). No LSC were found for ACC, neither for trained nor untrained problems. For RT, LSC were found for trained, as well as untrained problems. There were no differences between the two language training groups (German vs. Italian) regarding RT or ACC with respect to LSC. These behavioral results fit the findings by Spelke and Tsivkin (2001) regarding the language-dependency of NFK. The finding of LSC for untrained problems replicated the findings of the first experiment of Spelke and Tsivkin, which were not replicated in their second experiment. Regarding the neurophysiological data and therefore possible underlying mechanisms of LSC, results contrasted the previous data by Venkatraman et al. (2006). It was found that frontal and precentral regions, associated with increased executive and working memory demand, as well as parietal regions, associated with magnitude processing and calculation, were more activated when solving switching trials compared to no switching trials. As a result, it was argued that LSC arise due to additional numerical processing, which would suggest that participants calculate the problem at least in part anew. Since then, no further data on underlying mechanisms have been published.

The most recent study on LSC prior to the current research project was published in 2013 by Saalbach et al.. In contrast to previous studies, participants were pupils between 15 and 17, following an CLIL program in Switzerland. Their native language was German (L1). The second language in use for the study was French (L2). Participants had between 6 and 8 years of formal education in French, and were able to speak French fluently, with no knowledge of French prior to school entry. For the study, participants underwent three training sessions over a period of four days (participants were allowed to train either on day 2 or day 3) in NFK (i.e., multiplication and subtraction problems), either in German or in French. The test session took place right after the third training session on day four. Again, stimuli were provided in written number format and solutions provided by choosing among three options – same for training and testing. As in the previous study by Grabner et al. (2012), the test-design was a random trial-by-trial switching between tasks and languages. Testing consisted of four blocks containing 28 trials each, with each single problem presented four times. Results showed no differences regarding RT and ACC between the two training groups. Trained problems were solved faster and more accurate than untrained problems. LSC were found for RT for trained multiplication and subtraction problems. Regarding untrained problems, LSC were only found for multiplication. Further, LSC were stronger for the German compared to the French training group (i.e., switching from the native language to a foreign language led to stronger LSC). ACC was higher for trained compared to untrained problems. For the first time, LSC were also found for ACC. This was true for trained as well as untrained problems, with no differences between the two language-groups (i.e., training in German vs. training in French). For an overview, see Table 1 for a summary of previous findings on LSC in the field of arithmetic.

Table 1. *Summary of the key features of previous studies on LSC in the field of arithmetic and LSC found..*

	Spelke & Tsivkin (2001)	Venkatraman et al. (2006)	Grabner et al. (2012)	Saalbach et al. (2013)
Language 1	Russian	English	German	German
Language 2	English	Chinese	Italian	French
Training	2 days	5 days	4 days	3 days
Task 1 Task 2	Exact calculation Approximation	Exact calculation N/A	Exact calculation N/A	Exact calculation N/A
Stimuli in form of	Written-number words	Written-number words	Written-number words	Written-number words
Testing via	Verification task	Verification task	Verification task	Verification task
LSC for ACC	No	N/A	No	Yes
LSC for RT	Exact Calculation: Yes Approximation: No	Exact Calculation: Yes Approximation: No	Exact Calculation: Yes	Exact Calculation: Yes
LSC from L1 to L2	Yes	Yes	Yes	Yes
LSC from L2 to L1	Yes	Yes	Yes	Yes (weaker than from L1 to L2)
LSC for untrained problems	1st Experiment:Yes 2nd Experiment: No	N/A	Yes	Yes
Underlying mechanisms of LSC	N/A	Translational processes	Numerical processing, WM, EF	N/A

Note. N/A: not applicable; WM: working memory; EF: executive functions

What do we know about LSC

Aforementioned research on LSC revealed that performance is influenced when previously acquired knowledge has to be retrieved in a language different from the language of acquisition. The following bullet points summarize the appearance of LSC in relation to RT.

LSC appear:

- ✓ when testing different language combinations (i.e., Russian (L1) vs. English (L2), English vs. Chinese, German vs. Italian, German vs. French)
- ✓ in both directions (i.e., from L1 to L2 and vice versa; cf. Saalbach et al., 2013, for mixed results).
- ✓ for balanced and unbalanced bilinguals.
- ✓ across different training lengths (i.e., 2, 4 or 5 days)
- ✓ using different test-designs (i.e., block-wise language switching, trial-by-trial language switching)
- ✓ for exact calculation (i.e., NFK)
- ✓ for untrained problems in exact calculation
- ✓ not for approximation tasks (e.g., estimating cube roots)
- ✓ using visual stimuli in form of written number words
- ✓ using a verification task to assess RT

Thus, there is already profound evidence that NFK is at least to some extent acquired in a language-dependent way. Despite conveying powerful insights on LSC, findings likewise raise new questions. The following sections shed light on the limitations of previous research, open questions and new approaches to broaden the empirical evidence on LSC.

What do we not know about LSC

The area of the unknown will be split into four fields of interest. First, there are methodological characteristics of previous research that may or may not have led to LSC. In the same vein, these methodological features may lead to difficulties drawing implications for the field. Second, there is missing information about the underlying mechanisms of LSC. Third, LSC were mainly investigated for NFK. How about other types of knowledge, such as procedural knowledge? Fourth, little to nothing is known about individual characteristics that may lead to LSC on an individual level. In the following, the four fields will be discussed in more detail. Each section will already include the ways with which the studies of the present project will encounter these issues.

Methodological issues of previous research

With respect to ecological validity, the current data on LSC give rise to the question on how conclusions can be drawn from these laboratory studies to the field. Specifically, in all studies, subjects were confronted with tasks in a way that they will never encounter in the real world. In order to place each mathematical problem into a linguistic context, numbers were presented in written number format. This was supposed to secure the language dependent knowledge acquisition, which intuitively seems to be a great idea, but inevitably diminishes the ecological validity of the outcomes. Further, for the analyses of data, written number words come with the problem that individual reading speed becomes an issue. This issue makes the data hard to interpret, since the advantage for matched training and testing language may come from being trained in reading number words in that specific language. Grabner et al. (2012), as well as Saalbach et al. (2013) tried to tackle this concern by including a number reading exercise for both languages before the training sessions. Nonetheless, the amount of number reading was larger for the specific language groups in the end. For the present project we decided to use auditory stimuli continuously throughout all three studies. On the one hand, we hope to investigate LSC in a more representative way. Daily life usually demands solving mathematical problems in an auditory context, such as when being asked in a classroom conversation. On the other hand, we chose auditory stimuli to broaden our knowledge on LSC to another context. Hence, we will extend the knowledge of LSC from visual stimuli to auditory stimuli.

Another aspect in which all previous studies coincide is the test format for data collection. Subjects were asked to choose as fast as possible between two or three options. This kind of verification task is certainly easy to install, especially for examination in fMRI, since movement patterns are reduced. Nonetheless, there is a risk that subjects may a) work with an exclusion procedure, or b) have the opportunity to guess. This could be particularly the case towards the end of the testing session, when subjects feel a certain tiredness or even boredom. Here there is the danger that guessed trials will be included in the analysis, influencing the measures of RT and ACC. At this stage, it is speculation, but it cannot be ruled out. Again, we are changing the approach to tackle potential problems from previous studies and to expand knowledge about LSC using alternative methodological features. In all three studies, we will switch from verification tasks to production tasks (i.e., participants have to produce the solution themselves, no options being available). Study 1 will ask participants to type in the answer themselves, diminishing the likelihood that a correctly given answer was due to guessing. In studies, 2 and 3 we are going even one step further and work with oral responses. With the help of oral responses in studies 2 and 3, we can thus also check that the result was in the correct language. Within previous research, as well as Study 1, it cannot be ruled out that participants did not yet solve the task when pressing the key. It may be even the case that the problem is present in the wrong language. By giving the oral response, we do know for sure if the collected RT represents the moment of having the answer present. So overall, we thereby almost exclude the possibility of randomly correct trials being included in the analysis. Lastly, the oral response represents a more realistic situation, such as an oral exam, or ordinary classroom conversation. Finally, we decided to give participants sufficient time to answer each trial. Therefore, we wanted to lower the possibility that our data are affected by individual stress levels. If participants are put under pressure by only having a short amount of time to answer – which was done in previous research –, errors can be triggered. Certainly, this will be true for both conditions (i.e., no switching, switching), still, it may affect the sensitive measures of RT and ACC and in consequence even the appearance of LSC. In sum, Study 1 will be setting the groundwork by changing several methodological aspects to assess LSC within the laboratory context. Studies 2 and 3 will then add specific features.

Underlying mechanisms of LSC

The second area of interest involves the underlying mechanisms of LSC. To date, only two studies tried to investigate this question with the help of neurophysiological instruments (i.e., fMRI). Unfortunately, the two studies show different results, leading in one study to the assumption of additional translational processes (Venkatraman et al., 2006), while in the other study to additional numerical processing. Already in Study 1, we will integrate artificial tasks, which represent pure fact learning, but are comparable in their designed format with conventional arithmetical tasks. Are LSC present for tasks that are represented as pure fact knowledge? If so, then it would at least rule out that numerical processing alone is the underlying mechanism of LSC. Additionally, the question of underlying mechanisms will be the main focus of Study 2, in which participants will give insight in their proceeding via two different self-reports on a trial-by-trial basis. Against common criticism (e.g., Kirk & Ashcraft, 2001), such self-reports were tested to present a valid method to assess individual strategy use in arithmetic (e.g., Grabner & De Smedt, 2011).

LSC for procedural knowledge

Research on LSC almost exclusively concentrated on NFK. Very little to nothing is known about procedural knowledge. Within the context of mathematics, NFK is related to knowing the answer to a given problem, whereas procedural knowledge relates to knowing the series of steps or rules in order to come up with the answer of that given problem (Canobi, 2009; Rittle-Johnson, Schneider, & Star, 2015). To this end, there is contradictory evidence on procedural knowledge and LSC. In 2001, Spelke and Tsivkin examined an estimation task (i.e., estimating the cube root of a number), finding no LSC. It was concluded that there is neither a facilitation nor an impairment when language of instruction and application differ for approximating. In contrast, two studies found LSC for untrained multiplication and subtraction problems (Grabner et al., 2012; Saalbach et al., 2013). This most likely reveals that calculation in the no switching condition was accelerated by the simple fact that the same language was used to solve arithmetic problems of the same operation before. This finding may suppose a language-dependency of procedural knowledge. However, as the authors pointed out, methodological characteristics of the studies may lead to these effects, such as the fact that participants had more training to read number words in the no-switching condition. This may explain the acceleration in the no switching condition. Overall, it is very difficult to derive conclusions from these studies since participants did not truly learn a new procedure as well as the missing data on strategy use. Taken together, the explanatory power of existing data remains vague regarding LSC for procedural knowledge. After all, no experimental study exists that turned its attention truly on learning the single steps of a new procedure in a foreign language while being tested later in their mother tongue. Yet, this is exactly what is happening in the field: within the CLIL context, pupils are on a regular basis challenged to learn new procedures from scratch in a foreign language. How do they perform when applying these new procedures in a different language context? Study 3 will try to approach this interesting question.

Individual Characteristics and LSC

The fourth and final main field of interest in our project is related to the question of individual characteristics that promote or prohibit LSC. The question arises if there are individual characteristics that make one person more likely to show LSC than another. Thus, by revealing connections between individual characteristics and LSC, we may also get more insight about possible underlying mechanisms. Hitherto, mixed evidence exists for a possible directional effect. Directional effect means that switching from L1 to L2 will lead to LSC, while switching from L2 to L1 does not. Saalbach et al. (2013) found weaker LSC for the latter. Still, LSC were present. Marian and Fausey (2006), though not in the context of arithmetic, found no LSC for the study group that was considered as being balanced bilingual. In this case, LSC were found in both directions, but only if the participants were almost equally fluent in both languages. Thus, as already pointed out in previous research, the language ability of L2 may play a role. Unfortunately, either the studies did not measure language ability of L2 (e.g., Grabner et al., 2012; Saalbach et al., 2013), or they contained a sample of balanced bilinguals, not having any variety to investigating the relationship (e.g., Spelke & Tsivkin, 2001; Venkatraman et al., 2006). The need to consider language ability is further supported by the evidence that translational processes may underlie LSC (Venkatraman el al., 2006). In contrast, considering the divergent evidence that additional numerical processing may cause LSC (Grabner et al., 2012), a measure for arithmetic ability may provide further insight. In all three studies of the present project, individual characteristics will be assessed during a pre-meeting before the actual training starts. Therefore, we hope to gain further insights into LSC and individual differences.

Research objectives of this dissertation

The general aim of the present project is to gain more insights into the nature of LSC. First and foremost, we explore whether LSC occur when the specific methodological changes in our laboratory studies are changed. Therefore, Study 1 will be crucial by setting the basis for the whole project. Second, we will use auditory instead of visual stimuli (Study 1, 2, and 3). Third, we examine whether LSC occur when the testing format is changed (i.e., using a verification task instead of a production task; Study 1, 2, and 3). Previous studies have shown no variation here. Fourth, we will test different block designs during the test session (block-wise language and task switching in Study 1; trial-by-trial language and task switching in Study

2 and 3). Fifth, we are investigating especially the German context, thus, using German and English as language combination (Study 1, 2, and 3), which has never been done before. Sixth, we examine whether LSC occur in the context of procedural learning (Study 3). Seventh, each study will include test batteries to assess individual characteristics.

We are aware that questions are likely to remain open after this project. In the course of this project, it was of special interest us to proceed in a step-wise and precise manner, replicating results of the previous studies as far as possible, rather than continue working on the basis of unique findings. With regard to the subject content, the project will focus on the field of arithmetic, mainly declarative knowledge (i.e., NFK; Study 1, 2, and 3) as well as procedural knowledge (Study 3).

Study 1[1]

Language-dependent knowledge acquisition: investigating bilingual arithmetic learning

Christian G. K. Hahn

Institute of Psychology, University of Göttingen, Göttingen, Germany

Faculty of Education, University of Leipzig, Leipzig, Germany

Henrik Saalbach

Faculty of Education, University of Leipzig, Leipzig, Germany

Roland H. Grabner

Institute of Psychology, University of Graz, Graz, Austria

Abstract

Previous studies revealed language-switching costs (LSC) in bilingual learning settings, consisting of performance decreases when problems are solved in a language different from that of instruction. Strong costs have been found for arithmetic fact knowledge. The aim of the present study was to investigate whether LSC in arithmetic also emerge in an auditory learning task and in pure fact learning. Furthermore, we tested whether LSC are influenced by the direction of language-switching. Thirty-three university students learned arithmetic facts of three different operations (i.e., multiplication, subtraction, artificial facts) over a period of four days. The training was either in German or English. On day five, participants solved problems in both languages. Results revealed LSC in response latencies for all three types of problems, independent of the direction of language-switching. These findings suggest that LSC are modality-unspecific and occur independent of the type of arithmetic fact knowledge.

Key words: bilingual learning, language-switching costs, arithmetic learning, fact knowledge

[1] Published in a peer-reviewed journal: Hahn, C. G., Saalbach, H., & Grabner, R. H. (2017). Language-dependent knowledge acquisition: investigating bilingual arithmetic learning. Bilingualism: Language and Cognition, 22 (1), pp. 47-57. https://doi.org/10.1017/S1366728917000530

Introduction

Bilingual learning receives increasing attention by both public and research. One well-known example within the educational field is Content and Language Integrated Learning (CLIL), which represents a dual-focused instructional approach to teach content while simultaneously improving language skills in a foreign language (Eurydice, 2006; Lasagabaster & Sierra, 2010). For example, mathematics or geography are taught in English to German native speakers who have learned English as a second language. Despite the great success of these programs to foster language learning (e.g., Zaunbauer, Bonerad, & Möller, 2005; Zaunbauer & Möller, 2009), it is an open question whether and to what extent the acquired knowledge is represented in a language-dependent or language-independent way. This question is not only of theoretical but also of practical relevance. Language-dependent knowledge representations may cause cognitive costs if the language of instruction differs from the language of knowledge retrieval and application. For instance, a student who acquires mathematical knowledge in a foreign language may not be able to use this knowledge in his native language as effectively as when he had learned it in his mother tongue. The costs commonly consist of longer solution times and higher error rates. So far, the so called language-switching costs (LSC) have been reported for retrieving arithmetic (Spelke & Tsivkin, 2001; Grabner, Saalbach, & Eckstein, 2012; Saalbach, Eckstein, Andri, Hobi, & Grabner, 2013), and other numerical and non-numerical fact knowledge (Marian & Fausey, 2006), as well as recalling autobiographic information (Marian & Neisser, 2000). The present paper aims to further investigate the extent, correlates and mechanisms of LSC in the domain of arithmetic.

Language and knowledge representation in arithmetic

Language affects how people process information and knowledge is stored in memory (e.g., Gentner & Goldin-Meadow, 2003; Gumperz & Levinson, 1996; Malt & Wolff, 2010; Wolff & Holmes, 2011 for review). As a consequence, cognitive differences between speakers of different languages can be detected across a wide range of domains (e.g., Boroditsky, Fuhrman & McCormick, 2010; Fausey & Boroditsky, 2011; Saalbach & Imai, 2007). For mathematics, Miller, Smith, Zhu and Zhang (1995) found that the structure of the numerical system affects how quickly children develop basic counting and arithmetic abilities. For instance, compared to Chinese children, U.S. children had more problems understanding the base-10 structure, committing more counting errors (e.g., counting “twenty-eight, twenty-nine, twenty-ten, twenty-eleven"; see also Fuson & Kwon, 1992; Park, 1999 for an overview). In addition, the phonological structure of number words affects performance. For instance, cross-language performance differences have been reported between Mandarin and English (Chen, Cowell, Varley, & Wang, 2009) and between English and Welsh speaking language groups (Ellis & Hennelly, 1980). In the study by Chen and colleagues (2009), thirty native Mandarin Chinese and thirty native English speakers were tested on verbal and visuo-spatial working memory span (e.g., forward and backward digit span task). Results revealed significantly higher scores in the Mandarin Chinese speaking group for verbal working memory span than in the English-speaking group. The advantage of Mandarin was associated with the shorter articulation time for digits in spoken Mandarin Chinese. In arithmetic, the association between language and numerical cognition has been found predominantly for exact calculation (exact solution of an arithmetic problem) rather than approximate calculation (Dehaene & Cohen, 1997; Spelke & Tsivkin, 2001; Lemer, Dehaene, Spelke, & Cohen, 2003). These findings are in line with neuroimaging studies, showing that the retrieval of (exact) arithmetic facts is in close connection to brain circuits associated with language processing and storage of verbal information (e.g., Lee, 2000; Dehaene, Molko, Cohen, & Wilson 2004; Domahs & Delazer, 2005; Venkatraman, Siong, Chee, & Ansari, 2006; cf. Benn, Zheng, Wilkinson, Siegal, & Varley, 2012; Klessinger, Szczerbinski, & Varley, 2012).

In bilinguals, arithmetic knowledge seems to be strongly related to the language of acquisition, which is typically the mother tongue. For instance, Frenck-Mestre and Vaid (1993) required bilingual participants to perform simple addition problems (e.g., 2 + 5) as well as simple multiplication problems (e.g., 7 x 3). Performance was slower and less accurate when calculating in their second language (L2) than in their first language (L1). Similarly, German-French bilingual adolescents showed better performance when arithmetic tasks were presented in L1 (German) compared to L2 (French), even though later, in secondary education, mathematics had been taught in French. The effect was greater for complex addition problems (e.g., 56 + 32) compared to more simple addition problems (e.g., 4 + 2; van Rinsveld, Brunner, Landerl, Schiltz & Ugen, 2015). Taken together, research in the field of bilingual mathematics learning suggest that language is relevant for task performance. What is the implication for bilingual learning settings when language of encoding and language of retrieval differ?

Bilingual arithmetic learning and language-switching costs

According to the ENCODING-SPECIFICITY HYPOTHESIS the effectiveness of retrieving facts from memory is in close relation to the context in which information had been encoded (e.g., Barber, Rajaram, & Aron, 2010; Tulving, & Thomson, 1973). With respect to bilingual learning, this would suggest that the retrieval and application of knowledge is most effective in the language of encoding. When a person needs to solve a task in a language that is different from the language of encoding (or instruction, respectively), cognitive costs may emerge. Such LSC have been reported in previous research (Spelke & Tsivkin, 2001, Grabner et al., 2012; Saalbach et al., 2013). Spelke and Tsivkin (2001), for example, had Russian-English bilinguals undergo two training sessions consisting of different set of problems including exact calculations (e.g., "What is the sum of fifty-four and forty-eight?"), and approximation tasks (e.g., "Estimate the approximate cube root of twenty-nine!"). The testing situation included two kind of verification tasks in which participants had to decide which one was the exact answer (exact number task), or which one is closest to the exact number (approximation number task). LSC were specific to the exact number tasks as opposed to the approximation tasks as well as to a third task, including non-numerical information. The authors concluded that exact arithmetic is more strongly language-dependent than approximate arithmetic. Saalbach and colleagues (2013) investigated to what extent LSC in arithmetic are moderated by the arithmetic operation and whether they generalize to untrained problems. Thirty-nine bilingual high school students underwent a three-day training of fourteen multiplication and fourteen subtraction facts either in German (L1) or in French (L2). During training and test, problems were displayed in number-words (e.g., "twelve times seven"). In the test session, participants were presented with the trained as well as untrained problems in both languages. Results revealed that participants had longer RT as well as lower ACC for both multiplication and subtraction problems when language-switching was required. To notice, LSC for the trained problems did not depend on the arithmetic operation. This was unexpected, since it is commonly argued that multiplication problems rely more strongly on a verbal coding than subtraction problems, which are associated with mental manipulation of magnitude (e.g., Dehaene et al., 2004; Ischebeck, Zamarian, Siedentopf, Koppelstätter, Benke, Felber & Delazer, 2006). Thus, by manipulating the language, stronger LSC for multiplication problems had been expected. Interestingly, LSC also emerged in the untrained problems, suggesting that the impact of the language of instruction may not only affect fact retrieval but also the recall of other kinds of knowledge such as procedural knowledge. In addition, LSC were stronger when participants switched from their dominant language (L1, German) to the non-dominant language (L2, French) than vice versa (see also Marian and Fausey, 2006, for similar findings). The mechanism underlying LSC in arithmetic were investigated by Grabner et al. (2012). They used functional magnetic resonance imaging (fMRI) to scrutinize which neuro-cognitive processes might be associated with LSC. During a four-day training, twenty-nine participants learned ten subtraction and ten multiplication facts presented in number-words either in German or Italian. Throughout the test session, participants had to solve trained and untrained problems in both languages. In line with Saalbach and colleagues (2013), LSC were found both for trained and untrained problems in RT and ACC as well as for multiplication and subtraction problems. Moreover, results revealed an association between LSC and activation in areas related to magnitude processing, implying that LSC may be due to additional numerical processing rather than to mere language translation. As for the behavioral results, the association between LSC and neural correlates was independent of the arithmetic operations.

In sum, previous studies on language-dependency in arithmetic learning consistently reveal LSC in RT and ACC. In addition, LSC appear to be independent of the arithmetic operation, arguing for a similar cognitive cause regarding rote learned information (i.e., fact knowledge). Furthermore, LSC were found for different language combinations, highlighting the important role that a mismatch of the language of instruction and language of application can have on performance. Findings also suggest a directional effect in that LSC are higher when switching is required from L1 (first language) to L2 (second language).

Even though previous research has provided first important insights into the language-dependency of knowledge representation in arithmetic, some methodological limitations and open questions need also to be taken into consideration. First, in all three studies on arithmetic, stimuli were presented in written form (e.g., “three times twelve”), which is hardly used in educational practice and thus represents a substantial limitation of ecological validity. Second, verification tasks were used, which do not resemble authentic arithmetic problem solving and may even produce undesired effects. Indeed, the solutions to problems could be guessed instead of calculated by applying certain strategies (e.g., eliminating obviously wrong answers). Moreover, verification tasks produce an interference effect in which RT increase and ACC decrease the closer the numerical distance is between the correct answer and the distractor (Ashcraft & Bataglia, 1978; Ashcraft & Stazyk, 1981). Third, the cognitive mechanism underlying LSC in processing arithmetic problems is still unclear. Based on neuroimaging data as stated previously, Grabner and colleagues (2012) concluded that LSC may be the result of additional quantity processing (such as calculation) rather than mere translation into the testing language after fact retrieval in the language of training. However, earlier findings by Marian and Fausey (2006) revealed that LSC also apply to the retrieval of non-arithmetic knowledge, showing that mere quantity processing is unlikely to account for the appearance LSC alone. Finally, the potential interaction of second language proficiency and LSC requires exploration. Marian and Fausey (2006) argued that participants rely more on the higher-proficiency language during the encoding phase, therefore finding higher LSC when switching from the dominant to the non-dominant language than vice versa. However, language proficiency was assessed by means of self-reports to categorize participants as dominant or non-dominant speakers but not with an objective measure of language proficiency. As other research revealed, language-proficiency is critical for cognitive performance across different domains of academic learning including mathematics (e.g., Kempert, Saalbach, & Hardy, 2011). Thus, it is important to take language proficiency into account when studying LSC within arithmetic learning.

The main aim of the present study was to further our knowledge about the language-dependency of arithmetic knowledge and the nature of the LSC. In particular, we first investigated whether previous findings in German-English bilinguals can be replicated by using the ecologically more valid auditory stimuli (research question 1). Second, we further examined the mechanisms underlying LSC by comparing the extent of LSC after learning artificial vs. real arithmetic facts (research question 2). Specifically, in addition to multiplication and subtraction problems, we included artificial problems requiring pure fact learning (i.e., ## box # = ##). Arithmetic problems, even if extensively trained, may not only be solved by fact retrieval but also by other (e.g., magnitude-related) processes. These processes have even been discussed as a major cause of LSC in arithmetic (Grabner et al., 2012). In the artificial problems, however, such alternative strategies can be precluded. Third, we investigated whether the extent of LSC depend on the direction of language switching (from L1 to L2 or v.v.; research question 3). Finally, we explored whether and to what extent an indicator for L2 proficiency modulates the size of the LSC (research question 4).

We hypothesized that problems for all three tasks involving auditory material are solved more slowly and less accurately when the language of instruction differs from the language of application (i.e., when language-switching is required). (Hypothesis 1). Furthermore, we predicted LSC to appear for all three tasks (i.e., multiplication, subtraction and artificial problems) since all problems are likely to represent fact knowledge after a training period of four days, independently of their individual type (Hypothesis 2). In line with previous research

(Marian & Fausey, 2006; Saalbach et al., 2013) we expected to find stronger LSC when knowledge, which has been encoded in the dominant language, is retrieved in the non-dominant language as compared to a situation when knowledge, acquired in the non-dominant knowledge, needs to be retrieved in the dominant language (Hypothesis 3).

Methods

Participants

Thirty-three university students at the University of Göttingen, Germany, underwent the training and test procedure. One participant had to be excluded due to missing the last training session. The final sample consisted of 32 participants, with 20 being female and 12 being male. Half of the participants received the training in German (dominant language, L1), the other half in English (non-dominant language, L2). All participants were native German-speaking and had at least seven years of formal English education. The LexTALE test for vocabulary knowledge showed a mean score of 73% (SD = 12), which is supposed to indicate language proficiency at an upper intermediate level (B2; Lemhöfer & Broersma, 2012).

Material

Three different arithmetic tasks were included as experimental stimuli, comprising a) six multiplication problems: two-digit times one-digit numbers with two-digit solutions (00 x 0 = 00); b) six subtraction problems: two-digit minus two-digit numbers, including only carry-over calculations with two-digit solutions (00 - 00 = 00); and c) six artificial problems: two-digit and one-digit numbers connected via an arbitrary symbol (box) with two-digit solutions (00 box 0 = 00). The multiplication and subtraction problems resembled those used in previous arithmetic training studies (Delazer, Domahs, Bartha, Brenneis, Lochy, Benke, & Trieb, 2003; Ischebeck et al., 2006; Grabner & De Smedt, 2012; Saalbach et al., 2013). The first operand ranged from 14 to 17, and the second operand ranged from 3 to 7, keeping two-digit outcomes, and excluding solutions divisible by 5 or 10. Subtraction problems matched the multiplication problems regarding their difficulty (Ischebeck et al., 2006) to be comparable to each other. Artificial problems had the same structure as the multiplication problems – while having different operands – to account for best comparability. Auditory stimuli were created with the professional audio software Voice Reader Studio 15, a widely used text-to-speech program

(Linguatec, 2015). The training and test program was created using *E-Prime 2.0 Professional stimulus presentation software* (Schneider, Eschmann, & Zuccolotto, 2002). The online version of the LexTALE was chosen to indicate general English vocabulary knowledge. It has been developed to account for the increasing need in experimental studies to assess language vocabulary knowledge within a short time scale and been validated by Lemhöfer and Broersma (2012). The test consists of 60 items (40 words, 20 nonwords). Nonwords are orthographically legal and pronounceable, but represent nonsense strings. Participants have to indicate whether the word is an existing English word or not. Assessing language proficiency in its full detail would go beyond the scope of our present study. Nevertheless, this vocabulary test has not only been shown to be a better predictor than commonly used self-ratings for vocabulary knowledge (Lemhöfer & Broersma, 2012), but also shows a substantial correlation with more widely used vocabulary tests (e.g., r = .80 in Mochida & Harrington, 2006) as well as with the Quick Placement Test (QPT) that is frequently used to predict language proficiency (Quick Placement Test, 2001).

Training Procedure

Throughout the training period, participants were trained and tested in a room of the local Institute for Psychology. Training and testing was held in group sessions with up to four participants. Participants used headphones and were seated separately in cabins to eliminate distracting factors. Participants were randomly assigned to either the German training group or the English training group. They were instructed to learn 18 problems (i.e., 6 multiplications, 6 subtractions, and 6 artificial problems) over a period of four consecutive days. Instructions were given in written form in an extra session on the day of the first training session. Instructions were also repeated before each session to guarantee a problem-free run. Each training session consisted of three task-blocks (i.e., multiplication, subtraction, and artificial; see Figure 2a) including the six problems with six repetitions each. Thus, 108 problems were solved in each training session (6 problems x 6 repetitions x 3 tasks). At the end of the four training days, each problem had been repeated 24 times respectively. This number of repetitions has previously been shown to produce strong training effects (Ischebeck et al., 2006; Grabner et al., 2012; Saalbach et al., 2013). A one-minute break was included after each task-block. Overall, one training session lasted about 20 to 25 minutes. The order of the six different problems within each task block was randomized. No problem was repeated two times in a row. Trials started with a white fixation point for 1000 milliseconds on a black screen, followed by the auditory presentation of the problem while the screen remained black without the fixation point (see Figure 1). Participants were asked to press the ENTER key when having the correct answer in the instructed language ready to speak out loud. The maximum time-frame to answer a problem was ten seconds from the start of the auditory stimuli. Next, the correct answer had to be typed in via a key-pad, confirming it with a key-press within five seconds. The typed numbers were visible on the screen in order to correct the answer if needed. Corrective feedback was given by a green or red display, followed by the auditory presentation of the correct answer, irrespectively of whether the given answer was correct or not. Thereby, we attempted to strengthen the training process and the connection with the training language. Further, the auditory feedback was essential in order for the artificial problems to be learned in the first place. Before the next trial started, an inter-trial-interval (ITI) of one second appeared. RT and ACC were recorded.

Test Procedure

On day five, participants underwent the test session, taking about 45 minutes. Compared to the training session, each block was presented in both, the trained language (no language-switching required) and the untrained language (language-switching required). The test consisted of six task blocks. In order to avoid additional executive-control processes and unlike in previous studies, participants did not switch between operations or languages within blocks. Each block consisted of 36 problems which were either presented in the language of training or in the untrained language. Test order was counter-balanced, so that half of the sample started the test session with a task-block (e.g., multiplication) in the no language-switching condition (ns for no-switching) followed by a block of the same task (e.g., multiplication) in the language-switching condition (s for switching; order A) and vice versa (order B, see Figure 2b). Neither corrective feedback nor auditory presentation of the correct results was provided. RT and ACC were recorded.

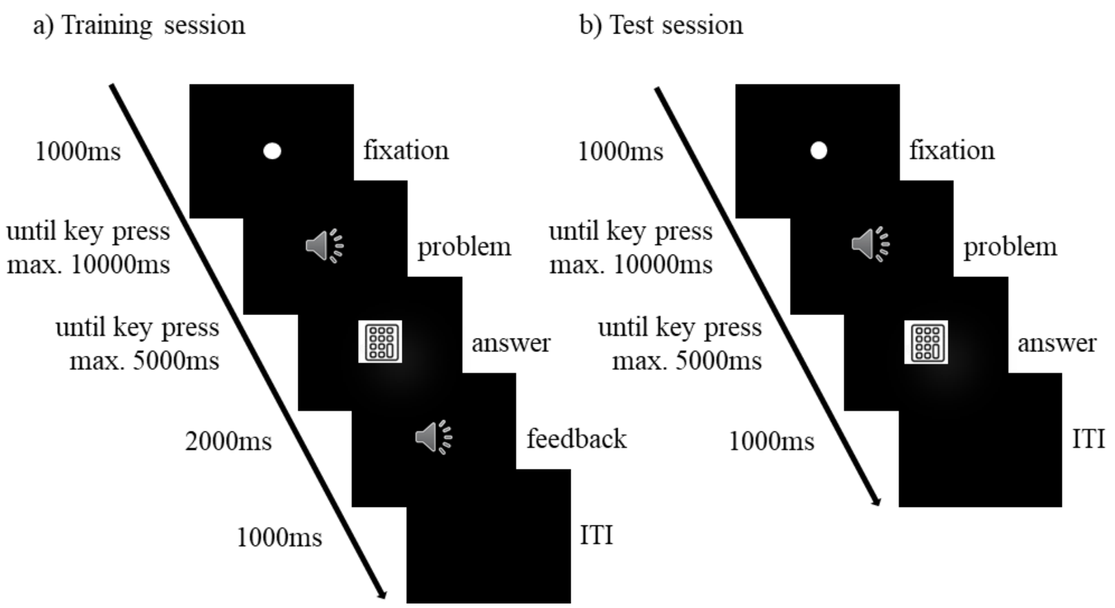


Figure 1. Schematic display of the trial time course during (a) the training and (b) the test session.

a) Training session

	Block 1	Block 2	Block 3
Training 1	ART	MUL	SUB
Training 2	MUL	SUB	ART
Training 3	ART	SUB	MUL
Training 4	SUB	ART	MUL

b) Test session

	Order A	Order B
Block 1	ART (ns)	ART (s)
Block 2	ART (s)	ART (ns)
Block 3	SUB (ns)	SUB (s)
Block 4	SUB (s)	SUB (ns)
Block 5	MUL (ns)	MUL (s)
Block 6	MUL (s)	MUL (ns)

Figure 2. Schematic display of the block design for a) the training session and b) the test session. Within the test session, the two different orders are represented (ns: no switching condition; s: switching condition).

Data Analysis

For statistical analyses, IBM SPSS Statistics 20 was used. To analyze the data, mixed design ANOVAs for RT and ACC were computed. To analyze the training data, the ANOVA contained the two within-subject factors Arithmetic Task (artificial vs. multiplication vs.

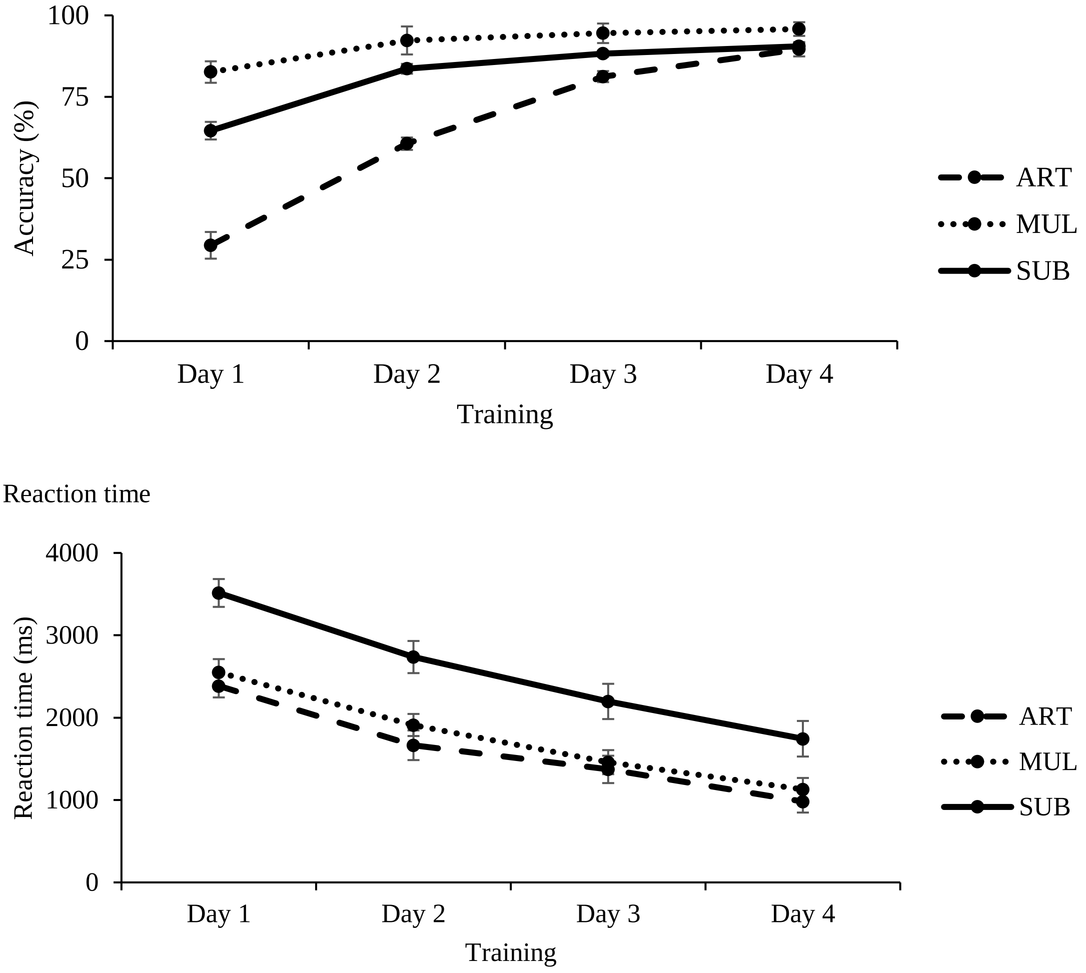
subtraction) and Training Day (day 1 vs. day 2 vs. day 3 vs. day 4), and the between-subject factor Language of Training (German vs. English). The ANOVA for the testing data contained the two within-subject factors Arithmetic Task (artificial vs. multiplication vs. subtraction) and Language Switching (switching vs. no switching), as well as the between-subject factors Language of Training (German vs. English). Test Order (i.e., switching followed by no-switching vs. no-switching followed by switching) was included as covariate. RT data was only analyzed for correct trials. For effect sizes, Cohen's *d* and partial eta-squared ($\eta_p{}^2$) were computed. In case of violation of the assumption of sphericity (Mauchly's test), degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity.

Results

Training data

Training data for ACC and RT are displayed in Figure 3. For ACC, analyses revealed a main effect for the factor Training Day**:** *F*(2.10, 62.84) = 142.96, *p* < .001, $\eta_p{}^2$ = .83. Post-hoc t-tests showed a persistent effect for each consecutive day (all *p*s < .001), indicating that ACC increased significantly in the course of the training. Moreover, a main effect was found for Arithmetic Task, *F*(1.40, 41.81) = 64.33, *p* < .001, $\eta_p{}^2$ = .68, indicating that ACC were not equal for all three arithmetic tasks. Post-hoc t-tests showed that – averaged over all four days – artificial problems were solved less accurate than multiplication problems (65% vs. 91%; *t*(31) = -10.962, *p* < .001, *d* = 1.94) and subtraction problems (65% vs. 82%; *t*(31) = -5.831, *p* < .001, *d* = 1.03), while subtraction problems were solved less accurate than multiplication problems (82% vs. 91%; *t*(31) = -6.601, *p* < .001, *d* = 1.17). Finally, an interaction effect was observed for Arithmetic Task and Training Day**,** (*F*(3.85, 115.50) = 44.03, *p* < .001, $\eta_p{}^2$ = .59, indicating that the magnitude for the daily training effects was different depending on the arithmetic task.

RT-analyses revealed a main effect for Training Day, *F*(1.89, 56.61) = 113.85, *p* < .001, $\eta_p{}^2$ = .79. Post-hoc t-tests showed that the effect was persistent for each consecutive day (all *p*s < .001), indicating that response time decreased significantly in the course of the training. We further found a main effect for Arithmetic Task, *F*(2, 90) = 33.63, *p* < .001, $\eta_p{}^2$ = .53. Post-hoc t-tests showed no significant difference in RT between artificial and multiplication problems (1601 ms vs. 1762 ms, *t*(31) = -1.239, *p* = .225, *d* = 0.22), but between artificial problems and subtraction problems (1601 ms vs. 2548 ms; *t*(31) = -6.858, *p* < .001, *d* = 1.21), and between multiplication problems and subtraction problems (1762 vs. 2548 ms; $t(31) = -6.615$, $p < .001$, $d = 1.17$). Finally, we found an interaction effect between Arithmetic Task and Language of Training, $F(2, 60) = 3.99$, $p = .024$, $\eta_p^2 = .12$. Post-hoc analyses revealed that subtraction problems account for this interaction since they were solved more slowly in the English training group than in the German training group (2928 ms vs. 2167 ms; $t(30) = -2.12$, $p = .042$, $d = 0.37$).


a) Accuracy

b) Reaction time

Figure 3. Training data for a) ACC and b) RT. Separate lines represent the three different tasks. Error bars indicate the standard error (SE).

Test data

The ACC- and RT-results are shown in Table 2.

Table 2. *Mean RT in milliseconds (upper rows) and ACC in percentage correct (lower rows) as a function of arithmetic task, and type of switching condition. Standard errors are enclosed in parentheses.*

	Artificial	Multiplication	Subtraction
RT in milliseconds			
No language switching	889 (117)	973 (154)	1715 (286)
Language switching	1156 (141)	1240 (162)	1879 (249)
ACC in percentage correct			
No language switching	93.5 (1.9)	97.5 (0.9)	92.3 (1.8)
Language switching	92.8 (2.4)	97.2 (0.9)	90.6 (2.4)

Hypothesis 1: Problems are solved more slowly and less accurately when the language of instruction differs from the language of application.

ACC and RT data for language-switching are depicted in Figure 4. For ACC, no effect for Language Switching was found ($F(1, 29) = 1.36$, $p = .25$, $\eta_p^2 = .05$), demonstrating that the answers to language-switching problems (93.5%) were as accurate as the answers to problems were no language-switching was required (94.4%). For RT, however**,** we found the predicted main effect for Language Switching ($F(1, 29) = 7.26$, $p = .012$, $\eta_p^2 = .20$), indicating that problems requiring language-switching were solved more slowly than problems requiring no language-switching (1425 ms vs. 1192 ms).

Unexpectedly, a significant interaction effect between *language switching* and the covariate Test Order emerged ($F(1, 29) = 17.97$, $p < .001$, $\eta_p^2 = .38$). To break down this interaction, additional post hoc t-tests were performed, showing that LSC only appeared in test order B, when – for each operation separately (see Figure 2b) – participants were tested first in the language-switching condition followed by the no language-switching condition (1492 ms vs. 1044 ms; $t(16) = 5.94$, $p < .001$, $d = 1.44$). No significant LSC emerged for the other sequence of testing, i.e., when testing started with the presentation of problems in the no language-switching condition followed by the language-switching condition (1360 ms vs. 1350 ms; *t*(16) = 0.14, *p* = .89, *d* = 0.04). The comparisons between test order A and order B in the switching condition (1350 ms vs. 1492 ms; *t*(30) = -0.428, *p* = .67, *d* = 0.03) and in the no-switching condition (1360 ms vs. 1044 ms; *t*(30) = 0.951, *p* = .35, *d* = 0.17) did not reveal significant effects.

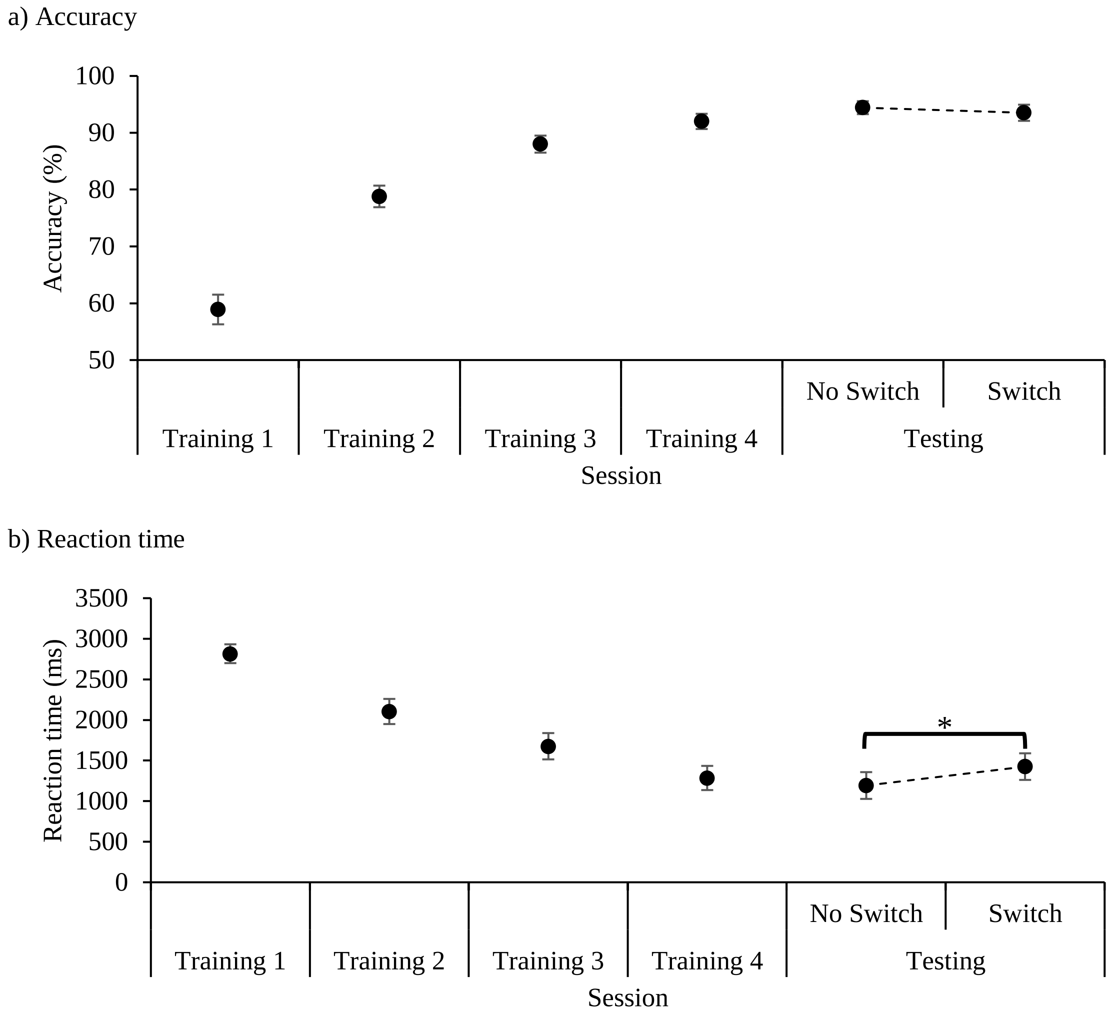


Figure 4. Illustration of a) ACC and b) RT from training to testing. Error bars presenting the standard error. **p* < .05. ***p* < .01.

Hypothesis 2: LSC emerge in all three task operations

As expected there was no interaction between Arithmetic Task and Language Switching (for RT: *F*(1.64, 47.65) = .48, *p* = .59, η_p^2 = .02, for ACC: *F*(1.31, 37.91) = .09, *p* = .83, η_p^2 < .01), indicating that LSC appear in all three included operations (i.e., multiplication, subtraction and artificial problems).

Hypothesis 3: More LSC when switching to the non-dominant language as compared to switching to the dominant language

Inconsistent to our prediction, no interaction was found between Language of Training and Language Switching (for RT: *F*(1,28) = 0.32, *p* = .86, for ACC: *F*(1,28) = 0.24, *p* = .63) indicating that LSC do not significantly differ across training languages.

Explorative analysis: Is L2 vocabulary knowledge related to LSC in general?

Individual scores for vocabulary knowledge were correlated with the respective LSC for each arithmetic task separately and for the overall LSC (i.e., including all three arithmetic tasks). There was no correlation between vocabulary knowledge and overall LSC (*r*(32) = -.08, *p* =. 68) nor between vocabulary knowledge and operation-specific LSC (for artificial problems: *r*(32) = -.15, *p* =. 42; for multiplication problems: *r*(32) = .20, *p* =. 28; for subtraction problems: *r*(32) = -.14, *p* =. 46).

Discussion

The main aim of the present study was to further investigate language-switching costs (LSC) in the domain of arithmetic. Therefore, thirty-two university students learned eighteen problems of three different arithmetic operations in German (L1) or English (L2) over four consecutive training days and were tested in both languages on the fifth day. We found significant LSC for RT but not for ACC. Results further revealed LSC for RT in all three task (i.e., multiplication, subtraction and artificial problems). However, LSC due to learning in the dominant language and retrieval in the non-dominant language did not differ from LSC due to learning in the non-dominant language and retrieval in the dominant language. Finally, there was no significant relation between vocabulary knowledge of L2 and LSC.

The present design provides an important extension of prior research. While previous studies on LSC in arithmetic learning used visual stimuli in the form of written number words (Spelke & Tsivkin, 2001, Grabner et al., 2012; Saalbach et al., 2013), the present study was the first to show that LSC appear when arithmetic problems are learned and tested auditorily. LSC using auditory stimuli is an important finding, since numerical information is commonly presented either auditorily or as digits rather than as words during instruction. Further, it was the first study to show LSC in a block-wise language switching design, compared to random switching of language and task within blocks (e.g., Grabner et al., 2012; Saalbach et al., 2013; see also Meuter & Allport, 1999; and Campbell, 2005, for studies on cued language switching). Especially, if we are interested in making implications for bilingual educational programs, a closer look on testing formats is necessary.

The first hypothesis, expecting problems to be solved more slowly and less accurately when the language of instruction differs from the language of application, was partly confirmed. In contrast to previous studies, LSC were limited to RT. The absence of LSC for ACC might be explained by adaptations made in the present study design, which led to a ceiling effect in ACC (ranging between 90% and 98%). The preceding studies used verification tasks, which required participants to choose among two or more answers. In the present study, a production task was administered in which participants had to type in their answers after they indicated the completion of problem solving by keypress. In addition, due to having only one language and one specific arithmetic operation within each block during testing, participants did not have to switch the language or operation type from trial to trial (but block-wise), which was required in the previous studies. This lower level of cognitive load within each block may have facilitated problem-solving, resulting in comparably high ACC for all three tasks, even in the language-switching condition.

Interestingly, another methodological change during the testing phase led to unexpected results regarding LSC. LSC were only found in test order B, when participants started with a block in the language-switching condition followed by a block in the no-switching condition (see Figure 2b). No LSC emerged in the reversed order A. It could be speculated that these results are due to a differential overlay of language-switching and practice effects. Overall, we found that RT for earlier trials within each block were significantly longer than for later trials, indicating a typical practice effect over the test session (post-hoc analysis showed a training effect within each block (all *p*s < .001, all *d*s > .98). In test order A, the practice effect may have counteracted the LSC resulting in similar RT in the switching-blocks (blocks 2, 4, and 6) compared to the respective no-switching blocks (blocks 1, 3 and 5). In test order B, however, the practice effect may have even amplified LSC as already the first blocks of each operation required language switching. Thus, despite of the clear advantages of the block design in examining LSC (e.g., preventing item-wise switching) it may partly have resulted in a confounding of practice and language-switching effects. This post-hoc finding and the following interpretation on practice effects remains still vague. It may give us a first insight on possible interventions to lower the likelihood of LSC within a short period of time. We may then ask the question whether only one or two short training session in the untrained language can prevent LSC to appear. Future studies may directly compare different designs to shed more light on the question of the robustness of LSC and possible interventions.

Regarding our second hypothesis, predicting LSC to appear for all three tasks, we found LSC for RT not only for typical arithmetic problems (i.e., multiplication and subtraction problems; replicating findings from Grabner et al., 2012, and Saalbach et al., 2013), but also for atypical arithmetic problems (i.e., artificial problems). Notably, LSC for artificial problems did not differ from LSC for multiplication or subtraction problems. This finding suggests that LSC cannot be solely explained by additional magnitude processing as suggested by the fMRI findings in previous research (Grabner et al., 2012). To identify underlying mechanisms, studies on LSC might benefit from the use of strategy reports after each trial. It is well understood that individuals use different strategies when performing arithmetic problems (e.g., LeFevre, Sadesky, & Bisanz, 1996; Campbell & Xue, 2001). Overlearned problems are commonly retrieved from memory as facts, while new or large problems are indicated to be solved by the use of procedural strategies. Different strategies have also been found to be accompanied by specific neural correlates in fMRI as well as EEG (e.g., Dehaene, Piazza, Pinel & Cohen, 2003; Jost, Beinho, Hennighausen & Rosler, 2004; Núñez-Peña, Cortinas & Escera, 2006; De Smedt, Grabner & Studer, 2009; Grabner & De Smedt, 2012). Thus, future research should employ strategy reports to further study the cognitive mechanisms underlying LSC. Such reports could also indicate what length of training is sufficient for problems to be rote-learned.

According to our third hypothesis, we expected more LSC for the German training group than for the English training groups as participants of the former group have to switch from their dominant language (i.e., German, L1) to their non-dominant language (i.e., English, L2). Results revealed that LSC did not depend on whether the training was carried out in the dominant language or the non-dominant language. This finding contrasts with one of the previous studies, showing more LSC when switching from the dominant to the non-dominant than vice versa (Saalbach et al., 2013), but is in line with the study by Grabner and colleagues (2012) on Italian-German bilinguals. Contradicting results concerning the directional effect may be attributed to the specific language combination used in the previous studies. So far, training studies on LSC used different language combinations (i.e., German and Italian; German and French; German and English). Importantly, the order of the Arabic digit notation for two-digit number words differs cross-linguistically (Campbell and Xue, 2001). German uses a unit-ten order (e.g., “24” = four-and-twenty), whereas Italian, French, and English (for numbers higher than twenty) use a ten-unit order (e.g., “24” = twenty-four). This difference in word structure has been shown to influence arithmetic performance (e.g., Ellis & Hennelly, 1980; Göbel, Moeller, Pixner, Kaufmann, & Nuerk, 2014; van Rinsveld et al., 2015). French, however, adds a second interference by making use of a base-20 structure for numbers between 70 and 99, while the other languages have a clear base-10 structure. This additional interference might have led to a directional effect of LSC in Saalbach et al. (2013). Finally, in an exploratory analysis, no relationship between L2-vocabulary knowledge scores and LSC was found. The validity of these findings is limited in two ways: First, the LexTALE represent only an indication of language-proficiency and is not equal to the concept of language proficiency. Second, the present sample represents a rather homogeneous group with regard to L2 vocabulary knowledge. Thus, future research on LSC needs to assess language proficiency in a more comprehensive way within a group of bilingual speakers being also heterogeneous with respect to their L2 proficiency.

The present study provides both theoretical and practical implications. With regard to the former, our findings give further insights into the interplay of language and arithmetic knowledge acquisition. So far, different arithmetic operations were considered to rely differently on language-based processing. For example, previous research suggests that multiplication problems rely more strongly on a verbal coding than subtraction problems (e.g., Dehaene et al., 2004; Ischebeck et al., 2006; see introduction). The present study, however, does not reveal differences between these two operations with regard to LSC using auditory stimuli (see also Saalbach et al., 2013, using visual stimuli). Thus, we find no indication that auditorily presented multiplication problems rely more strongly on verbal coding than subtraction problems. Furthermore, finding no difference between LSC in the two arithmetic operations and LSC in the artificial task, requiring pure fact retrieval, suggests that arithmetic problems are stored as factual knowledge after an extended time of rote-learning. However, this assumption requires further and more direct examination, for example, by means of strategy reports or specific neuroscientific approaches.

Findings of the present study also provide implications for CLIL settings. A lot of content learned in school represents factual knowledge (e.g., rote-learning the multiplication table, remembering capital cities, historical dates, etc.). Given our findings that rote-learned information is applied more efficient in the language of instruction, LSC may also occur in school settings when language of application differs from language of instruction. This effect may be particular relevant for learners performing a task in limited time, such as classroom exams or other assessments. However, we need to be cautious in drawing inferences from laboratory studies to real-life classrooms. Teaching at school does not normally contain such massive rote learning as in the paradigm of the present study. Furthermore, the content used in this study and in most previous studies on language switching costs are limited to the effects on factual knowledge, representing only a part of what is learned in school. Thus, future research needs to examine the effects of language switching across learning and testing on the acquisition of conceptual as well as procedural knowledge within more complex kinds of task. In other words, highly controlled experimental studies on LSC should be complemented by research in more authentic settings. One way would be the scientifically based evaluation of implemented CLIL programs. Although a large evaluation of a specific CLIL program is being carried out (the *Europe School Berlin* program; Möller, Hohenstein, Fleckenstein, Köller, & Baumert, 2017), they do not include an examination of possible LSC yet.

To conclude, the present study revealed that cognitive costs arise when the language of instruction is different from the language of knowledge retrieval and application in the domain of arithmetic. This finding adds new evidence that language affects the way knowledge is stored in memory. To widen the extent to which these assumptions can be generalized, future research on cognitive costs through switching languages across instruction and retrieval needs to target other kinds of knowledge and more complex task settings. Then, it may be possible and justified to draw important implications for the design of effective CLIL programs.

Study 2

Language-dependent knowledge acquisition: mechanisms underlying language-switching costs in fact learning.

Introduction

Study 1 as well as previous research on LSC in bilingual learning settings mainly focused on the appearance of LSC but not on the underlying mechanisms. Understanding the mechanisms behind LSC is not only of interest to understand fundamental mechanisms in cognition 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 regarding the underlying mechanisms of LSC. On the one hand, LSC may emerge due to the translation of the acquired knowledge from the language of instruction to 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 RT and ACC are not enlightening with regard to the underlying cognitive mechanisms. One approach to gain further insights into them is to use neurophysiological data. This was done in two functional magnetic resonance imaging (fMRI) studies that have been outlined within the general introduction (see page 15 et seq. for a detailed description). In summary, both fMRI studies found increased activation in the language-switching condition, but were inconsistent regarding the involved brain network. Therefore, it remains in question what additional processes lie behind LSC. Furthermore, both studies applied visual stimuli in the form of written number words, which do not represent an ecologically valid learning material. Thus, Study 1 has already been an important groundwork showing that LSC appear within an auditory learning context.

An alternative way to examine the underlying mechanisms of LSC – compared to neurophysiological measures – are self-reports. Self-reports have a long tradition in research on arithmetic and are typically used to validate the problem-solving strategy that is applied to solve an arithmetic problem (e.g., LeFevre et al., 1996; Campbell and Xue, 2001; Imbo & Vandierendonck, 2007; Grabner & De Smedt, 2011; Vanbinst, Ghesquiere, & De Smedt, 2012, cf. Kirk & Ashcraft, 2001; Smith-Chant & LeFevre, 2003). 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 12 = 72; Siegler, 1996). Retrieval strategies are common in single-digit multiplications, which were rote-learned in school (e.g., Imbo & Vandierendonck, 2007; Grabner & De Smedt, 2011), and after a training on arithmetic facts (e.g., Grabner & De Smedt, 2012). Procedural strategies, in contrast, are used whenever the solution cannot be retrieved because of the problem size (e.g., in two-digit multiplications) or the operation (e.g., subtraction facts are typically not stored in a fact network). 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 or not translation processes were involved in problem-solving. Such self-reports have not been used before but directly address the question of whether LSC are due to translation processes. 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.

First and prior to the training sessions, we assessed individual characteristics of each participant regarding vocabulary knowledge of L2, arithmetic fluency, and general intelligence. In Study 1, there was no relation between the scores for vocabulary knowledge and LSC. We discussed that the test itself might not be a valid instrument to determine L2 proficiency precisely. However, we will stick to this measure throughout all three studies, since only then results of the project are comparable with each other. Further, within this project, a detailed test on L2 proficiency was not actionable. Even so, we added a second indicator for L2 proficiency, to corroborate the score of the LexTALE. Both tests have equal intention with an equivalent procedure, but different web-design and word compilation. Second, we assessed arithmetic fluency, an indicator for performance speed in different arithmetic operations. Since one of the possible mechanisms underlying LSC are additional numerical processes, it is reasonable to examine a possible relationship between arithmetic fluency and LSC. Finally, a measure for general intelligence was added, including measures for numerical intelligence and memory capacity.

Aims and Hypotheses of Study 2

The aim of Study 2 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 (ART), multiplication (MUL), and subtraction (SUB) problems using auditory stimuli similar to Study 1. As a key feature of Study 2, participants were required to provide two kinds of trial-by-trial self-reports, one on the problem-solving strategy and one on the use of translation processes. Finally, we assessed participants' vocabulary knowledge in L2, general intelligence, as well as arithmetic fluency.

Based on previous research on LSC, especially Study 1, we hypothesized finding longer RT for fact learning when the language of training differs from the language of testing, independently of the arithmetic task and the language of training (Study 1), while no LSC were expected for ACC (Hypothesis 1). According to the view that LSC are caused by additional numerical processing (Grabner et al., 2012), we expected (a) a significantly higher frequency of procedural strategy use in the switching condition compared to the no-switching condition (Hypothesis 2a). Moreover and according to the view that LSC are caused by additional translational processes (Venkatraman et al., 2006), we expected a higher frequency of self-reported translated trials for problems in the switching condition compared the no-switching condition (Hypothesis 2b). Moreover, we explored the possible relationship between trial-by-trial self-reports and LSC. Our expectations depended on the outcomes of hypotheses 2a) and 2b). In case of a higher frequency of procedural trials in the switching condition compared to the no-switching condition, we expected that this frequency of procedural trials predicted LSC. This expectation was based on the assumption that solving an equation with a procedure takes more time than retrieving an answer from memory. On the other hand, if additional translation processes predominantly characterize correctly solved trials in the switching condition, we expected that the frequency of trials solved with the help of translation processes predicted LSC. Again, this expectation was based on the assumption that additional translation will take longer and may lead to more errors than no translation processes. Finally, we explored the relationship between individual characteristics and LSC. Once more, depending on the outcomes of hypothesis 2a) and 2b), we expected a predictive effect of arithmetic fluency on LSC in case of more procedural strategies in the switching condition. On the other hand, we expected a relationship between vocabulary knowledge and LSC in case that participants use more translational processes within the switching condition.

Methods

Participants

The study included 47 right-handed students at the University of Göttingen. 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 artefacts throughout the test session[2]. The final sample consisted of 36 participants, aged between 20 and 28 (M=22.97, SD = 2.10). Participants were randomly assigned to either a German or English training group. All participants studied English Linguistics, had German as mother-language, and received their previous math education in a German school. They gave written informed consent and were paid for their participation.

Material

In line with Study 1, Study 2 included three different arithmetic tasks as experimental stimuli, comprising a) six ART (exact same as in Study 1); b) six MUL problems (exact same as in Study 1); and c) six SUB problems. SUB problems were two-digit and one-digit numbers with two-digit solutions (00 - 0 = 00), in contrast to two-digit and one-digit numbers with two-digit solutions (00 - 0 = 00) in Study 1. This change was reasonable due to the circumstance, that SUB problems in Study 1 had significantly longer RT than ART and MUL problems and therefore seemed to be more difficult than previously expected (see Results of Training Data, Study 1). Since we focused on NFK, we decided to use an easier version of SUB problems to ease the process of rote-learning. Furthermore, in contrast to Study 1, where the length of problems had been slightly different (i.e., problem length within 250ms differences), each problem in Study 2 had been modified to have the exact same length (i.e., 1850 milliseconds).

Assessment Instruments

Before the first training session, a battery of ability tests was administered in a separate screening session, including English vocabulary knowledge, arithmetic fluency and intelligence profile.

[2] EEG data within Training session 1 and the test session were collected for investigating additional research questions irrelevant to the present research project. These data will not be reported further.

Vocabulary knowledge

In order to assess L2 vocabulary knowledge, the same test was used as in Study 1 (for detailed description see page 30). Further, we added a second brief test for vocabulary knowledge, the Dialang (Huhta, Luoma, Oscarson, Sajavaara, Takala, & Teasdale, 2002). The Dialang placement test including 75 words that need to be markes as existing or non-existing in the English Language. In contrast to the LexTALE, answers can be corrected once marked, since 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 correlated with r = .80 (p < .001).

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. For each page, the time limit was two minutes. All subtests have two pages. The first subtest contains 60 three-term addition problems with multi-digit addends (e.g., 50+42+15= ...). The second subtest contains 60 multi-digit division problems per page (e.g., 56:8= ...). The third subtest contains six alternating rows of 10 multi-digit subtraction and multiplication problems per page (e.g., 42-17= ... , and 62x6= ...). The fourth subtest contains 60 multi-digit addition and subtraction problems with a suggested answer (e.g., 22+29=41) that have to be verified. The final score for arithmetic fluency is calculated as the total number of correctly solved problems.

General Intelligence

Participants' intelligence profile was 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 min. The raw scores of the individual tests are aggregated to an IQ score for general intelligence with 100 being the mean and 15 being the standard deviation.

Procedure

In line with Study 1, Study 2 consisted of four training days and one test session. All sessions took place at the Institute for Psychology at the University of Göttingen, Germany. Training session 1 and the test session took place 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 rote 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 activation was recorded by means of EEG, and the applied strategies were assessed by means of self-reports as described below. Figure 5 displays the schematic time course of the different sessions. The block order for training sessions was randomized. In consequence of several adjustments and a more complex procedure compared to Study 1, training and test procedures will be outlined in full detail. In contrast to Study 1, we used a mixed-design for the test session in line with previous studies on LSC (e.g., Grabner et al., 2012; Saalbach et al., 2013). As discussed in Study 1, the block-design used led to an unexpected order-effect (see Figure 2 on page 32 for a description of the design and page 39 for the discussion). With respect to our research goals, we chose to go back to the most commonly used test design in order to have a higher likelihood in finding LSC for our whole sample. Thus, each of the six test blocks contains all arithmetic problems as well as half the trails in the switching and half the trials in the no switching condition.

a) Training session

	Block 1	Block 2	Block 3
Training 1	ART	MUL	SUB
Training 2	MUL	SUB	ART
Training 3	ART	SUB	MUL
Training 4	SUB	ART	MUL

b) Test session

Block 1	ART, MUL and SUB in German and English
Block 2	
Block 3	
Block 4	
Block 5	
Block 6	

Figure 5. Schematic display of block order during a) Training sessions and b) Test session. Within the training, participants trained either in German or in English. During the test session, participants faced all problems in both languages.

Training Procedure for 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, or ART problems), 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 6, 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. A voice key collected RT. Timeout was set to 8.150 seconds after stimulus presentation (i.e., 10 seconds minus 1.850 seconds stimulus length). The examiner – seated outside the EEG cabin – typed in the given answer, after which the participants received a 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, 2001; 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 ART problems at the first training session, all ART problems were presented two times each with the solution. Thereafter, participants were required to solve these problems on their own as for MUL and SUB problems. The first session lasted between 30 and 40 minutes, depending on the individual speed of each participants.

Training Procedure for Sessions 2, 3 and 4[3]

Over the next three consecutive days, there were three additional training sessions to rote-learn the 18 problems. Each session had a duration between 25 and 35 minutes. Before each session, participants received again instructions on how to proceed during the session. As for the *Training session 1*, the three blocks were counterbalanced across the sample. The fixation point lasted for two seconds, and the problems were presented via headphones, so that about four participants could work on the task at the same time. Further, participants were instructed to press the ENTER-key in the moment they had the answer in mind. This was used as an alternative measure of RT in contrast to the voice-key in sessions 1 and 5. Afterwards, they were asked to enter the solution using a numerical keypad. Then, participants received a corrective feedback in visual form (correct or incorrect) followed by the correct solution being presented auditorily. Along the training sessions 2 to 4, 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 rote-learned the answer to each problem (Study 1; Grabner & De Smedt, 2012).

Test Procedure

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 had to go through six blocks, including both English and German problems. Within each block, the three operations were mixed as well as the two languages. 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. Moreover, as the posttest included a constant switching of language, 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. The duration of the test session was between 35 and 40 minutes.

[3] The following procedure is identical to the procedure of the training procedure in Study 1.

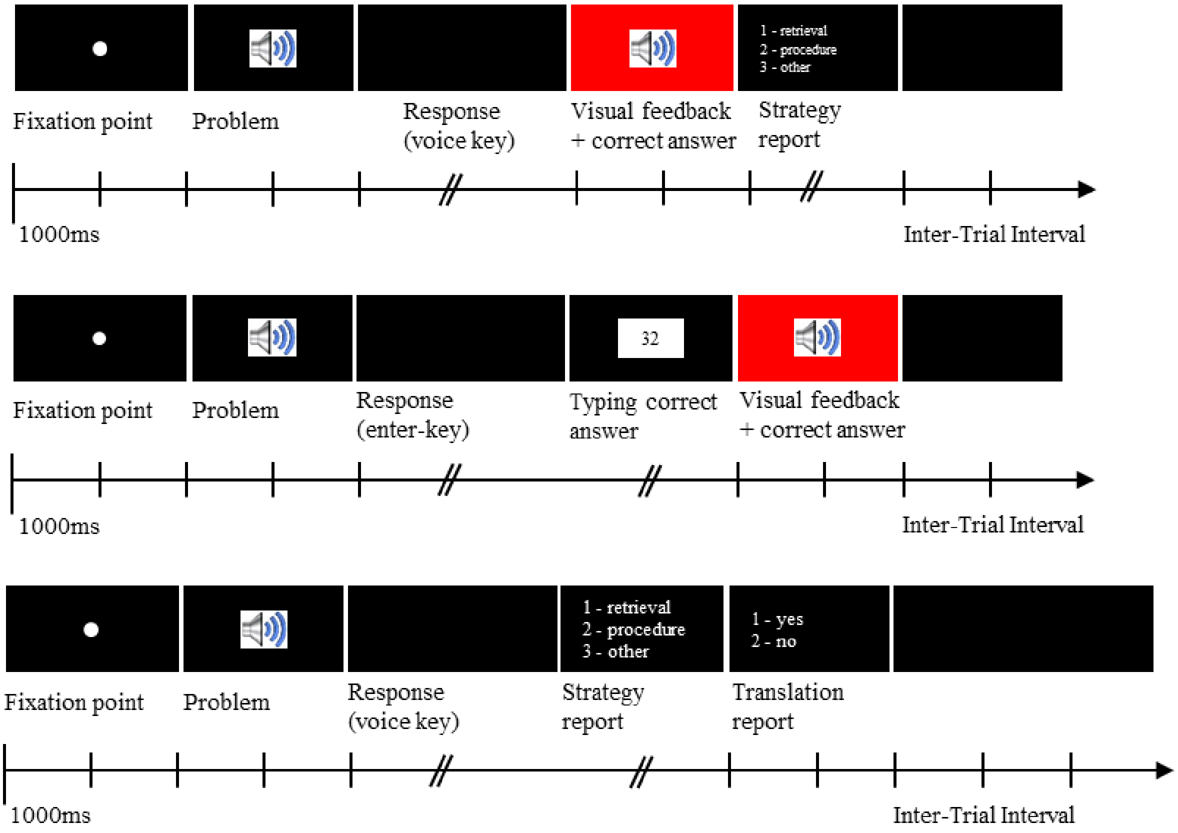


Figure 6. Schematic display of the trial time course separated for the training session 1, the training sessions 2 to 4, and the test session.

Data Analysis

IBM SPSS Statistics 20 was used for statistical analyses. ACC and RT for correctly solved trials were analyzed by means of mixed repeated-measure ANOVAs. Trials with voice-key errors in *Training session 1* and the *Test session* were excluded from analyses. For the training data, the ANOVA included the two within-subject factors Arithmetic Task (ART vs. MUL vs. SUB) and Training Day (day1 vs. day2 vs. day3 vs. day4), and the between-subjects factor Language of Training (German vs. English). The testing data ANOVA comprised the two within-subject factors Arithmetic Task (ART vs. MUL vs. SUB) and Language Switching (no switching vs. switching). 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 (i.e., .05 divided by the number of tests). For the analyses of strategy and translation reports, we conducted mixed repeated-measure ANOVAs, including the within-subject factors Arithmetic Task (ART vs. MUL vs. SUB) and Language Switching (no switching vs. switching), and the between-subject factors Language of Training (German vs. English). 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 ($\eta_p{}^2$). from analysis. In order to analyze the relationship of LSC with our assessment instruments, Pearson´s *r* was used.

Results

Table 3 summarizes the individual characteristics, 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 3. *Mean scores (standard errors) for the German and English training group (N=18 for each group) are displayed. Scores for Vocabulary Knowledge in percentage terms. Raw scores for Arithmetic Fluency. Standardizes IQ scores for General Intelligence.*

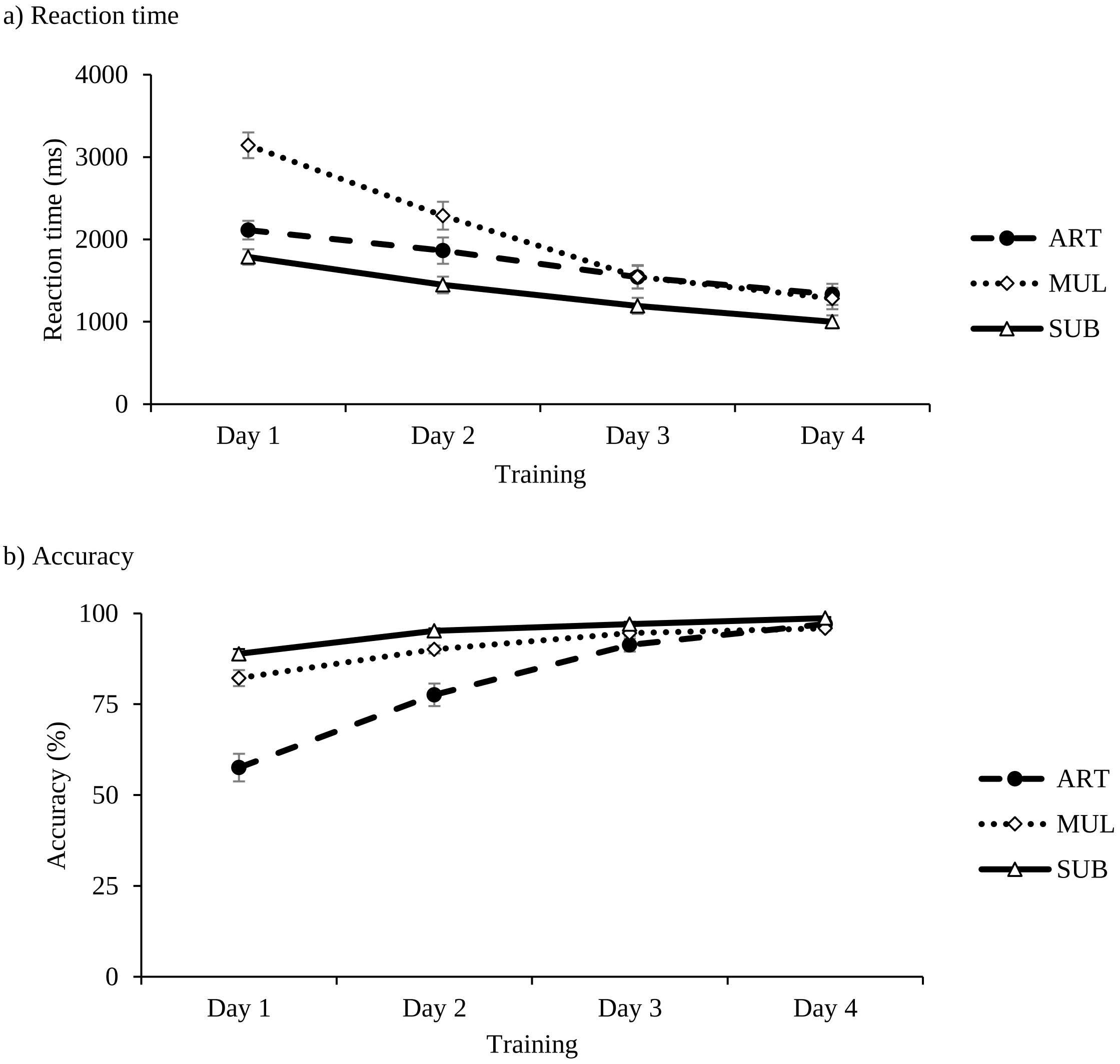
Measure	German Training (L1)	English Training (L2)	*p*
Vocabulary Knowledge L2	80.4 (3.1)	85.9 (2.2)	.16
General Intelligence	94.8 (2.0)	97.1 (1.6)	.37
Arithmetic Fluency	128.0 (9.5)	128.8 (5.5)	.94

**p < .05. **p < .01.*

Training Data

The Training data were analyzed for 18 participants trained in German (L1) and 18 participants trained in English (L2). Data for RT and ACC are displayed in Figure 7. In both measures, performance improved significantly. For RT, there was a strong main effect of Training Day ($F(2.13,72.31) = 95.86$, $p < .001$, $\eta_p{}^2 = .74$), with a significant decrease for each consecutive day (all *p*s < .001). In addition, there was a main effect of Arithmetic Task in RT ($F(2,68) = 31.47$, $p < .001$, $\eta_p{}^2 = .48$). ART were solved faster than MUL (1713 ms vs. 2065 ms; $t(35) = -4.20$, $p < .001$, $d = 0.48$), but more slowly than SUB (1713 ms vs. 1357 ms; $t(35) = 3.87$, $p < .001$, $d = 0.61$), and MUL more slowly than SUB (2065 ms vs. 1357 ms; $t(35) = 7.68$, $p < .001$, $d = 1.08$). Further, there was an interaction between Training Day and Arithmetic Task ($F(2.83, 96.12) = 22.10$, $p < .001$, $\eta_p{}^2 = .39$), which goes back to the specific improvement for ART from training 1 to training 2. All interaction effects were not significant (*p*s > .30). It can be concluded that the answers to all three arithmetic tasks were sufficiently rote-learned (i.e., all response times < 1333 ms).

For ACC, there was a main effect of Training Day (*F*(1.58,58.38) = 117.22, $p < .001$, $\eta_p{}^2 = .78$), with a significant increase for each consecutive day (all *p*s < .001). Further, there was a main effect of ARITHMETIC TASK (*F*(1.34,45.58) = 36.40, $p < .001$, $\eta_p{}^2 = .52$). ART were solved less accurately than MUL (80.1% vs. 90.1%; *t*(35) = -5.19, $p < .001$, $d = 1.03$) as well as SUB (80.1% vs. 95.0%; *t*(35) = -7.12, $p < .001$, $d = 1.56$), while MUL were solved less accurately than SUB (90.1% vs. 95.0; *t*(35) = -4.50, $p < .001$, $d = 0.96$). As for RT, there was an interaction between Training Day and Arithmetic Task (*F*(2.59, 88.10) = 31.77, $p < .001$, $\eta_p{}^2 = .48$), which goes back to the specific improvement for ART from training 1 to training 2. All interaction effects were not significant (*p*s > .61). At day four, problems of all three arithmetic tasks were sufficiently rote-learned (i.e., all ACC > 95%).


a) Reaction time

b) Accuracy

Figure 7. Training data for RT and ACC. Error bars indicate the standard error (SE).

Test data

Language Switching Costs

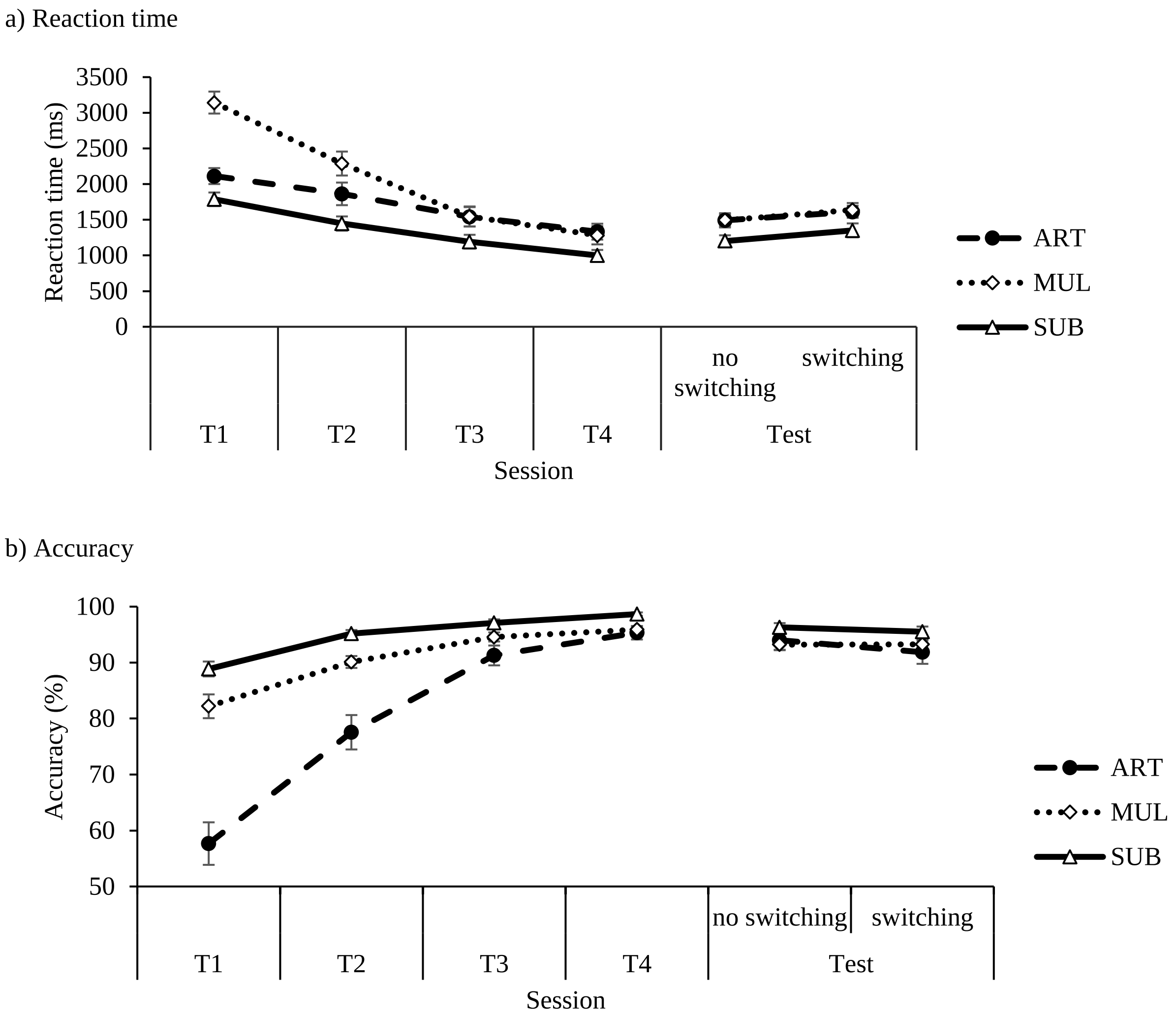
The ACC- and RT-results are shown in Table 4. A detailed overview of LSC for RT and ACC – separated by task – are displayed in the Supplementary Material (see page 105 et seq.).

Table 4. *Mean RT in milliseconds (upper rows) and ACC in percentage correct (lower rows) as a function of arithmetic task, and type of switching condition. Standard errors are enclosed in parentheses.*

	Artificial	Multiplication	Subtraction
RT in milliseconds			
No language switching	1492 (82)	1493 (100)	1203 (80)
Language switching	1613 (84)	1638 (97)	1352 (97)
ACC 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)

Hypothesis 1: We expected longer RT for fact learning when the language of training differs from the language of testing, independently of the arithmetic task and the language of training. No LSC were expected for ACC.

In line with hypothesis 1, there was a strong main effect for Language Switching on RT ($F(1,34) = 22.38$, $p < .001$, $\eta_p{}^2 = .40$), showing that problems in the no switching condition were solved faster (1396ms) than problems in the switching condition (1534ms). In line with hypothesis 1, there was neither an interaction between Arithmetic Task and Language Switching ($F(1.60,54.38) = 1.40$, $p = .254$, $\eta_p{}^2 = .04$), nor between Language Switching and Language of Training ($F(1,34) = .17$, $p = .685$, $\eta_p{}^2 = .01$). In addition, there was a significant main effect for the factor Arithmetic Task ($F(1.60,54.37) = 8.28$, $p = .001$, $\eta_p{}^2 = .20$). Post-hoc pairwise comparison revealed that SUB were solved more slowly than ART (-274.48, 95%-CI[-498.44,-40.52]) and MUL (-287.76, 95%-CI[-513.42,-62.10]). All other effects were not significant (all *p*s > .25). For ACC and in line with hypothesis 1, all effects were not significant (all *p*s > .07).



a) Reaction time

b) Accuracy

Figure 8. Training progress and test performance of a) RT and b) ACC. Error bars indicate the standard error (SE).

Strategy and translation reports

Figure 9 displays the distribution of (a) procedural strategy use and (b) translation 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.

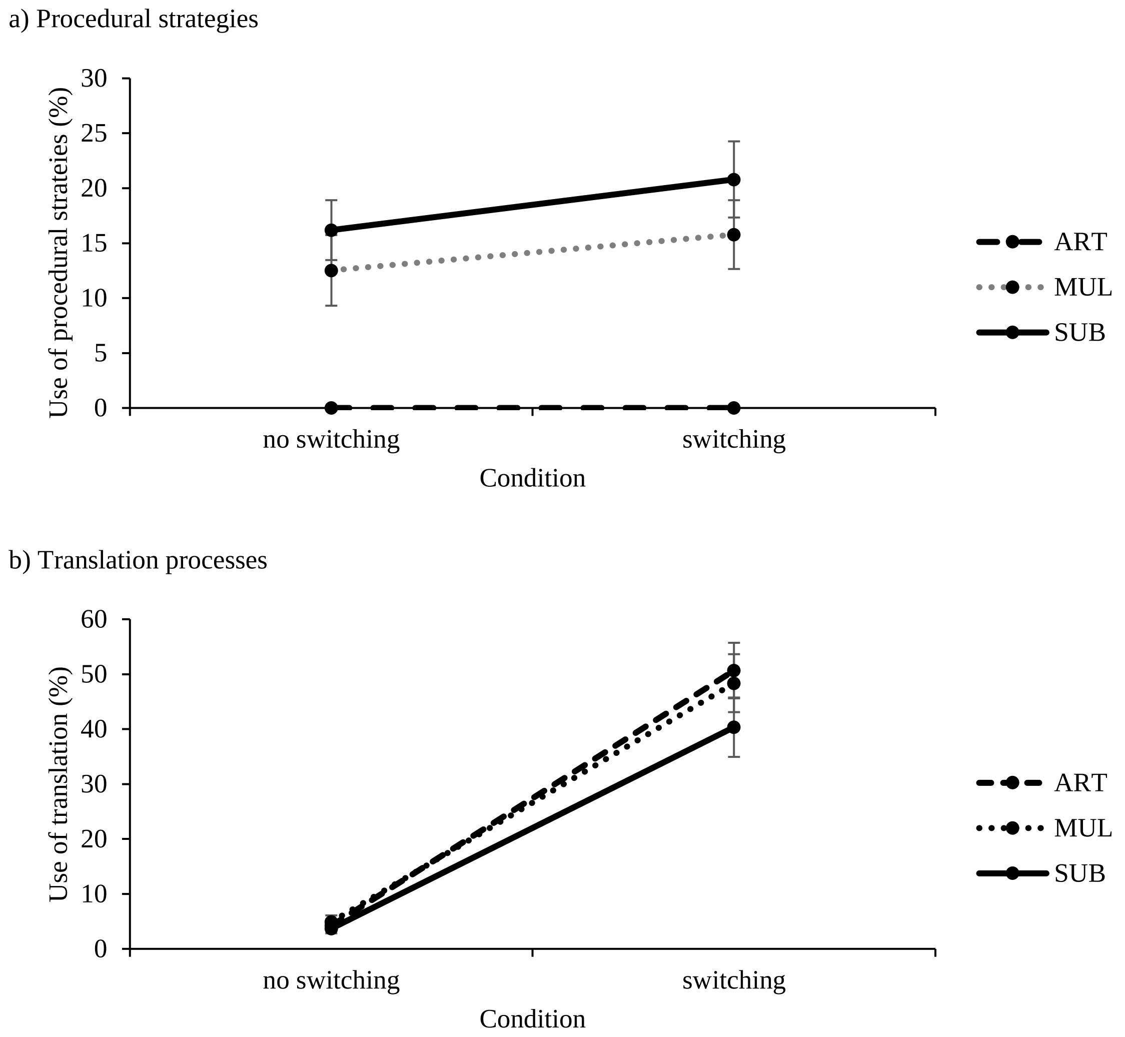


Figure 9. Distribution of self-reports during the test session for a) strategy reports and b) translational processes. Error bars indicate the standard error (SE).

Hypothesis 2a: According to the view that LSC are caused by additional numerical processing (Grabner et al., 2012), we expected a significant higher frequency of procedural strategy use in the switching condition compared to the no switching condition.

In line with hypothesis 2a, the repeated measures ANOVA on strategy reports revealed a main effect for Language Switching (F(1,35) = 12.12, p = .001, η_p^2 = .26), indicating that the frequency for procedural strategy use was higher in the switching condition (11.69%) compared to the no switching condition (9.36%). Further, there was a main effect of Arithmetic Task (*F*(2, 70) = 19.49, *p* < .001, η_p^2 = .36). Post-hoc pairwise comparison revealed a higher frequency of procedural strategy use for MUL and SUB compared to ART (MUL > ART: 13.68, 95%-

CI[6.37,21.00]; SUB > ART: 17.89, 95%-CI[10.55,25.23]). There was no interaction between the use of procedural strategies for the two training groups (i.e., German vs. English: *F*(1,34) = 3.00, *p* = .092, η_p^2 = .08). All other effects were not significant (all *p*s > .08).

As validation of the strategy reports, we conducted an additional RT analysis. This revealed that trials in which retrieval strategies were reported were solved significantly faster compared to procedural strategies (1391ms vs. 2236ms; *t*(25[4]) = -6.90, *p* < .001, *d* = 1.18).

Hypothesis 2b: According to the view that LSC are caused by additional translational processes (Venkatraman et al., 2006), we expected a higher frequency of translated trials for problems in the switching condition compared the no switching condition.

In line with the hypothesis 2b, the repeated measures ANOVA on translation reports showed a main effect of Language Switching (*F*(1,34) = 66.77, *p* < .001, η_p^2 = .66), indicating that the frequency of translation use was higher in the switching condition (46.46%) compared to the no switching condition (4.26%). Further, there was a main effect of Arithmetic Task (*F*(2,78) = 7.42, *p* = .001, η_p^2 = .18). Post-hoc pairwise comparison revealed that the frequency of translation use was higher for ART compared to SUB (5.41, 95%-CI[1.30,9.52]), as well as higher for MUL compared to SUB (4.61, 95%-CI[1.07,8.15]). Most importantly, there was an interaction of Arithmetic Task and Language Switching (*F*(2,68) = 5.57, *p* = .006, η_p^2 = .14). Post-hoc t-tests showed that the frequency of using translation during switching was lower for SUB (40.35%) compared to ART (50.66%; *t*(35) = -3.03, *p* = .005, *d* = .33) and MUL (48.36%; *t*(35) = -3.30, *p* = .002, *d* = .25). All other effects were not significant (all *p*s > .07).

As validation of the translation reports, we conducted an additional RT analysis. Overall, RT in trials without reported translation was significantly shorter (1560ms) than in trials with reported translation (2012ms ; *t*(33[5]) = -6.34, *p* < .001, *d* = .75).

Additional Analyses

Multiple regression analysis was used to test whether the trial-by-trial strategy and translation reports within the language-switching condition predicted LSC for RT. The

[4] Eleven participants were excluded from analysis reporting less than ten procedural trials
[5] Three particpants were excluded from analysis reporting less than ten trials including additional translation

regression model explained 19.4% of the variance in LSC ($R^2 = .19$, $F(2,33) = 3.97$, $p = .03$). The translation report (translation) turned out to be a significant predictor (*ß* = .44, $p = .008$), whereas the strategy report (procedures) was unrelated to LSC (*ß* = .03, $p = .87$). Hence, the more participants used translation processes in the switching condition, the higher the LSC were. On the other hand, despite the fact that participants used significantly more procedural strategies during the switching condition and procedural strategies had significantly longer RT, this factor did not predict LSC for RT. The same analyses was conducted for the data on LSC for ACC. The regression model showed no explanatory value for the prediction of LSC ($R^2 = .02$, $F(2,33) = .37$, $p = .69$).

Table 5 summarizes the relationship of LSC for RT and ACC for both training groups with our assessment instruments. None of the measurements showed a significant relationship with LSC. Neither within the German training group, nor the English training group.

Table 5. *Pearson correlation for individual characteristics with LSC for RT and ACC separated for the two training groups.*

	LSC for RT	LSC for ACC	LSC for RT	LSC for ACC
	German training group (n=18)		English training group (n=18)	
Vocabulary Knowledge L2	-.18	.35	-.26	.01
Arithmetic Fluency	.04	-.08	- .26	.13
General Intelligence	- .22	-.10	- .14	.02

**p < .05. **p < .01.*

Discussion

The aim of the present study was to provide further insights into the mechanisms underlying LSC in arithmetic fact learning. Therefore, thirty-six university students were trained on four consecutive days to learn eighteen problems of three different operation in either German (L1) or English (L2). On a fifth day, all participants were tested in both languages. LSC were found for RT but not for ACC. Further, participants used more procedural strategies as well as more translation processes in the language-switching condition compared to the no language-switching condition. Additional analyses revealed that only the participants´ use of translation during language switching significantly predicted LSC. No relationship was found between LSC and the individual characteristics measured (i.e., vocabulary knowledge for L2, arithmetic fluency, and general intelligence).

The first hypothesis was confirmed, finding longer RT for problems in the language-switching condition compared to the non-switching condition. Further, LSC were found for all three types of operations (i.e., multiplication, subtraction and pure fact learning), replicating the finding of Study 1 (see hypothesis 2, Study 1). No LSC were found for ACC. This is in line with our hypothesis, since it was previously shown that LSC do not appear for ACC when participants are given sufficient time to respond (see hypothesis 1, Study 1). As discussed in Study 1, a possible ceiling effect might be the reason (ACC >90%). Within the present study participants had an even more generous time frame to answer each trial (i.e., 13 seconds with an average RT < 2 seconds). Overall, the confirmation of hypothesis 1 replicates earlier findings of LSC in NFK in ecologically most valid design tested so far. It is the first study to find LSC combining auditory stimuli presentation and a voice-key for data collection. Previous research either collected data via visual stimuli and numeric keyboard via verification or production task (e.g., Grabner et al. 2012; Saalbach et al., 2013) or auditory stimuli and numeric keyboard (Study 1). In addition, it was the first study to find LSC for auditory stimuli in a test design including randomized switching of language and task (i.e., Study 1 used a block-wise language-switching design). Therefore, the study provides further evidence for the robustness of the appearance of LSC for NFK and amplifying ecological validity.

Regarding the main aim of the study, it was the first study to use self-reports to take a closer look at possible mechanisms behind LSC. In line with our expectations, participants not only used more procedural strategies within the language-switching condition (hypothesis 2a), but also indicated to use more additional translation processes (hypothesis 2b). Thus, both hypotheses were confirmed. The confirmation of hypothesis 2a implies that LSC might be explained by additional numerical processing as suggested by Grabner et al. (2012). However, it has to be mentioned that overall only about 12% of the trials in the language-switching condition had been identified as procedural strategies. Even though trials using procedural strategies took participants longer to solve a problem compared to retrieval strategies, it is unlikely that procedural strategies alone can account for the overall LSC found in our sample. Further, LSC were found for ART problems – which can only be retrieved from memory – to the same amount as for MUL and SUB. Finally, analyses did not mark procedural strategies as a predictor for LSC regarding RT. The confirmation of hypothesis 2b adds empirical evidence to the general assumption 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. As for procedural trials, translation trials also showed significantly longer

RT than its counterpart (i.e., no translation), therefore raising the average response time considerably. Further, analyses revealed the amount of translation trials as a predictor for overall LSC. Overall, the dominant change in solution strategy when confronted with the task in the language-switching condition can be pinned down to additional translation processes, but not to them alone.

It is critical to note that for ART trials, about 49% of the trials in the language-switching condition were indicated to not include additional translation, even though translation might be the only way to speak out the solution in the language asked for. There are at least two ways to explain this result. On the one hand, when considering that all problems were recurring six times, participants might have had a training effect in the switching condition during the test session. This means that at some point during the test session (e.g., after solving an arithmetic problem two or three times in the switching condition) participants knew the answer to a problem in the previously untrained language and did not need any additional processes. Therefore, an additional translation or procedural step had not been necessary anymore as was the case for the first or second the same problem had to be solved in the switching condition. In addition to this, participants in the English training group may have already been partly training the equation in their mother tongue from session 1 on. This is based on the assumption that when participants leave a training session or prepare for the next one, they think about the training items in their mother tongue, irrespective of the fact that the language of training is English. Therefore, a strong connection to only one language (the language of training respectively) might had never taken place for some participants or specific problems. A second consideration in order to explain the finding is connected to the fact that the test session included constant switching of language and three operations. In consequence, it is likely that some participants might have had a hard time reliably indicating for each trial what exactly had taken place. Nonetheless, the two self-reports show a clear tendency towards additional translation processes playing a key role in the appearance of LSC.

Regarding individual characteristics, the present study found no effect of L2 vocabulary knowledge (as an indicator for language proficiency), intelligence profile, and math fluency on later performance measures. Considering these findings in relation to the data received through the self-reports, the following explanations are possible. The fact that arithmetic fluency does not show a connection to LSC might be explained by the fact that additional numerical processes in the switching condition took place only in small proportion (i.e., < 20% for multiplication and subtraction problems). Most of the subjects almost exclusively used the retrieval strategy. It further adds evidence to the finding of our study that additional numerical processing does not represent a major player in explaining the underlying processes of LSC in NFK. Concerning vocabulary knowledge of L2, solving arithmetic problems only required limited language skills because problems and solutions consist of only one number words. Thus, the language ability in need was likely to be perfectly present for all participants. Even if this was not true, as mentioned in the discussion of Study 1 (see page 39 et seq.), the sample in Study 2 also exists of a rather homogeneous sample (i.e., all participants were following English as a study subject), opening a door for sample bias. Finally, the circumstance remains for our project that the tests used do not represent a direct test of language proficiency. For the purpose of Study 2, it was important to replicate findings of Study 1, which we did. Essentially, a recent publication by our research group (Volmer, Grabner, & Saalbach, 2018) revealed a negative correlation between LSC and L2 vocabulary knowledge. There were no noteworthy differences within the study design with respect to the auditory stimuli used (i.e. they were designed by the same person and program) and the training and test design. The fact that the same tests for vocabulary knowledge were used and a relation to L2 vocabulary knowledge was found for NFK as well as NFK embedded in text problems refute the argument that the tests do not fit the purpose. The only remarkable difference between the studies where the difference in vocabulary knowledge scores. Whereas our sample had an average score of about 83%, the sample of Volmer et al. only had an average score of about 62%. Thus, it might be the case that our sample was a more balanced sample with regard to language proficiency (i.e., miniscule difference between language proficiencies of L1 and L2), in contrast to a rather unbalanced sample in the divergent study. Nevertheless, the argument for sample bias remains speculative until tested empirically, contrasting two groups with significant different levels of language proficiency.

To conclude, the present study found LSC for multiplication, subtraction and a pure fact learning task using auditory stimuli and an oral response task. Further, by adding self-reports (i.e., strategy and translation reports), we were able to shed new light on the question of why LSC in NFK appear. The evidence suggests that additional translation processes play a key role in the origination of LSC in NFK. Self-reports therefore indicate that rote learned information is at least partially tied to the language of acquisition.

An elaborated discussion of theoretical and practical implications of Study 2 is integrated into the general discussion of the dissertation (see page 86 et seq.).

Study 3

Language-dependent knowledge acquisition: effects of language switching on procedural knowledge.

Introduction

Results of Study 1 and 2 corroborate the view that LSC with respect to RT are robust for NFK within the field of arithmetic using different methodological setups. Further, the implementation of self-reports revealed that the mechanisms underlying LSC are comparable between pure fact learning and arithmetic fact learning and consist to a major part of additional translational processes.

Aims and Hypotheses of Study 3

Study 3 aimed at broadening the picture on LSC by investigating whether LSC appear when participants have to learn a new arithmetic procedure in English (L2) and are then tested in German (L1) and English. We decided on focusing solely on an English training group, since Study 1 and 2 did not find any effect regarding language direction and LSC (i.e., switching from L1 to L2 vs. vice versa), neither on RT, ACC nor the distribution of self-reports. Thus, we concentrated only on an English training group, not only to increase the sample size, but also to represents at best the circumstances in the German context of CLIL. In order to entail comparable elements within Study 3 to Study 1 and 2 we also included the exact same artificial problems (ART) to assess LSC for NFK. Therefore, providing a second chance for replication of LSC in pure fact learning. The innovative feature was then to include a novel task, introducing addition in the base-7 number system (i.e., septimal system). Within this task, the numbers 7, 8 and 9 do not exist[6]. Hence, participants needed to inhibit common routines when performing addition tasks, therefore, learning a new routine. This task was used in studies before, providing evidence that it can be learned within a short time frame (Spelke & Tsivkin, 2001; Venkatraman et al., 2006; Nussbaumer, Grabner, Schneider & Stern, 2013). A set of base-7 addition was also rote-learned over the training days to become NFK (OLDADD). Therefore, we were able to compare the pure fact learning task with an arithmetic fact learning

[6] For a detailed description of the instruction, see Supplementary Material (page 109).

task. Previous research already found LSC for RT for exact base-6 and base-8 problems using written number words as visual stimuli and a verification task to assess RT and ACC (Spelke & Tsivkin, 2001). Further, a different set of base-7 additions (NEWADD) was included in each training session anew to automatize the procedure, with no ability to rely on fact retrieval. Therefore, we used the base-7 tasks as an indicator for LSC for NFK as well as procedural knowledge. Moreover, within the test-session, subtraction problems in base-7 were introduced to examine possible transfer effects of the procedural training in addition to the execution of subtraction problems in base-7 (NEWSUB). Overall, it will be the first study to question whether LSC appear when learning a novel arithmetic procedure. As in Study 2, identical self-reports were used to collect data for strategy use and additional translational processes, looking for a replication of our recent findings. Finally, individual cognitive markers were assessed (i.e., vocabulary knowledge (L2), general intelligence, arithmetic fluency, inhibitory control, and working memory) during an introductory meeting of the study. The relationship between individual markers and LSC is in its preliminary stage. So far, findings remain inconclusive (e.g., *Discussion* of Study 2). Data on inhibitory control and working memory were collected in comparison to Study 2. The reason behind this decision was that base-7 problems require the inhibition of the former automatizes number system. Recently, Volmer et al. (2018) found a correlation between the ACC of the same task we were using for inhibitory control (i.e., Simon Task) and LSC for NFK (i.e., multiplication problems), whereas no effect was found for the working memory measure. It has to be noted that Volmer and colleagues (2018) found LSC exclusively for the group that switched from the native language to the foreign language. Since we do not entail a group training in their native language in our study, we only aim at exploring the relationship for our present sample, with no clear expectations. Further, the new procedure may stress working memory to a different extent, compared to Volmer and colleagues (2018), since participants have to decide on a trial by trial basis and only for base-7 problems whether to add an additional number or not.

First, we hypothesized that LSC appear for both NFK tasks (i.e., ART, OLDADD) when the language of training differs from the language of testing. No LSC were expected for ACC, based on Study 1 and Study 2 (Hypothesis 1). Moreover, we explored whether LSC appear for procedural knowledge (i.e., NEWADD), and possible transfer effect of LSC to another arithmetic operation (i.e., subtraction problems; NEWSUB). Second, based on our findings of Study 2, we expected a higher percentage of procedural strategies within the switching condition for both NFK tasks (i.e., ART, OLDADD; Hypothesis 2a) as well as a higher percentage of translational processes within the switching condition for both fact learning tasks (i.e., ART, OLDADD; Hypothesis 2b). Moreover, we explored the distribution of self-reports for NEWADD and NEWSUB. Third, based on findings in Study 2, we expected that the individual frequency of additional translational processes (collected via self-report) predicts the size of LSC (Hypothesis 3). Finally, we explored the relationship of individual characteristics and LSC.

Methods

Participants

The study included 40 right-handed psychology students at the University of Göttingen, Germany. Five participants were excluded, missing one training session. The final sample consisted of 35 participants, aged between 18 and 28 years (M = 21.60, SD = 2.16). All participants had German as mother-language, and received their previous math education in a German school. They gave written informed consent and were paid for their participation in form of subject hours that were mandatory for their study subject.

Material

Three blocks of arithmetic problems were trained within each training session. The following order of description represents the order of presentation within each training block. First, six artificial arithmetic problems (ART): two-digit and one-digit numbers connected via an arbitrary symbol (box) with two-digit solutions (17 box 2 = 93). The exact same problems have been used in Study 1 and Study 2, where it was shown that these problems can be rote-learned in comparable time and fashion as typical arithmetic problems. Second, base-7 addition problems (OLDADD): two-digit + two-digit problems with two-digit solutions (00 + 00 = 00). Third, base-7 addition problems (NEWADD): two-digit + two-digit problems with two-digit solutions (00 + 00 = 00) as well as two-digit + one-digit problems with two-digit solutions (00 + 0 = 00). ART and OLDADD problems stayed the same for each training day to promote rote-learning, with six repetitions for six different problems, while the third training block NEWADD contained 36 new problems each day to promote the learning process of the new arithmetic operation. For the test session, 36 subtraction problems (SUB) – to be calculated also in the base-7 system – were included to test for transfer effects: two-digit minus two-digit problems with two-digit solutions (00 – 00 = 00) and two-digit minus one-digit problems with two-digit solutions (00 – 0 = 00). Auditory stimuli were created with the professional audio software Voice Reader Studio 15, a widely used text-to-speech program (Linguatec, 2015). The training and test program were created using E-Prime 2.0 Professional stimulus presentation software (Schneider, Eschmann, & Zuccolotto, 2002). As was true for Study 2, each problem was modified to have the same length (i.e., 1850 milliseconds).

Assessment Instruments

During an instruction meeting in which also the appointments for training and test session were made, a battery of ability tests was assessed.

Vocabulary Knowledge L2

In order to assess the vocabulary knowledge of L2, the same instruments were used as in Study 2 (for detailed description see page 47)

Arithmetic fluency

In order to assess arithmetic fluency, the same instruments were used as in Study 2 (for detailed description see page 47)

Intelligence Profile

In order to assess the general intelligence profile, the same instruments were used as in Study 2 (for detailed description see page 47)

Inhibitory Control

The Simon Task by Simon and Rudell (1967) was used as a measure of inhibitory control. In general, the Simon effect is known as the difference in ACC or RT between trials in which stimulus and response are displayed on the same side of the screen (congruent trials), compared to trials in which they are on the opposite sides (incongruent trials). Participants had to press the “a” key (QWERT-keyboard with the “a” positioned on the far left side) when seeing a red square on the screen and to press the “ä” key (positioned at the far right side of the keyboard) when seeing a green square on the screen. The squares appeared either on the right or on the left side of the screen. Therefore, the position of the square may be congruent with the positioning of the key (e.g., pressing “s” key when the red square appears on the left side of the screen) or incongruent (e.g., pressing the “s” key when the red square appears on the right side of the screen). The training consisted of 20 items, followed by 200 test items, separated in five blocks of 40 items. Final score was the percentage of correct trials.

Working Memory

The 3-back task by Mackworth (1959) was used to account for a measure of working memory. Task requirements are the storage of information in an updating manner, therefore engaging working memory constantly (Conway, Miura, & Colflesh, 2007). Within this computer task, participants have to judge whether a number that appears on the screen matched the number presented three numbers before. Participants were instructed to press the SPACE-bar when the numbers did match. The test consisted of 24 training numbers, followed by a break of 30 seconds, and 240 test numbers. The final score represents the sum of the proportion of found matches and the proportion of correct rejections (i.e., to prevent participants from constantly pressing the SPACE-base).

Procedure

In line with Study 1 and Study 2, Study 3 consisted of four training sessions and one test session. All sessions took place at the Institute for Psychology at the University of Göttingen, Germany. Training sessions took place in the same computer lab as training and test session of Study 1 and training sessions 2, 3 and 4 of Study 2. The test session in the same EEG-lab than Study 2. Since Study 1 and 2 showed no difference between the German or English training groups, there was no German training group within Study 3. During the training sessions, all participants had to solve sets of problems in English (L2). In the posttest session, the problems were presented in English and German (L1; Figure 10 displays the schematic time course of training and testing.)

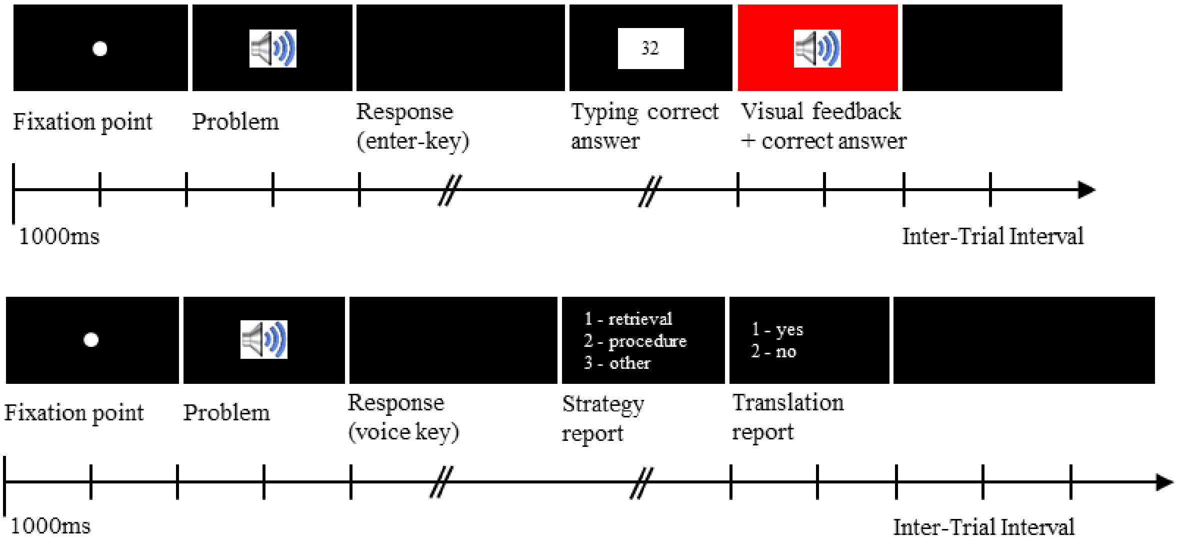


Figure 10. Schematic display of the trial time course separated for the training sessions and the test session.

Training Procedure

Figure 11 displays the sequence of the training and test session. In order to learn the new procedure (i.e., addition in base-7), participants were instructed via a written outline of the procedure (see Supplementary Material on page 109). After reading the description in English, participants sat down at the computer and started with the first training session. Before each training session, the procedure was explained anew to make sure that all participants followed the same instructions. The training procedure for single trials was identical to the training sessions 2, 3 and 4 as described in Study 2. Working time for all training session was between 25 and 35 minutes.

Test Procedure

The test procedure for single trials was identical to the procedure described in Study 2, except for the application of EEG in Study 2. In contrast to Study 2, the test session consisted of eight blocks. In blocks one to three and five to seven, the three training tasks were mixed as well as the two languages. Block four and eight consisted only of base-7 subtraction problems in both, also presented in German and English. After block four, a break of five minutes was included in order to opening the windows and having the chance to use the bathroom, located next door. Depending on the individual speed, the duration of the test session was between 45 and 60 Minutes.

a) Training sessions

	Block 1	Block 2	Block 3
Training 1	ART	OLDADD	NEWADD
Training 2	ART	OLDADD	NEWADD
Training 3	ART	OLDADD	NEWADD
Training 4	ART	OLDADD	NEWADD

b) Test session

Block 1	ART, OLDADD and NEWADD in German and English
Block 2	
Block 3	
Block 4	NEWSUB in German and English
5 minute break	
Block 5	ART, OLDADD and NEWADD in German and English
Block 6	
Block 7	
Block 8	NEWSUB in German and English

Figure 11. Schematic display of block order during a) Training sessions and b) Test session. Within the training, problems were only presented in English (L2), whereas problems during the test session were presented in both languages within each block.

Data Analysis

IBM SPSS Statistics 20 was used to statistical analyses. To evaluate behavioral effects of language-switching, we analyzed ACC, RT and the information from the two self-reports (strategy report and translation report; frequency of selected strategies). Training data were analyzed in an ANOVA, including the two within-subject factors Arithmetic Tasks (ART vs. OLDADD vs. NEWADD) and Training Day (day1 vs. day2 vs. day3 vs. day4). Testing data were also analyzed in an ANOVA, including the two within-subject factors Arithmetic Task (ART vs. OLDADD vs. NEWADD vs. NEWSUB) and Language Switching (no switching vs. switching). Additional post-hoc t-tests with alpha-correction were performed in cases of main effects including three or more means to provide additional specific information about the data. Trials with voice-key failures were excluded from analyses. RT data was only included for correct trials. For effect sizes, Cohen’s d and partial eta-squared were computed. In case of violation of the assumption of sphericity (Mauchly’s test), degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. In order to analyze the relationship of LSC with our assessment instruments, Pearson´s *r* was used.

Results

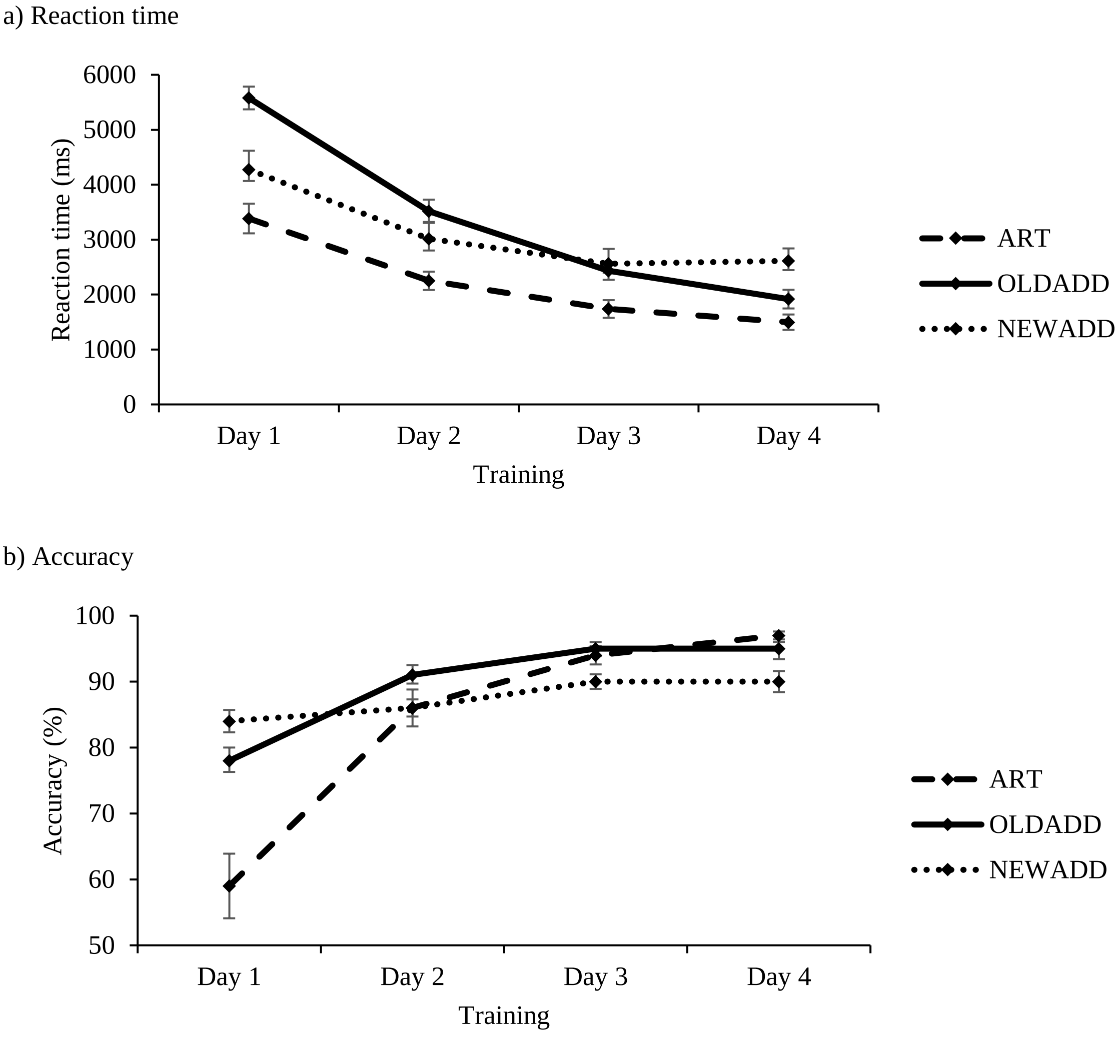
Table 6 summarizes the individual characteristics collected during the pre-meeting.

Table 6. *Mean scores (standard errors; N=35) for individual characteristics are displayed. Scores for Vocabulary Knowledge, Inhibitory Control, and Working Memory in percentage terms. Raw scores for Arithmetic Fluency. Standardizes IQ scores for General Intelligence.*

Measure	English Training Group
Vocabulary Knowledge L2	72.1 (2.1)
General Intelligence	92.7 (1.6)
Arithmetic Fluency	182.3 (6.6)
Inhibitory Control	93.3 (0.6)
Working Memory	68.3 (2.7)

Training Data

Training data for ACC and RT are displayed in Figure 12. For ACC, there was a main effect for Training Day (*F*(1.83, 62.17) = 88.82, *p* < .001, η_p^2 = .72). This effect was significant for the first three training days (all *p*s < .001) and non-significant comparing day three and four (*t*(34) = -1.23, *p* = .27, d = 0.42). Further, there was a main effect for Arithmetic Task (*F*(1.40, 47.56) = 7.39, *p* = .004, η_p^2 = .18). Post hoc analysis revealed that overall, ART problems were solved less accurate than OLDADD problems (84% vs. 89%, *t*(34) = -3.15, *p* = .003, *d* = 1.08), and NEWADD problems (84% vs. 87%, *t*(34) = -2.13, *p* = .04, *d* = 0.73), while OLDADD problems were solve more accurate than NEWADD problems (89% vs. 87%, *t*(34) = 2.50, *p* = .017, *d* = 0.86). After the last training session, participants had a higher ACC for ART than for OLDADD (97% vs 95%, *t*(34) = 2.19, *p* = .035, *d* = 0.75), as well as for NEWADD (97% vs. 90%, *t*(34) = 5.28, *p* < .001, *d* = 1.99), and a higher ACC for OLDADD than for NEWADD (95% vs. 90%, *t*(34) = 3.73, *p* = .001, *d* = 1.28). For RT, there were also the two main effects for Training Day and Arithmetic Task. For Training Day, post hoc analysis revealed a significant increase for each consecutive day (all *p*s < .001). For the main effect of Arithmetic Task, post hoc analysis displayed faster RT for ART than for OLDADD (2217 ms vs. 3360 ms, *t*(34) = -5.83, *p* < .001, *d* = 2.00), and NEWADD (2217 ms vs. 3116 ms, *t*(34) = -6.34, *p* < .001, *d* = 2.17). Further, equal RT was found for OLDADD compared to NEWADD (3360 ms vs. 3116 ms, *t*(34) = 1.46, *p* = .16, *d* =). Looking only at the last training day, answering ART was faster than OLDADD (1497 ms vs. 1917 ms, *t*(34) = -2.32, *p* = .026, *d* = 0.80) as well as NEWADD (1497 ms vs. 2615 ms, *t*(34) = -8.55, *p* < .001, *d* = 2.66), and OLDADD faster than NEWADD (1917 ms vs. 2615 ms, *t*(34) = -4.61, *p* < .001, *d* = 1.58).


a) Reaction time

b) Accuracy

Figure 12. Training data for a) RT and b) ACC. Error bars indicate the standard error (SE).

Test Data

Language Switching Costs

The RT- and ACC-results are shown in Table 7. A detailed overview of LSC for RT and ACC – separated by task – are displayed in the Supplementary Material (see page 105 et seq.).

Table 7. *Mean RT in milliseconds (upper rows) and ACC in percentage correct (lower rows) as a function of arithmetic task, and type of switching condition. Standard errors are enclosed in parentheses.*

	Artificial	Old Addition	New Addition	New Subtraction
RT in milliseconds				
No language switching	2015 (147)	2606 (210)	2655 (122)	3562 (148)
Language switching	2054 (134)	2543 (191)	2364 (122)	3240 (152)
ACC in percentage correct				
No language switching	85.3 (1.0)	89.7 (1.5)	81.7 (1.9)	72.1 (2.5)
Language switching	85.3 (0.9)	89.9 (1.1)	86.4 (1.2)	79.5 (2.3)

Hypothesis 1: We expected longer RT for both NFK tasks (i.e., ART, OLDADD) in the switching condition compared to the no switching condition. We did not expect any difference for ACC, based on findings of Study 1 and Study 2.

In relation to hypothesis 1, we explored whether LSC appear for procedural knowledge (i.e., NEWADD), and possible transfer effect of LSC to another arithmetic operation (i.e., subtraction problems; NEWSUB).

Results for RT and ACC are displayed in Figure 13. For RT, there was a main effect for Language Switching (*F*(1,34) = 6.74, *p* = .014, $\eta_p{}^2$ = .17), indicating that problems in the no switching condition had longer RT than problems in the switching condition (159.02, 95%-CI[34.54,283.50]). The effect was opposite than expected, indicating that overall participants were faster in solving problems in the switching condition than in the no switching condition. Further, there was an interaction between the factors Arithmetic Task and Language Switching (*F*(3,102) = 4.38, *p* = .006, $\eta_p{}^2$ = .11). Post-hoc t-tests showed shorter RT for NEWADD and NEWSUB in the Switching condition compared to the no switching condition (NEWADD: *t*(34) = 3.80, *p* = .001, *d* = 1.30); NEWSUB: *t*(34) = 3.20, *p* = .003, *d* = 1.10). No LSC were found for ART and OLDADD. Both tasks were solved equally fast in the no switching compared to the switching condition (ART: *t*(34) = -.62, *p* = .54, *d* = 0.21); OLDADD: *t*(34) = .66, *p* = .51, *d* = 0.23). Finally, there was a main effect of Arithmetic Task (*F*(2.49/84.68) = 37.04, *p* < .001 , $\eta_p{}^2$ = .52). Pairwise comparison with Bonferroni correction revealed that ART problems had shorter RT than OLDADD (-539.55, 95%-CI[-922.17,-156.93]), NEWADD (-474.98, 95%-CI[-797.09,-152.87]), and NEWSUB (-1366.79, 95%-CI[-1755.15,-978.44]).

Finally, NEWSUB had longer RT than OLDADD (827.24, 95%-CI[372.68,1281.81]) and NEWADD (891.81, 95%-CI[610.25,1173.38]).

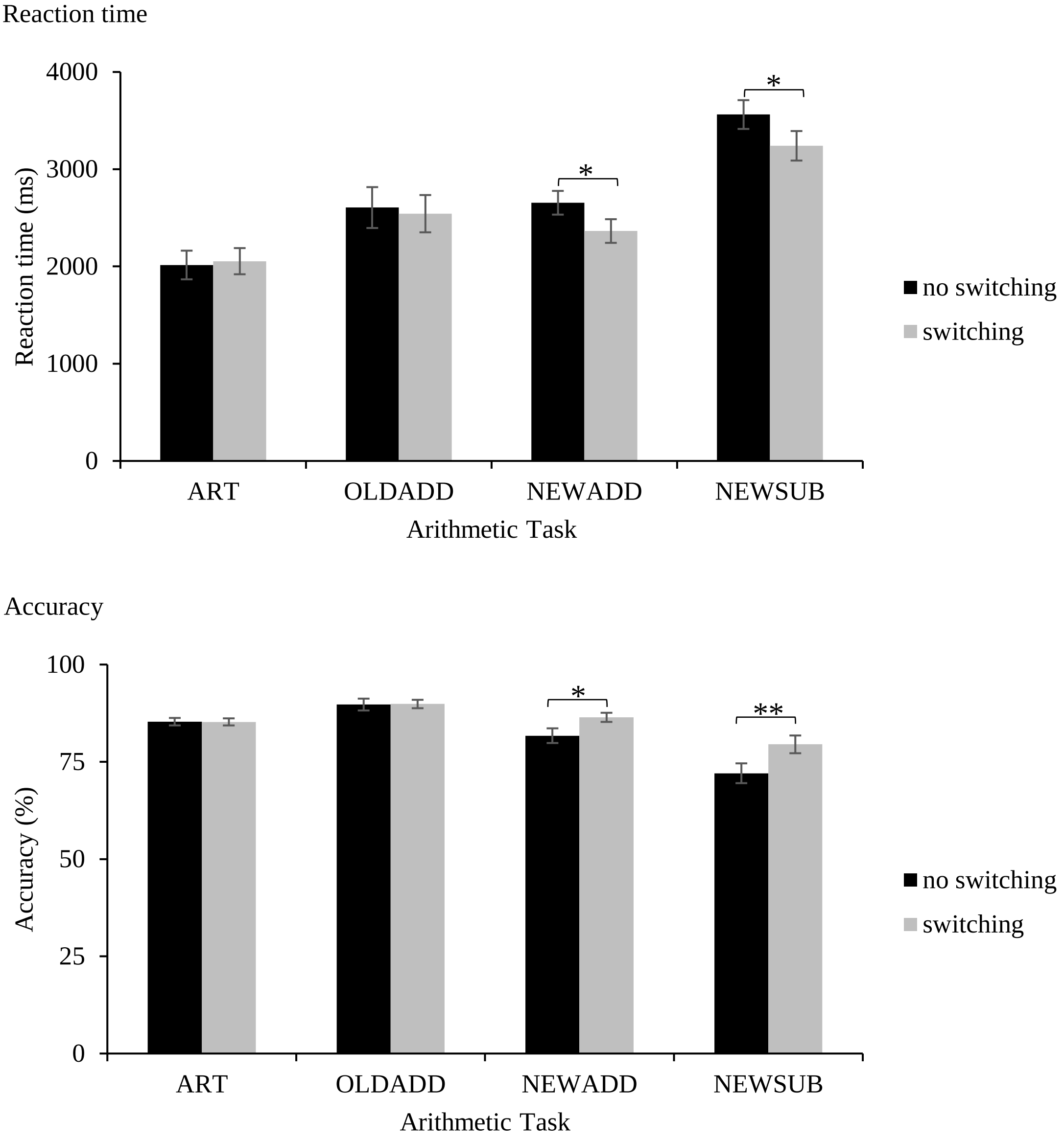


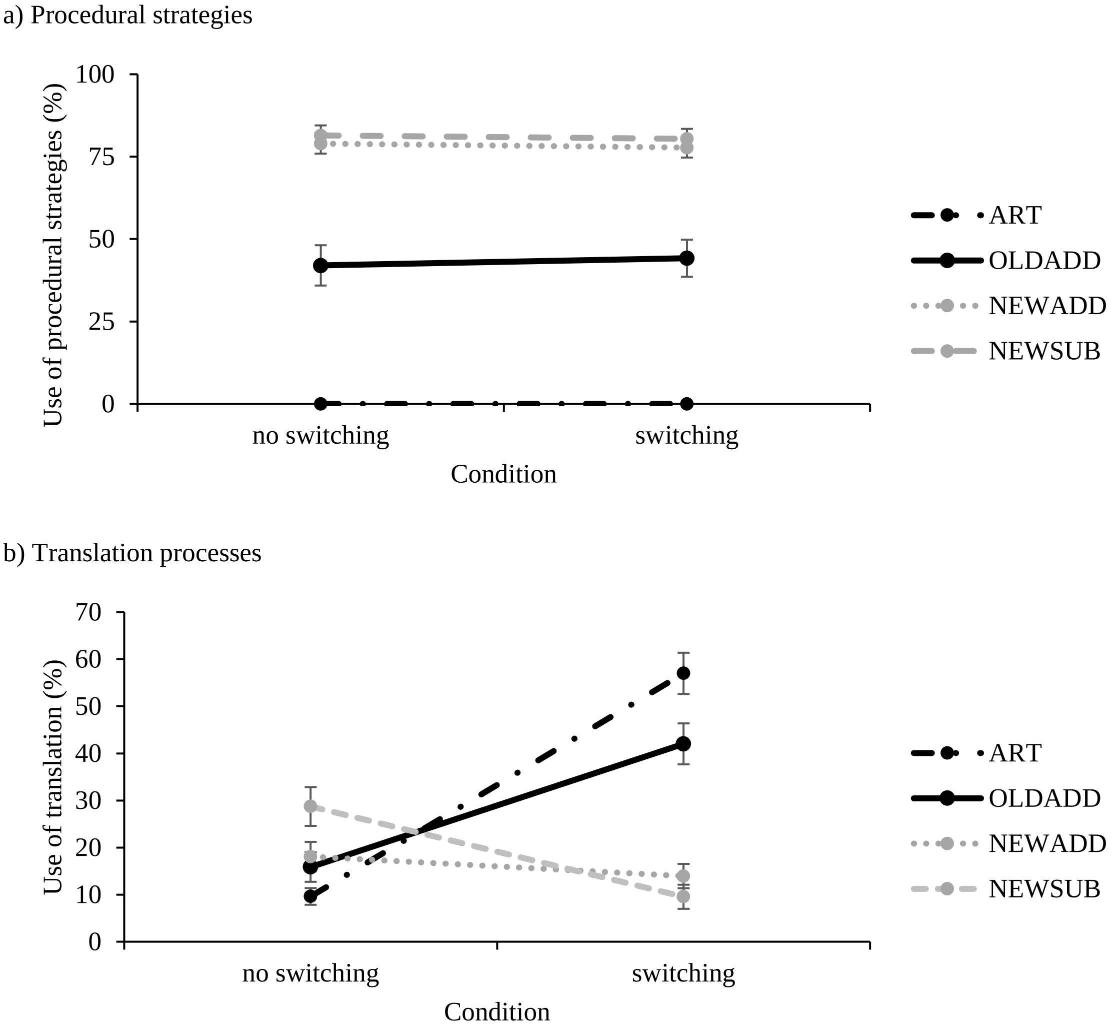
Figure 13. Test performance displayed for each tasks regarding a) RT and b) ACC. Error bars indicate the standard error (SE). *p < .05. **p < .01.

For ACC, results revealed a main effect for Language Switching ($F(1,34) = 20.50$, $p < .001$, $\eta_p^2 = .38$), indicating that overall problems in the no switching condition were solved less accurate than problems in the switching condition (-3.1, 95%-CI[-4.4,-1.7]). Further, there was an interaction between Arithmetic Task and Language Switching ($F(3,102) = 8.11$, $p < .001$, $\eta_p^2 = .19$). Post-hoc tests revealed that ACC for NEWADD and NEWSUB was higher in the switching condition compared to the no switching condition (NEWADD: $t(34) = -3.00$, $p = .005$, $d = 1.03$; NEWSUB: $t(34) = -5.45$, $p < .001$, $d = 1.87$). These results are opposite to

Hypothesis 1, since we expected no LSC for ACC. Finally, there was a main effect of arithmetic task ($F(1.81,61.69) = 30.37$, $p < .001$, $\eta_p^2 = .47$). Pairwise comparison with Bonferroni correction revealed that ART were solved less accurate than OLDADD (-4.5, 95%-CI[-7.3,-1.7]), and more accurate than NEWSUB (9.5, 95%-CI[3.9,15.1]). OLDADD were solved more accurate than NEWADD (5.7, 95%-CI[3.3,8.1]), and more accurate than NEWSUB (8.3, 95%-CI[3.2,13.4]).

Self-reports for strategy and translation

Figure 14 displays the distribution of (a) strategy reports and (b) translation use across all four tasks. Since the frequency of trials with the strategy “other” was low (< 1%), these trials were excluded from further analyses.


a) Procedural strategies

b) Translation processes

Figure 14. Distribution of self-reports during the test session for a) strategy reports and b) translational processes. Error bars indicate the standard error (SE).

Hypothesis 2a: We expected a higher frequency of procedural strategy use in the switching condition compared to the no switching condition for NFK (i.e., ART, OLDADD).

Moreover, we explored the distribution of strategy reports for NEWADD and NEWSUB.

Repeated measures ANOVA showed no main effect for Language Switching (F(1,34) = .00, p = .979, $\eta_p{}^2$ = .00) indicating that procedural strategies were used as often in the switching condition as in the no switching condition (0.18, 95%-CI[-1.35,1.39]). There was a main effect for TASK (F(1.54,52.22) = 138.44, p < .001, $\eta_p{}^2$ = .80), indicating that less procedural strategies were used for ART than for OLDADD (-43.11. 95%-CI[-59.36,-26.86]), NEWADD (-78.37, 95%-CI[-87.21,-69.52]), and NEWSUB (-80.93, 95%-CI[-89.24,-72.62]). Further, there were less procedural strategies for OLDADD than for NEWADD (-35.26, 95%-CI[-52.21,-18.31]), and SUB (-37.82, 95%-CI[-54.08,-21.57]). None of the other effects was significant (all *p*s > .05). As validation of the strategy reports, we conducted an additional RT analysis. This revealed that trials in which retrieval strategies were reported were solved significantly faster compared to procedural strategies (2269ms vs. 3401ms; t(34) = 7.91, p < .001, d = 1.14).

Hypothesis 2b: We expected a higher frequency of translation strategy use in the switching condition compared to the no switching condition for NFK (i.e., ART, OLDADD).

Moreover, we explored the distribution of translation reports for NEWADD and NEWSUB.

Repeated measures ANOVA showed a main effect for Language Switching (F(1,34) = 18.80, p < .001, $\eta_p{}^2$ = .36) indicating that more translation was reported in the Switching condition than in the no switching condition (12.56, 95%-CI[6.67,18.44]). Further, there was an interaction for Arithmetic Task and Language Switching (F(1.90,64.55) = 57.47, p < .001, $\eta_p{}^2$ = .63). Post-hoc analyses revealed a higher translation strategy use in the switching condition for ART (t(34) = 9.29, p < .001, d = 2.34), and OLDADD (t(34) = 4.86, p < .001, d = 1.17). Further, for NEWSUB there was less use for translational strategy within the switching condition ($t(34) = 4.82$, $p < .001$, $d = 0.95$). Finally, there was a main effect for Arithmetic Task ($F(2.07,70.39) = 17.13$, $p < .001$, $\eta_p^2 = .34$). Post-hoc analysis revealed that the frequency for use of translation was higher for ART compared to NEWADD (17.30, 95%-CI[8.10,26.50]) and NEWSUB (14.17, 95%-CI[4.25,24.10]), and higher for OLDADD than for NEWADD (12.93, 95%-CI[6.84,19.02]) and SUB (9.81, 95%-CI[2.67,16.95]). None of the other effects were significant (all *p*s > .05). As validation of the strategy reports, we conducted an additional RT analysis. This revealed that trials in which retrieval strategies were reported were solved significantly faster compared to procedural strategies (2294 ms vs. 3069 ms; $t(34) = 6.76$, $p < .001$, $d = 0.84$).

Hypothesis 3: We expected that the individual frequency of additional translational processes predicts the size of LSC for RT, whereas the frequency of additional numerical processes does not.

Multiple regression analysis was used to test whether the trial-by-trial strategy and translation reports within the language-switching condition (procedural strategies; additional translation processes) predicted LSC for RT. Despite the fact that no LSC were found, we conducted the analysis, since it is still possible that a relationship exists. The results of the regression indeed indicated that the model explained 15.5% of the variance. The effect was marginal significant in predicting LSC ($R^2 = .155$, $F(2,32) = 2.92$, $p = .068$). As predicted, the frequency of procedural strategy use did not predict LSC (*ß* = .30, $p = .098$), while the use of additional translation processes predicted LSC significantly (*ß* = .40, $p = .031$). Thus, the more often participants used translational strategies, the more likely they showed LSC for RT. The same analyses was conducted for the data on LSC for ACC. The regression model showed no explanatory value for the prediction of LSC ($R^2 = .02$, $F(2,32) = .29$, $p = .75$).

Additional Analyses

Table 8 summarizes the relationship of LSC for RT and ACC with all assessment instruments used. Based on previous findings in this study, we display correlations separately for all four tasks. Scatterplots are displayed for all significant findings in Figure 15. For RT, there was a positive correlation for arithmetic fluency and LSC for ART ($r = .38$, $p =.023$). For ACC, there was a strong correlation between the score in working memory and overall ACC (*r*

= .50, *p* = .002), indicating that the higher the working memory score, the more likely participants show overall LSC in ACC. This relationship was marginally significant for ART (*r* = .32, *p* = .061) and NEWUBS (*r* = .33, *p* = .051).

Table 8. *Pearson correlation for individual characteristics and LSC for RT and ACC.*

	LSC for RT					LSC for ACC				
	all	ART	OLD ADD	NEW ADD	NEW SUB	all	ART	OLD ADD	NEW ADD	NEW SUB
Vocabulary Knowledge L2	-.08	-.05	-.17	.24	-.14	.27	.16	.16	.09	.22
Arithmetic Fluency	.08	**.39***	.17	.01	-.17	.19	.27	-.01	.06	.22
General Intelligence	.12	.22	.19	.14	-.11	.32	.14	.20	.12	.30
Inhibitory Control	.08	-.07	.10	-.01	.13	-.01	-.05	-.32	.25	.07
Working Memory	-.18	-.04	.04	-.20	-.27	**.50****	.32	.21	.29	.33

**p* < .05. ***p* < .01.*

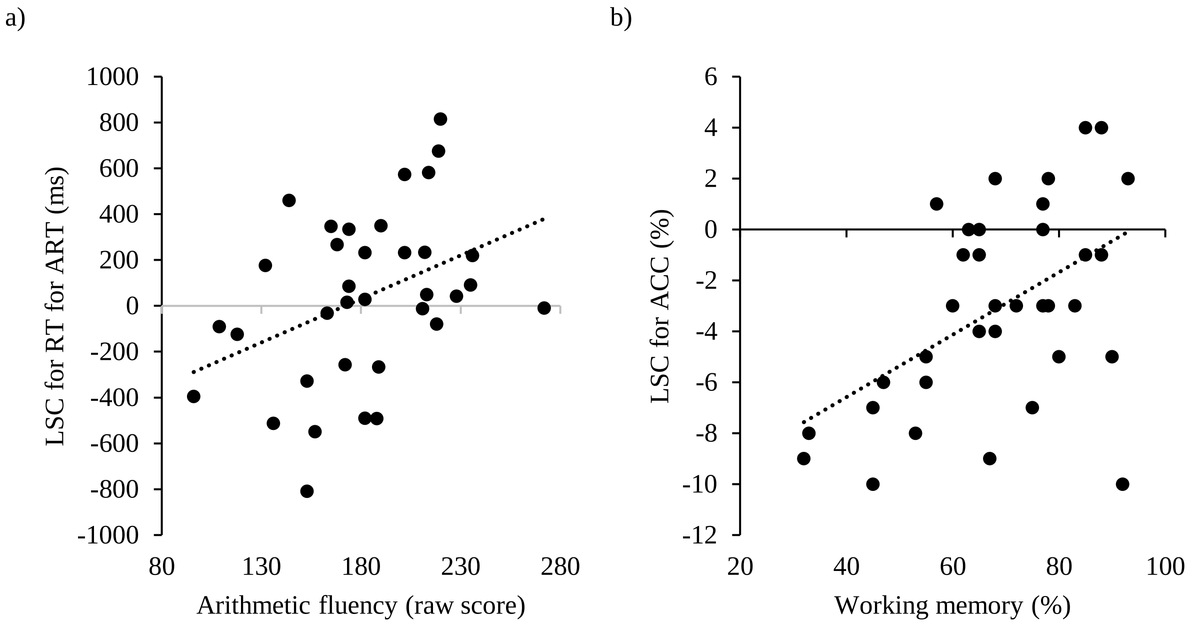
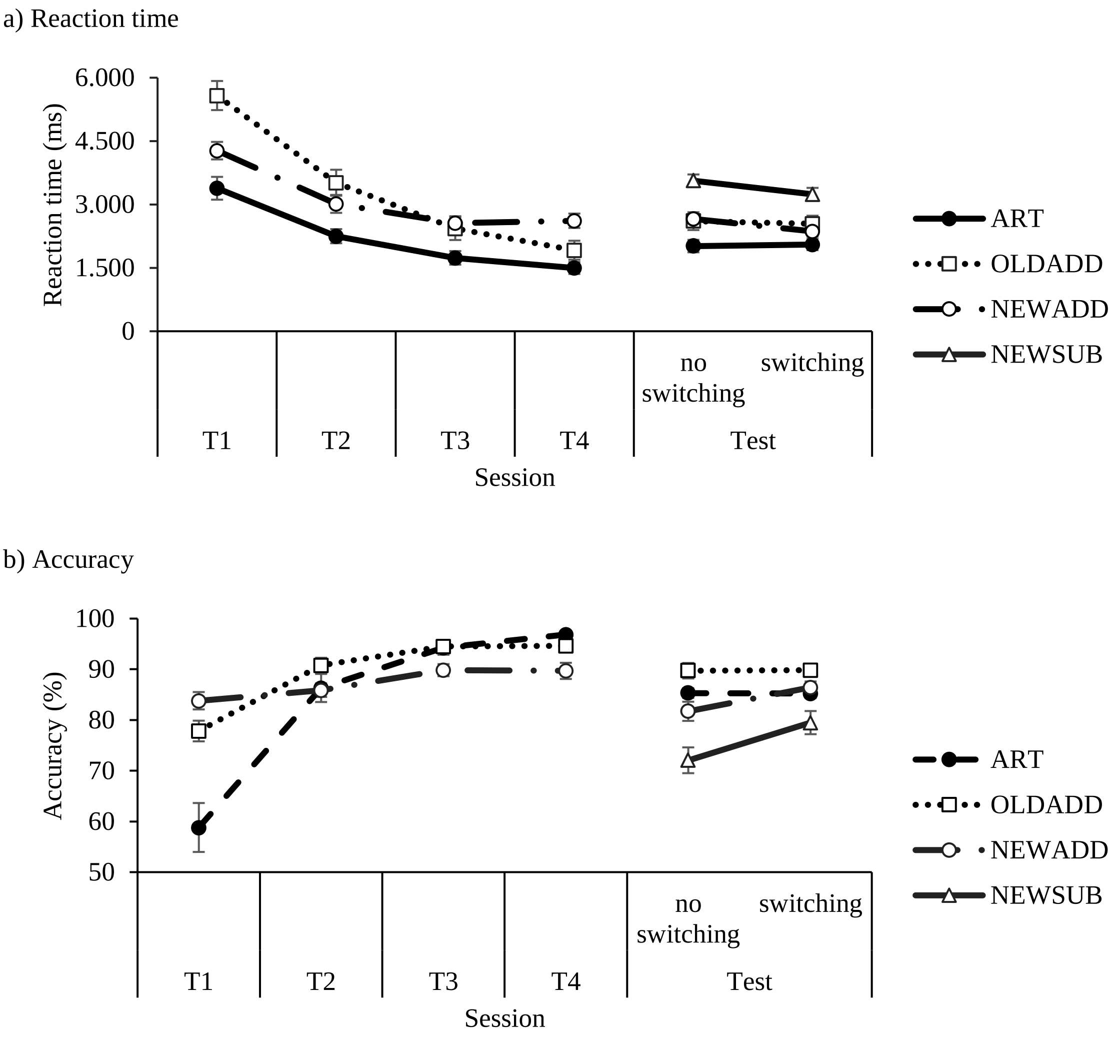


Figure 15. Relationship between a) LSC for RT for ART and scores for arithmetic fluency, and b) LSC for ACC and scores for working memory.

Explorative follow-up analyses

Since results of Study 3 were a big surprise, one possible way to reveal possible explanation is to compare the results of Study 2 and Study 3, since the design was very much alike and the exact same ART problems were used. Figure 16 displays the study progression of RT and ACC. What becomes visible by mere looking is that the performance for NFK (i.e., ART and OLDADD) dropped from the last training session to the no switching condition during testing (i.e., longer RT and lower ACC). To note, problems within training and the no switching condition during testing represent identical problems, presented in the same language. Considering overall results of Study 2 and 3, these follow-up analyses seemed not only important for the interpretation of Study 3 findings, but rather essential for the general discussion of this dissertation (see page 86 et seq.).


a) Reaction time

b) Accuracy

Figure 16. Progression of RT for ART problems over training sessions (T) and testing compared for Study 2 and 3. Error bars indicate the standard error (SE).

Statistical analyses confirmed the notion of a significant drop in performance from training four to the test session for both NFK tasks regarding RT (ART: $t(34) = -6.15$, $p < .001$, $d = 0.61$; OLDADD: $t(34) = -5.68$, $p < .001$, $d = 0.54$) and ACC (ART: $t(34) = 11.00$, $p < .001$, $d = 2.47$; OLDADD: $t(34) = 4.14$, $p < .001$, $d = 0.64$). Based on these findings and the circumstances that a) the test session took up to 60 minutes and b) the statements of the majority of participants expressing their exhaustion to the examiner right after the test session, a closer look seemed appropriate. Having the identical ART problems used in Study 2 – with having the same training and test design – made it possible to directly compare both studies. There was no further investigation of OLDADD since they were not included in Study 2. For the sake of completeness, all Figures contain the data for each of the four tasks.

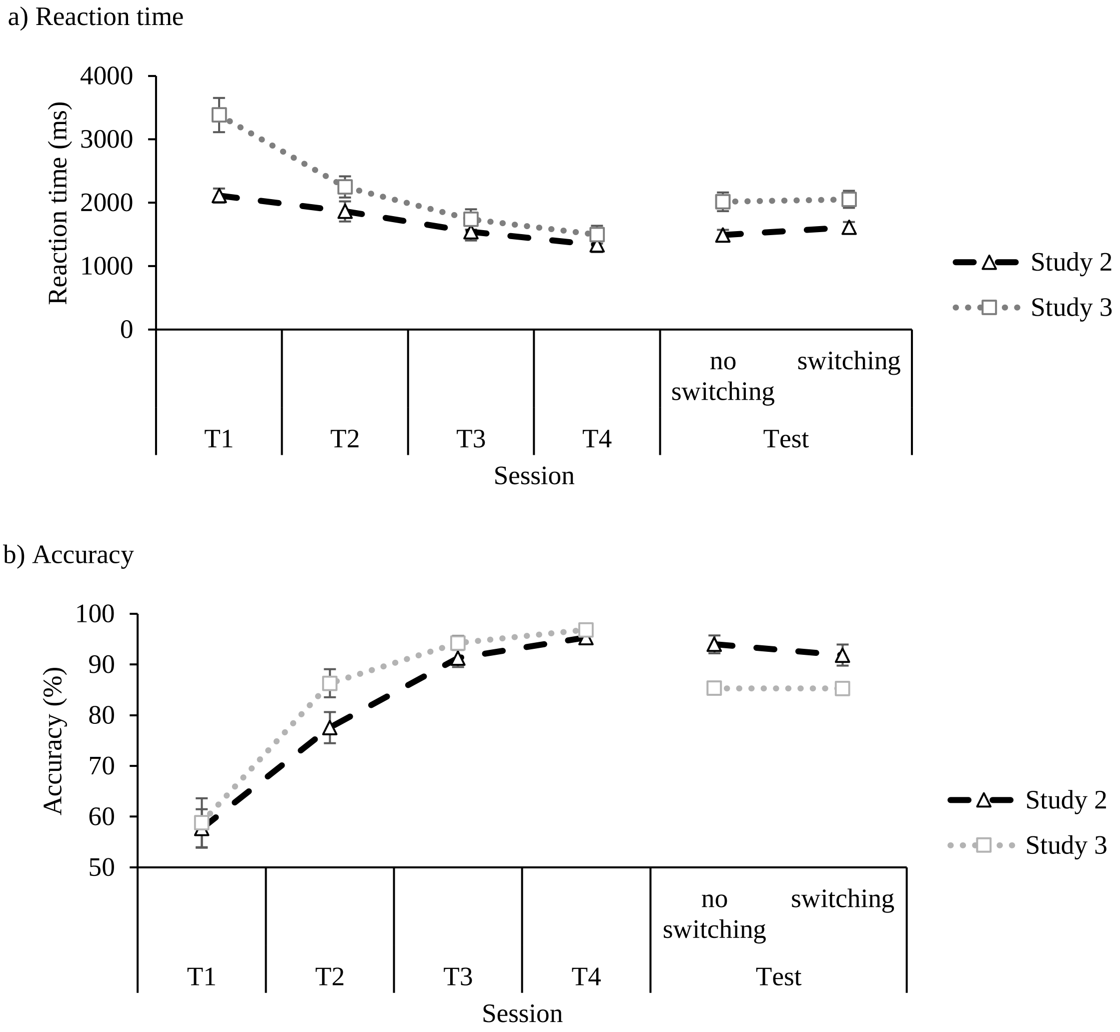


Figure 17. Training and test performance separate for Study 2 and 3 regarding a) RT and b) ACC. Error bars indicate the standard error (SE).

Within Study 2, the performance drop from the last training session to testing was only true for RT, but to a lesser degree (*t*(35) = -2.34, *p* = .025, *d* = 0.27). However, no significant drop in performance was found for ACC (*t*(35) = 1.30, *p* = .200, *d* = 0.15). In a next step, we directly compared Study 2 and 3 regarding training and test performance of ART, using independent sample t-tests. Figure 17 displays the differences for RT and ACC. Analysis showed that ART problems were retrieved significantly faster in Study 2 compared to Study 3 (no switching: *t*(69) = -3.13, *p* = .003, *d* = 0.74 ; switching: *t*(69) = -2.80, *p* = .007, *d* = 0.66), while showing no differences comparing the last training session (*t*(69) = -.92, *p* = .363, *d* = 0.21). The same pattern was true for ACC with higher ACC in Study 2 compared to Study 3 (no switching: *t*(69) = 4.29, *p* < .001, *d* = 1.02; switching: *t*(69) = 2.88, *p* < .001, *d* = 0.67), while finding no differences for the last training session (*t*(69) = -1.13, *p* = .261, *d* = 0.28). Thus, despite equivalent performance of ART on the last training day, performance remarkably dropped in Study 3 only.

In a further step, we compared the distribution of self-reports for ART in Study 2 and 3. This was done to check whether a different distribution of translational processes led to the change in performance of ART (i.e., translation trials had longer RT than no translation trials). The strategy-reports were left out since ART problems can only be solved by retrieval. Regarding the use of translation processes, there was no difference in the use of translation processes between Study 2 and Study 3 (43.0% vs. 49.3%; *t*(69) = .95, *p* = .35, *d* = 0.23)

Finally, we considered the overall distribution of self-reports within Study 2 and 3 including all correct trials. The rationale behind this investigation was the information about exhaustion given by most participants. Hereby, it was shown that participants in Study 2 were retrieving answers from memory in about 90% of the trials, finishing the session 40 minutes or less, while participants in Study 3 needed to calculate in about 50% of the trials – which can be considered more exhausting –, leading to a session duration of up to 60 minutes. Regarding the amount of additional translational processes, there were no differences between the two studies (about 25% translation within both Study 2 and 3). Overall, these analyses indicate that specific factors in relation to the study design influenced the performance of ART problems regarding RT and ACC (for further discussion see page 81 et seq.).

Despite not finding LSC overall, it has to be stressed that 58% of participants showed LSC for ART and 49% for OLDADD. In Study 2, 72% showed LSC for ART. Thus, the overall pattern of self-reports and LSC are very similar to Study 2 (see Figure 18).

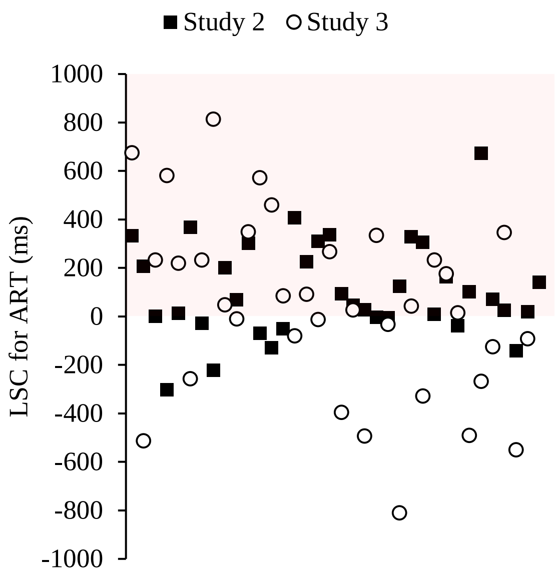


Figure 18. Display of individual LSC for ART in Study 2 (M = 109.5 ms; SD = 197.1 ms) and Study 3 (M = 39.2 ms ; SD = 374.6 ms). A positive value (shaded area) refers to longer RT in the switching condition.

Discussion

The third and final study of this dissertation aimed at looking at LSC in the context of procedural knowledge. Thirty-five psychology students were trained over a period of four days in learning addition problems in a previously unknown number system (i.e., base-7). For all participants, the training language was English (L2). On the fifth day, participants were tested in English as well as in German (L1). The test session included trained base-7 addition problems (OLDADD), untrained base-7 addition problems (NEWADD), and untrained base-7 subtraction problems. As a control task, the same ART problems were included as in Study 1 and Study 2.

Overall, results showed a different picture than expected, yet add crucial information about the nature of LSC. NFK problems (ART and OLDADD) were solved equally fast in the no switching and switching condition, while untrained problems (NEWADD, NEWSUB) were solved faster in the switching condition. Thus, with regard to the untrained problems, we cannot speak of LSC, but rather a language-switching advantage, with performance being overall easier in the switching condition (i.e., the native language). This outcome was true for RT as well as ACC. Self-reports revealed that procedural strategies were used to the same amount in no switching and switching condition. Importantly, within the switching condition, procedural strategies were significantly more present for untrained problems (NEWADD, NEWSUB) compared to trained problems (OLDADD; ART are left out her, since they can only be retrieved). Regarding the second self-report, we found that the percentage use of translation processes was higher in the switching compared to the no switching condition only for the trained problems (ART, OLDADD), while lower for the untrained problems (NEWADD, SUB). Further, the frequency of trials solved with additional translational processes predicted LSC, replicating the finding of Study 2. Finally, there was a positive correlation between working memory span and LSC for ACC.

At first glance, results may appear surprising since no LSC were found for NFK, contradicting the findings of Study 1 and Study 2. In order to find possible explanations for this outcome, follow-up analyses were conducted. Results of Study 1 were not considered here, because the testing design was different (i.e., trial-by-trial language and task switching in Study 2 and 3, compared to block-wise switching in Study 1). Analyses revealed a performance drop from training to testing of NFK problems for RT and ACC, whereas in Study 2, there was only a comparably small drop regarding RT. In Study 2, a slight decrease in RT of about 150ms was foreseeable, since the test session had a constant trial-by-trial switching between languages and tasks, compared to the training, where each block only contained the training language and one specific. In the present data, however, not only did the RT increase by about 500ms, also the error rate increased by roughly 12%. This remarkable change in performance is likely to be explained by the cognitive demand of the tasks used in combination with the block design and length of our test session. Participants needed to be in high alert during the test session, since untrained problems were included within each block. Previous research indeed also included untrained problems, but for well-known tasks (i.e., multiplication and subtraction; Grabner et al., 2012; Saalbach et al, 2013) and shorter test sessions (i.e., in contrast to our studies, test sessions in previous research had overall less trials, not repeating every single problem six times as it happened in our studies). Therefore, the overall cognitive demand was likely to be significantly lower. Base-7 problems demand working memory to a greater extend because only in specific cases the number three must be added compared to addition in base-10. Thus, it is likely that the cognitive demand was too high, therefore reducing overall performance. Several aspects within the present study corroborate this proposition. First, analysis of self-reports showed that roughly 80% of NEWADD had to be calculated, as well as 40% of OLDADD (see Figure 14). As a result and specifically in contrast to Study 2, participants did not get into a flow of retrieving information from memory (i.e., in Study 2, 90% of trials during testing were retrieved). Second, the amount of procedural strategies in Study 3 compared to Study 2 led to a significant raise of the overall duration. Third, and connected to the prior point, participants openly expressed their exhaustion or sometimes sleepiness in post-test conversation with the examiner. Finally, as can be seen in Figure 18 in the previous section, the question can be raised why less people in Study 3 showed LSC for NFK and what caused the higher variety in LSC overall (i.e., double standard deviation of LSC in Study 3 compared to Study 2). Once more, it points to a possible cognitive overload during the test session for at least part of the sample, which may lead to more extreme results.

How can the possible cognitive overload relate to the null result of LSC for NFK? The decrease in performance suggests that participants were at least struggling with their performance regarding problems that were almost flawless just one day before. Even in case that only parts of the sample struggle within the testing phase LSC are likely to disappear. This would further imply that LSC may only appear in situations of high performance when the full concentration can be held with the process of retrieving stored information from memory. From another viewpoint, it is important to note that OLDADD, in contrast to NEWADD and NEWSUB, were not solved faster and more accurate in the native language (i.e., switching condition). Further, OLDADD showed a similar pattern regards both self-reports than ART (see Figure 9 and 14). Thus, a language-dependent knowledge acquisition seemed present. As was true for Study 1, it might be the case that LSC for NFK were masked by the test design. Unfortunately, we are left with speculation.

With regard to the main research question, the present data do not support the idea that procedural knowledge is acquired in a language-dependent way, finding shorter RT and higher ACC for problems in the switching condition. It has to be noted, that the language within the switching condition was the participants´ native language (i.e., German). Since most of the trials for NEWADD and NEWSUB were indicated by the strategy report as being calculated, results show that calculation might be still faster in the mother language, even though the instructions to the procedure were given in English. A problem that jumps into awareness is to question whether participants actually learned the procedure in the language of instruction in the first place. Even though instructions were provided in English, and participants were constantly trained with English stimuli, it cannot be ruled out that in order to understand a new procedure the inner-speech of participants was German. This argument weighs heavily, seeing the significant advantage in the switching condition regarding both RT and ACC. It moreover limits the power of the present data to draw conclusions on LSC and procedural knowledge, thereby pointing towards a weakness of our present study design. We do not have an assurance that participants also adjusted the language of their inner-speech. Within our sample of unbalanced bilinguals (i.e., one language is dominant over the other), it is rather unlikely that this adjustment took place when bearing in mind the complexity of learning a new procedure. Thus, the inner-speech may be an important factor to consider. For the most part when participants perform low in the first training session, they may start to go through the steps the easiest way, which is the native language for unbalanced bilinguals. Within studies on NFK, this factor may take a smaller role, because participants had to rote-learn short equations where procedural steps were automatized due to previous school education in mathematics. However, also in previous research, a directional effect had been stated, arguing that unbalanced bilinguals may have to rely to a greater extent on their native language when learning in L2, therefore diminishing the language-content connection during training for the L2 training group (Marian & Fausey, 2006; Saalbach et al., 2013). Future research may consider testing balanced bilinguals, or at best directly contrasting unbalanced and balanced bilinguals. In sum, hypothesis 1 was not supported (no LSC for NFK), as well as no LSC were found for procedural knowledge (NEWADD) and transfer effects to a new task within the same new number system (NEWSUB).

Looking at the distribution of self-reports, it was found that participants used the same amount of procedural strategies during the switching condition and no switching condition (dismissing Hypothesis 2a). Whereas it was likely that participants calculate mainly for NEWADD and NEWSUB, it was surprising that participants calculated OLDADD to the same amount in both conditions. This may be interpreted that either OLDADD might not have been rote-learned to perfection or that the testing phase, as outlined before, influence the overall performance on even rote-learned trials. Thus, participants were not able to retrieve answers as easily as during the last training session or even decided to quickly calculate again when not being sure. This remains speculative, since no strategy reports were collected during training. Thus, it might be the case that also during the last training session, OLDADD were calculated in half of the trials. Considering the second self-reports, it was found that more translation processes were required during the switching condition (supporting Hypothesis 2b) In about 45% of the trials, additional translation was used in the switching condition compared to the no switching condition, thereby replicating the findings of Study 2. This finding corroborates the data of Study 2, indicating that an intensive learning of specific content will be connected to the language of instruction. Looking at the different tasks, this pattern was strong for ART and OLDADD, and opposite for NEWADD and NEWSUB. For NEWADD and NEWSUB, participants used more translation processes in the no switching condition, supporting the assumption that the new procedure was never truly acquired in the language of instruction or that there is no language-dependent learning for procedural strategies. This finding adds solid evidence for a language-dependent knowledge acquisition especially for NFK. Further evidence provides the finding that the frequency of translational processes predicted LSC (supporting Hypothesis 3; replicating findings of Study 2).

As was true for Study 2, ART trials were also in the present study indicated in about 43% of the trials as being solved without translation processes. The possible explanations have already been discussed before, concerning a possible training effect within the test session or the possibility of non-reliable self-reports (see page 60 for a detailed discussion).

With regard to individual characteristics, results are challenging to interpret as well. Regarding LSC for RT, there was a positive correlation between LSC of ART and arithmetic fluency. Since arithmetic fluency is a measure for speed in executing arithmetic problems and ART problems cannot be calculated, there is no rational argument for the relation. Thus, we refute to interpret this result further. With regard to LSC for overall ACC, we found a strong positive correlation with the measure for working memory span (i.e., 3-back task). The finding indicates that the higher the score for the 3-back task, the better the performance in trials in the trained language (L2) compared to the native language. Thus, the more likely you show LSC (i.e., shorter RT in L2 compared to the native language). Important to note is that this correlation was visible for all four tasks (i.e., significant for ART and NEWSUBB, marginally significant for OLDADD and NEWADD), therefore, showing a strong general pattern. If we consider that at least part of the participants may never truly connected the new procedure to the language of training (i.e., English), it can be assumed that base-7 calculation in the foreign language comes with a higher cognitive demand than performing it in the native language. Thus, despite the fact that training took place in L2, for base-7 addition, it seems likely that most of the procedural steps during the training session were self-instructed via the German language. The 3-back task represents a difficult version of possible n-back tasks. It requires participants to be highly alert while constantly updating numbers shown on a screen. Similar to the 3-back task, participants need to be highly alert in base-7 calculation to recognize when an extra addition has to take place to solve the problem. Thus, the lower your score for the 3-back task, the more struggle you have with calculation in base-7 overall, and possibly with calculation in the foreign language. This may in the end lead to better performance in the native, compared to the foreign language.[7] With regard to inhibitory control, the scores within our sample showed a clear ceiling effect (M = 93.3%), diminishing the meaningfulness of the test. In sum, the present data are not able to add reliable evidence regarding the relationship of individual characteristics and LSC, and ought to be interpreted with caution as it applies for all findings of the present study. The only significant correlational finding may tell that some people were better able to handle the high demanding test situation, yet may not deliver reliable information on the appearance of LSC. In sum, the study results are difficult to interpret and ask for further research in order to draw conclusions on the language-dependency of procedural knowledge. However, the study represents a good start in that respect, providing insight about methodological difficulties for laboratory studies.

As for the results of Study 2, the elaborated discussion of theoretical and practical implications are integrated into the general discussion of the dissertation (see page 86 et seq.).

[7] Figure 15 shows that the seven participants with the lowest scores for working memory performed 5 to 10% higher in the native language.

General Discussion

The present project was built upon prior research showing that performance decreases when the language of instruction and application differ. Due to several open questions in the field, the aim was to further the insights on LSC in the context of bilingual learning, with specific attention to the field of mathematics. For this purpose, three studies were conducted taking different aspects of LSC into focus. Study 1 focused on LSC for auditory stimuli as well as comparing pure fact learning with arithmetic fact learning. Study 2 set the focus on underlying mechanisms of LSC using self-reports. Study 3 shifted the focus to another knowledge type, namely investigating LSC for procedural knowledge. In all three studies, individual characteristics were collected in addition as a secondary objective to gain potential insight about possible predictors. The following sections provide an overview of the main findings. Further, these findings will be discussed in a broader context including possible theoretical and practical implications. Finally, we will address the limitations of the current project and explore opportunities for future research.

Study 1

In Study 1, thirty-two university students were trained to learn problems of three different operations in either their native language (German; L1) or their first foreign language (English, L2). After the four day training, problems were considered as numerical fact knowledge (NFK), with the solution of the equation simply being retrieved from memory in short time. On a fifth day, participants were tested in both languages (see Figure 2 for the detailed design). First, LSC were found regards RT for all three tasks. This finding marks a highlight of this dissertation by being the first data on LSC showing that LSC do not only occur for common arithmetic problems but also for a pure fact-learning task. Moreover, by finding no differences between MUL, SUB and ART problems, the study provided a first indication that rote-learned arithmetic problems may be comparable to pure facts regarding RT and ACC. Finding no difference between MUL and SUB problems replicate findings by Grabner et al. (2012) as well as Saalbach et al. (2013). Neuroscientific data may be used in future studies to examine whether this finding is only true on a behavioral level or even goes back to comparable or even identical neuronal processing of pure and arithmetic facts. Second, regards ACC, there were no differences between the no switching (i.e., when language of instruction and application match) and switching condition (e.g., when language of instruction and application differ). This finding is converging with previous research, almost exclusively finding LSC for RT (Spelke & Tsivkin, 2001; Venkatraman el al., 2006; Grabner et al., 2012, cf. Saalbach et al., 2013). The present test session design most likely promoted the null result, as participants had a long period to response for each trial. The long response window was primarily implemented to not stress participants additionally, which may confound results for ACC. Third, results did not differ between the two training groups, providing evidence that regards NFK the match of language itself is important, not the particular direction of switching (i.e., L1 to L2, or vice versa). This finding adds evidence to the inconclusive findings in the literature (Saalbach et al., 2013; Volmer et al., 2018, but see Grabner et al. 2012; Study 1). Fourth,, LSC did only appear for one test group, namely, when participants were confronted first with the switching condition before the no switching condition. The other half of the sample did first solved problems in the no-switching condition, therefore undergoing a kind of extra training (i.e. although without feedback), before solving the same problems of each task in the switching condition. The study leaves the question open whether the effect was simply masked by the current design, or LSC can already be prevented, when examination would include a short additional training to pre-activate the content. The latter would promote the view that LSC appear because participants had to wait one day in order to be tested. The argument is challenged by the study of Saalbach et al. (2013) who conducted the last training session prior to the trial-by-trial switching test session and did not find such effects. It has to be noted that it is problematic to compare results from studies using written number words with studies using auditory stimuli. Overall, what this tells us is that LSC are sensitive to the test design. A promising approach might be to contrast a block-wise task and language switching with a trial-by-trial task and language switching within the same study.

Study 2

Study 2 was built up upon the findings of Study 1. Study 1 provided first knowledge on LSC regarding pure fact learning as well as in the context of auditory stimuli. Study 2 went a step further, namely integrating self-reports to investigate underlying mechanisms of LSC. Therefore, thirty-six university students were trained and tested. Since previous research had been inconclusive about the mechanisms (see page 14 et seq. for summary of findings by Venkatraman et al., 2006, and Grabner et al., 2012), we integrated two different self-reports hoping to cover the two potentially involved mechanisms (i.e., use of procedural strategies and use of translation processes). First, we replicated the findings of Study 1. It has to be mentioned that in Study 2 the auditory stimuli were slightly adapted: Stimuli were not identical regards to their length in time. Further, SUB problems were made easier to solve by lowering the problem size. Moreover, the test design changed from a block-wise switching regards task and language to a cognitively more demanding trial-by-trial switching. Finally, data collection for RT and ACC changed from a key-press to a more sensitive voice-key (i.e., participants spoke the answer into a microphone). Thus, finding LSC for all three task regards RT and not for ACC, independent of the two training groups (i.e., German vs. English), adds further evidence to the robustness of LSC for NFK. It further corroborated the idea that NFK as well as pure facts are stored in long-term memory in a language-dependent knowledge format.

Most interestingly and the main aim of Study 2, results provided impressive insights into underlying mechanisms. It seems that self-reports hit the nerve in order to tackle that important question and provide support for the expressive power of behavioral data. While strategy reports and RT provided evidence that problems were easily and quickly accessible by retrieval from memory in both languages, translation reports revealed that translation processes were used as a mechanism to speak out the answer in the right language in about half of the trials of the switching condition. Combining the two reports, this means that the answer was often ready, but in the wrong language. This finding adds key evidence to the question if NFK is learned in a language-dependent way. To put it simple: If NFK is not connected to the language of instruction, why would there be any trial at all that includes translational processes? This is especially interesting for the English training group, because when confronted with an arithmetic problem in their native language, in about half the trials, participants made use of translational processes. Further evidence, that can unfortunately not backed up by protocolled data, is the circumstance that most of the few participants´ errors were caused by answering in the wrong language. Thus, when hearing a problem, for instance in English, participants gave the correct answer, but in the wrong language. Therefore, the first step was retrieving the answer in the wrong language. The second step included than a mere translation of that answer in order to respond. Obviously, this takes more time than retrieving an answer without translation, hence, leading to longer RT in the switching condition, called LSC. Thus, the inclusion of a voice-key in Study 2 was necessary to reveal these type of errors made by participants, therefore marking an important methodological improvement compared to Study 1. For the results of Study 1, we cannot be sure that the RT collected via the keypress always represents the moment in which participants had the answer present in the language of application. Hence, participants may have already pressed the key after retrieving the answer in the wrong language. The voice key in Study 2 rules out this possibility since the examiner was present during the whole session, marking this kind of errors to take them out of the analysis. It is likely that these errors mainly occurred in the switching condition, so why did we again not find LSC for ACC?

Study 2 corroborates the view that LSC for ACC may depend on the research design of the test session. If there is enough time and a production task, mistakes are expected to be rare, ending up in a ceiling effect (e.g., Study 1 and 2 with ACC > 90%). In contrast, if answering is forced by giving participants only a quick moment to answer, mistakes are more likely to happen (e.g., Grabner et al. (2012) with ACC of about 83% (switching condition) and 87% (no switching condition) for trained problems. Overall, by now it seems reasonable to proclaim that RT is to be preferred as a measure for LSC in contrast to ACC regards NFK. This is also supported by the results of self-reports, stating that LSC may be primarily due to additional translational, and to a lesser extent to numerical processing. Since these processes take time, they will more likely influence RT in contrast to leading participants to make errors. Finally, there was no evidence that the individual characteristics assessed (i.e., general intelligence, indicator for language proficiency of L2, and arithmetic fluency) show any relation with LSC.

Overall, from a theoretical point of view, Study 2 adds crucial evidence to the relation between language and arithmetic knowledge acquisition by not only again showing that language of instruction matters but also revealing further evidence on underlying mechanisms of LSC. From a practical point of view, we remain reserved with strong implication for CLIL programs. Despite improvements in ecological validity, the applied setting is still not easily comparable to a testing situation in a classroom context. However, the evidence suggests that if LSC are mainly caused by additional translational processes and reveal themselves in the form of longer RT, then the examination in the CLIL context should consider that pupils may need some additional time compared to the same exam in a traditional context. This may be only a few minutes, but it may provide that extra moment without additional stress to switch between languages and come up with the right answer.

Study 3

Study 3 was the first to investigate the possible language-dependency of procedural knowledge in contrast to previous research primarily focusing on NFK. In addition, it was aimed to replicate findings of Study 2 relating to LSC for NFK and the underlying role of individual strategy and translation use. Finally, individual characteristics were again considered, with additional tests on working memory and inhibitory control that were not used in Study 1 or 2. Therefore, thirty-five university students were trained and tested. First, only parts of the findings of Study 2 were replicated. In contrast to Study 1, no LSC for RT were found for ART and OLDADD, finding equal speed in performances for both conditions. Despite not finding LSC for the sample, the individual distribution of self-reports was again able to predict LSC as was shown in Study 2. As outlined in the discussion of Study 3, we propose that the null result for LSC for NFK may have been masked by the special circumstances of the study design. Overall, and in contrast to Study 1 and 2, the individual value for LSC varied widely within the sample. From the data and impressions collected, there seems to be no other explanation than to propose an overload of cognitive demand for at least part of the sample. This may question the validity of the RT data of Study 3. Did we really measure how fast and accurate participants retrieve facts from memory when the languages of instruction and application match compared to when they differ? Or did we measure how good participants are able to deal with a cognitive demanding task? On the other hand, Study 3 may give rise to an entire new question: do LSC only appear in highly performing situations, when participants are tested on rather simple tasks. Admittedly, previous research did not only include trained problems during testing, but also untrained problems (see Grabner et al., 2012; Saalbach et al, 2013). Still, those untrained problems were common arithmetic tasks and not comparable to addition in base-7, requiring to keep the concentration constantly high. Thus, it might be the case that the advantage of a match between language of instruction and language of application only shows when the subject content is rather easy and fast to access. In these situations, additional processing in the switching condition may manifest itself in the form of RT differences. In contrast, when the subject content is more complex or the testing situation becomes cognitively demanding and tiring, results reveal more variation in the data (see Figure 18), which then may mask effects and/or make them statistically disappear.

Considering the observation that LSC are mainly within the range of hundreds of milliseconds and were found in studies where from the outside look, participants were not asked to go through highly demanding test sessions, it appears that LSC are a phenomenon that shows up in high performing situations, when knowledge has to be retrieved in short time from memory. In the moment of a more demanding, rather tiring situation, there might be so many distinct factors influencing performance that the advantage of the match between language of instruction and language of application vanishes or is at least masked by additional circumstances. What this may imply is that for the field and especially individual cases it may not be relevant whether there are overall LSC in class, knowing that this specific individual will struggle a lot in CLIL context. Therefore, we urgently point to the need to investigate individual differences that may or may not lead to LSC more closely. With regard to individual characteristics, we found a strong positive correlation between working memory and LSC for the overall ACC. Thus, performance differences between no switching and switching increased in favor of the native language (switching) the lower the working memory score. To further interpret these results, a German training group would have been helpful. Unfortunately, we did not do so because of prior negative findings regarding a directional effect of LSC. Future research should again include both training groups in order to help interpreting such results.

Concerning the main research focus of Study 3, there was no evidence for a language-dependency of procedural-knowledge. It is important to note that we question that at least part of the participants learned the procedure in the training language in the first place. We assume that participants which had problems understanding the procedure at the beginning, self-instructed themselves in their native language German. This may explain why many participants – when looking on an individual level (see Figure 18 and Supplementary Material for more detail) – showed strong advantages for RT and ACC in the switching condition, therefore, turning our expectations upside down. Consequently, it is crucial to further investigate LSC in relation to procedural strategies and improve research designs that can track more appropriately whether or not participants self-instruct themselves in the training language. On the other hand, these findings raise the question whether newly acquired procedural knowledge, in the context of CLIL programs, is connected to the language of instruction or – in the case of unbalanced bilinguals – rather to the language that is more proficient. This is a very important aspect to optimize CLIL programs, requiring more precise field research. So far, it remains speculative, since the present research did not investigate CLIL itself. Thus, even though teachers are giving classes in English, pupils switch to their native language when the content is getting too complex to understand. Thus again, it is likely that the language-proficiency is a major factor, pointing to studies directly comparing unbalanced and balanced training groups.

Concisely, due to the mentioned limitation of the design, results of Study 3 are reported and interpreted with caution and have limited implications for practice. Much of the implications are based on speculation. However, they are crucial for the design of future research. We further refrain to draw further theoretical or practical implications from Study 3 based on the present limitations.

New insights on LSC

Starting this project off, we summarized findings of previous groundwork in the field of LSC, with the special focus on mathematics. The present project adds three new studies with each adding information to the field:

LSC for NFK:

- ✓ can be found for the language combination German and English (Study 1; Study 2; cf. Study 3)
- ✓ can be found for RT by using auditory stimuli (Study 1; Study 2; cf. Study 3)
- ✓ do not appear for ACC when the time to answer is lengthened (Study 1, Study 2)
- ✓ are comparable, rather identical for pure facts and arithmetic facts (Study 1; Study 2)
- ✓ may be masked or prevented depending on the research design (Study 1; Study 3)
- ✓ are mainly caused by additional translational processing (Study 2; Study 3)
- ✓ do not show an directional effect (Study 1; Study 2)
- ✓ may not stand in relation or are difficult to set into relation with individual characteristics (Study 1; Study 2; Study 3)

Further, LSC do not appear for procedural knowledge (Study 3).

Recently, Volmer et al. (2018) provided further evidence on LSC in a German-French sample consisting of fifty-eight university student. In their study, the same auditory stimuli were used as in Study 2 and 3. Only for the German training group, LSC were found for NFK for RT, not for ACC. Further, LSC were found for NFK that were integrated into mathematical text problems, also only for the German training group. Regarding individual characteristics, the study revealed that the LSC for the German training group negatively correlated with the individual score for vocabulary knowledge in French (L2). The results shed more light on the possible interplay between LSC and individual characteristics and may help to understand that some studies find a directional effect of LSC (e.g., Saalbach et al., 2012; Volmer et al., 2018), whereas others do not (e.g., Grabner et al., 2012; Study 1, Study 2). As was already mentioned within the discussion of Study 2 (see page 61), the average L2 score of the sample was reasonable lower than the score within the studies of the present project. Thus, it is likely that LSC are higher, the more unbalanced the sample. As was pointed out by Marian and Fausey

(2006), and already discussed above, unbalanced bilinguals may rely more heavily on their native language, when training in L2. Thus, LSC are not likely expected because the native language (i.e., switching condition) has already been used to a great extend during the learning phase. The sample of Volmer et al. (2018) rather represented an unbalanced sample with an average score of 62%, compared to scores above 80% in our studies. Thus, the better your L2 proficiency, the more likely you show LSC in both directions, since you are comfortable in either language during the learning phase. In other words, the less balanced a person, the more likely is a directional effect regarding LSC. The more balanced, the more likely are LSC in both directions. Future research may address this issue.

Limitations, open questions, and future research

Despite new insights provided by the present research project, new questions arise at the same time. It is crucial to point out that each of the three studies had a limited focus. The goal was to further the insights in the field of language-switching costs step by step, gaining specific insights from study to study. Further, studies 2 and 3 had the goal to at least confirm (i.e., replicate) results of the previous studies. Therefore, we tried to keep the experimental designs and stimuli in use constant as far as the specific research objectives allowed, to overall account for internal validity within the research project. One of the most critical limitations of the whole project was the compilation of the experimental groups in all studies. Despite the results, we are far from concluding that individual characteristics play no role for the appearance of LSC. Future research is needed trying to capture specific abilities in a more precise manner, instead of using the most convenient ways when it comes to the management of own resources (e.g., using a vocabulary knowledge test that is finished within a few minutes). With the benefit of hindsight, it would have been necessary and helpful for interpretation of data to assess individual characteristics in a clear screening session to set up a sample that shows variety in the test measures. Instead, the sample was fixated in a “first come, first served” fashion due to time-management reasoning. Unfortunately, the samples in all three studies were rather homogeneous, especially with regard to language proficiency, questioning the expressive power of results. Another major issue was the test design in Study 3. The overall testing time reached up to one hour. On second thoughts, the two additional blocks including a transfer task, may have overloaded the study and influenced the outcomes in a way that make them hard to interpret.

Taking the findings of the present project into account, we propose that future research needs to conduct studies that directly compare different testing designs and set up more heterogeneous training groups (i.e. especially with regard to individual differences). This is especially true for L2 language proficiency as was discussed in previous section, relating to the findings of Volmer et al. (2018). If lower language-proficiency leads to higher LSC, then this finding will be very important for CLIL, since it provides evidence against a thought that language and content can be easily taught simultaneously. The present project adds crucial evidence to rather fundamental questions, but makes it hard to draw conclusions for the field and therefore CLIL. Especially with regard to CLIL, different knowledge types need to be investigated in relation to possible LSC, because daily classroom interaction is not only concerned with fact knowledge, but also procedural as well as conceptual knowledge. Whereas it seems by now rather undebatable that the individual speed during performance will decrease, present data do not provide sufficient evidence to make a statement about the language-dependency of procedural knowledge, with no focus yet on conceptual knowledge at all. The second next step may then be to investigate LSC in real classroom settings, when having a well thought research design that proved to work in laboratory settings for different knowledge types.

Additional remarks on CLIL

At the beginning of this project, we pointed out that one of the main intentions of the growing number of bilingual education programs is the idea to kill two birds with one stone: learning specific content while simultaneously learning a second language. With research providing evidence that bilingualism as well as CLIL comes with benefits, there seems to be no apparent reason to view this development critically. However, not only the view on the positive effects that are related to bilingualisms (for a review see Bialystok, 2018), but also the positive outcomes of empirical research on CLIL, are starting to crock. More and more empirical evidence is accumulated, questioning the promising advantages of bilingual education due to probable publication bias and/or methodological shortcomings (for a critical overview on the effect of bilingualism on executive functioning see Paap and Sawi (2014), and de Bruin, Treccani, and Della Sala (2014).

With regard to CLIL, Roussel et al. (2016) provides a great overview of the current data on studies on CLIL. This whole issue cannot be discussed within this dissertation in close detail, but it is important to understand that research – such as the current projects´ investigation of LSC – that is concerned with advantages and disadvantages of bilingual education needs to get attention. Therefore, some observations will be stated in the following. Roussel et al. (2016) conclude that the evidence is at most mixed and very much inconclusive on CLIL. A recent two-year longitudinal research by Rumlich (2017) found no improvement of second language proficiency that could be attributed to the CLIL approach. Bruton already stated in 2011 that many of the advantages found in studies comparing CLIL classes to non-CLIL classes can be traced back to the selective nature of CLIL programs with regard to staff and pupil composition (see also Dallinger, Jonkmann, Hollm, & Fiege, 2016; cf. Lorenzo, Moore, & Casal, 2011). In consequence, we have to be careful supposing that positive research outcomes from studies examining CLIL programs are automatically generalizable to each and every school implementing a form of CLIL. In order to further question the CLIL approach. Roussel et al. (2016) took the cognitive-load theory into focus in their critical dispute, arguing that CLIL ignores the cognitive human architecture, and therefore evolutionary psychology. It is argued that learning a second language represents secondary knowledge. While primary knowledge comes rather effortless, like learning the native language by merely being in a specific language context, secondary knowledge requires conscious effort (see Geary and Berch, 2016 for an overview). In the context of CLIL, while learning specific content can be done easily in the combination with primary knowledge, it becomes highly demanding in the context of secondary knowledge. In their study, Roussel et al. (2016) showed that learning in a second language without additional language training comes to the expense of performance on the specific content taught. In three experiments, they showed that knowledge acquisition was better in the native language compared to two different conditions using the foreign language as instructional tool with further language instruction being withhold from the participants. In 2016, Piesche et al. even showed 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. Such evidence raises the question whether learning of basic concepts (e.g., “Floating and Sinking”) or basic arithmetic shall be learned in the language in which the knowledge will be applied.

Summarizing, the current project not only provides supplementary evidence on the fundamental question of the language-dependent acquisition of knowledge, but also adds evidence from a practical point of view. In line with the observations reported in the previous paragraph, it is argued that the implementation of bilingual education programs, especially with respect to learning basic numerical knowledge, should be well thought out. Empirical evidence provided indisputable evidence that language does play a role in knowledge acquisition and should be taken into account when teaching. Especially Study 3 showed that performance in untrained content in combination with a cognitively demanding context is best when performed in the native language (i.e. participants performed better in their native language even though training took place in another language). Generally, we can ask if we overcomplicate learning environments by including another language in the learning and/or testing context. On the one hand, learning new content and foreign language together may put unnecessary load on the working memory (e.g., Sweller, Ayres, & Kalyuga, 2011). Further, it may add additional stress in a school context that is already marked by increasing self-reported stress and stress related health problems (WHO, 2008; WHO, 2016). We might not kill two birds with one stone but may rather 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 rudimentary knowledge, which builds the foundation for future learning. Overall, the fundamental research in the area of knowledge acquisition remains important, to understand how subject matter content and language of acquisition interact. If it seems to be so relevant to focus on time efficient teaching, we then rather propose to change the content of current second language classes instead of changing the language in class such as mathematics. In order to learn a second language, teaching may use the content that was already learned in other classes. Therefore, the content will be covered again and most likely better consolidated.

References

Ashcraft, M. H., & Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning and Memory, 4*(5), 527.

Ashcraft, M. H., & Stazyk, E. H. (1981). Menatal addition: A test of three verification models. *Memory & Cognition, 9*(2), 185-196.

Baker, C. (2011). Foundations of bilingual education and bilingualism (Vol. 79). Multilingual matters.

Barber, S. J., Rajaram, S., & Aron, A. (2010). When two is too many: Collaborative encoding impairs memory. *Memory & Cognition, 38*(3), 255-264.

Benn, Y., Zheng, Y., Wilkinson, I. D., Siegal, M., & Varley, R. (2012). Language in calculation: A core mechanism? *Neuropsychologia, 50*(1), 1-10.

Bialystok, E. (2018). Bilingual education for young children: review of the effects and consequences. *International journal of bilingual education and bilingualism, 21*(6), 666-679.

Boroditsky, L., Fuhrman, O., & McCormick, K. (2011). Do English and Mandarin speakers think about time differently? *Cognition, 118*(1), 123-129.

Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System, 39*(4), 523-532.

Bruton, A. (2013). CLIL: Some of the reasons why… and why not. *System, 41*(3), 587-597.

Campbell, J. I., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General, 130*(2), 299-315.

Campbell, J. I. (2005). Asymmetrical language switching costs in Chinese–English bilinguals' number naming and simple arithmetic. *Bilingualism: Language and Cognition, 8*(1), 85-91.

Chen, Z. Y., Cowell, P. E., Varley, R., & Wang, Y. C. (2009). A cross-language study of verbal and visuospatial working memory span. *Journal of Clinical and Experimental Neuropsychology, 31*(4), 385-391.

Cheng, L., Li, M., Kirby, J.R., Qiang, H., & Wade-Woolley, L. (2010). English language immersion and students' academic achievement in English, Chinese and mathematics. *Evaluation & Research in Education, 23*(3), 151–169.

Claessens A., Duncan G., Engel M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review, 28*(4), 415–427.

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.

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

Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex, 33*(2), 219-250.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20*(3-6), 487-506.

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current opinion in neurobiology, 14*(2), 218-224.

Delazer, M., Domahs, F., Bartha, L., Brenneis, C., Lochy, A., Trieb, T., & Benke, T. (2003). Learning complex arithmetic—an fMRI study. *Cognitive Brain Research, 18*(1), 76-88.

De Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive advantage in bilingualism: An example of publication bias?. *Psychological science, 26*(1), 99-107.

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.

Domahs, F., & Delazer, M. (2005). Some assumptions and facts about arithmetic facts. *Psychology Science, 47*(1), 96-111.

Dowker, A. (2005). *Individual differences in arithmetic: Implications for psychology, neuroscience and education.* Hove: Psychology Press.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Sexton, H. (2007). School readiness and later achievement. Developmental psychology, 43(6), 1428.Ellis, N. C., & Hennelly, R. A. (1980). A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, *71*(1), 43-51.

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

Eurydice. (2006). *Content and language integrated learning (CLIL) at school in Europe.* Brussels, Belgium: Eurydice European Unit.

Fausey, C. M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic bulletin & review, 18*(1), 150-157.

Frenck-Mestre, C., and Vaid, J. (1993). Activation of number facts in bilinguals. *Memory and Cognition 21*, 809–818.

Fuson, K. C., & Kwon, Y. (1992). Learning addition and subtraction: Effects of number words and other cultural tools. In Bideaud, J., Meljac, C., & Fischer, J. P. (Eds.). (2013). *Pathways to number: Children's developing numerical abilities.* Psychology Press.

Geary, D. C. (2013). Early foundations for mathematics learning and their relations to disabilities. *Current directions in psychological science*, *22*(1), 23–27.

Geary, D. C., & Berch, D. B. (2016). Evolution and children’s cognitive and academic development. In *Evolutionary perspectives on child development and education* (pp. 217-249). Springer, Cham.

Gentner, D., & Goldin-Meadow, S. (2003). Whither whorf. *Language in mind: Advances in the study of language and cognition*, 3-14.

Göbel, S. M., Moeller, K., Pixner, S., Kaufmann, L., & Nuerk, H. C. (2014). Language affects symbolic arithmetic in children: the case of number word inversion. *Journal of experimental child psychology*, *119*, 17-25.

Grabner, R.H., & B. De Smedt. 2011. Neurophysiological evidence for the validity of verbal strategy reports in mental arithmetic. *Biological Psychology 87*(1): 128–136.

Grabner, R. H., & De Smedt, B. (2012). Oscillatory EEG correlates of arithmetic strategies: a training study. *Frontiers in psychology, 3*.

Grabner, R. H., Saalbach, H., & Eckstein, D. (2012). Language-Switching Costs in Bilingual Mathematics Learning. *Mind, Brain, and Education, 6*(3), 147-155.

Gumperz, J. J., & Levinson, S. C. (1996). Introduction to part I. In Gumperz, J. J. (1996). *Rethinking Linguistic Relativity*, pp. 21-36. UK: Cambridge University Press.

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.

Hüttner, J., & Smit, U. (2014). CLIL (Content and Language Integrated Learning): The bigger picture. A response to: A. Bruton. 2013. CLIL: Some of the reasons why… and why not. System 41 (2013): 587–597. *System, 44*, 160-167.

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.

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

Johnson, R. K., Swain, M., & Long, M. H. (Eds.). (1997). Immersion education: International perspectives. Cambridge University Press.

Jost, K., Beinhoff, U., Hennighausen, E., & Rösler, F. (2004). Facts, rules, and strategies in single-digit multiplication: evidence from event-related brain potentials. *Cognitive Brain Research, 20*(2), 183-193.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 615.

Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of educational psychology, 103*(3), 547.

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.

Klessinger, N., Szczerbinski, M., & Varley, R. (2012). The role of number words: the phonological length effect in multidigit addition. *Memory & cognition*, *40*(8), 1289-1302.

Köller, O., Leucht, M., & Pant, H. (2012). Effekte bilingualen Unterrichts auf die Englischleistungen in der Sekundarstufe I. Unterrichtswissenschaft, 4(4), 334-350.

Lasagabaster, D., & Sierra, J. M. (2009). Immersion and CLIL in English: more differences than similarities. *ELT journal*, *64*(4), 367-375.

Lee, K. M. (2000). Cortical areas differentially involved in multiplication and subtraction: a functional magnetic resonance imaging study and correlation with a case of selective acalculia. *Annals of neurology*, *4*8(4), 657-661.

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.

Lemer, C., Dehaene, S., Spelke, E., & Cohen, L. (2003). Approximate quantities and exact number words: Dissociable systems. *Neuropsychologia*, *41*(14), 1942-1958.

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.

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.

Lorenzo, F., Moore, P., & Casal, S. (2011). On complexity in bilingual research: The causes, effects, and breadth of content and language integrated learning—a reply to Bruton (2011). *Applied Linguistics*, *32*(4), 450-455.

Mackworth, J.F. (1959). Paced memorizing in a continuous task. *Journal of Experimental Psychology*, *58*(3), 206–211.

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.

Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory and Language*, *51*(2), 190–201.

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, *129*(3), 361.

Meuter, R. F., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of memory and language, 40*(1), 25-40.

Miller, K. F., Smith, C. M., Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science*, *6*(1), 56-60.

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

Möller, J., Hohenstein, F., Fleckenstein, J., Köller, O., & Baumert, J. (Eds.). (2017). Erfolgreich integrieren-die Staatliche Europa-Schule Berlin. Waxmann Verlag.

Núñez-Peña, M. I., Cortiñas, M., & Escera, C. (2006). Problem size effect and processing strategies in mental arithmetic. *Neuroreport, 17*(4), 357-360.

Nold, G., Hartig, J., Hinz, S., & Rossa, H. (2008). Klassen mit bilingualem Sachfachunterricht. Englisch als Arbeitssprache. Beltz.

Otoya, M. T. (1987). A study of personal memories of bilinguals: The role of culture and language in memory encoding and recall. Unpublished doctoral dissertation, Harvard University.

Paap, K. R., Johnson, H. A., & Sawi, O. (2014). Are bilingual advantages dependent upon specific tasks or specific bilingual experiences?. *Journal of Cognitive Psychology, 26*(6), 615-639.

Park, M. (1999). Linguistic influence on numerical development. *The Mathematics Educator, 10*(1).

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

Pladevall-Ballester, E., & Vallbona, A. (2016). CLIL in minimal input contexts: A longitudinal study of primary school learners' receptive skills. *System*, *58*, 37-48.

Quick Placement Test. (2001). Oxford: Oxford University Press.

Roquet, H., & Pérez-Vidal, C. (2015). Do productive skills improve in content and language integrated learning contexts? The case of writing. Applied Linguistics, 38(4), 489-511.

Roussel, S., Joulia, D., Tricot, A., & Sweller, J. (2017). Learning subject content through a foreign language should not ignore human cognitive architecture: A cognitive load theory approach. *Learning and Instruction*, 52, 69-79.

Saalbach, H., & Imai, M. (2007). Scope of linguistic influence: Does a classifier system alter object concepts? *Journal of Experimental Psychology: General, 136*(3), 485.

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.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). E-Prime v1. 1. *Pittsburgh, PA: Psychology Software Tools Inc.*

Schrauf, R. W., & Rubin, D. C. (1998). Bilingual autobiographical memory in older adult immigrants: A test of cognitive explanations of the reminiscence bump and the linguistic encoding of memories. *Journal of Memory and Language*, *39*(3), 437-457.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: a bilingual training study. *Cognition*, *78*(1).

Stanescu-Cosson, R., Pinel, P., van de Moortele, P. F., Le Bihan, D., Cohen, L., & Dehaene, S. (2000). Understanding dissociations in dyscalculia: a brain imaging study of the impact of number size on the cerebral networks for exact and approximate calculation. *Brain*, *123*(11), 2240-2255.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37-76). Academic Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *8*0(5), 352.

Van Rinsveld, A., Brunner, M., Landerl, K., Schiltz, C., & Ugen, S. (2015). The relation between language and arithmetic in bilinguals: insights from different stages of language acquisition. *Frontiers in psychology, 6.*

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.

Watts T. W., Duncan G. J., Siegler R. S., Davis-Kean P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352–360.

WHO (2008). Inequalities in young people health. Health behavior in school-aged children. International report from the 2005/2006 survey.

WHO (2016). Growing up unequal: gender and socioeconomic differences in young people's health and well-being. International report from the 2013/2014 survey. WHO: Regional office for Europe.

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

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. Wiley Interdisciplinary Reviews: *Cognitive Science, 2*(3), 253-265.

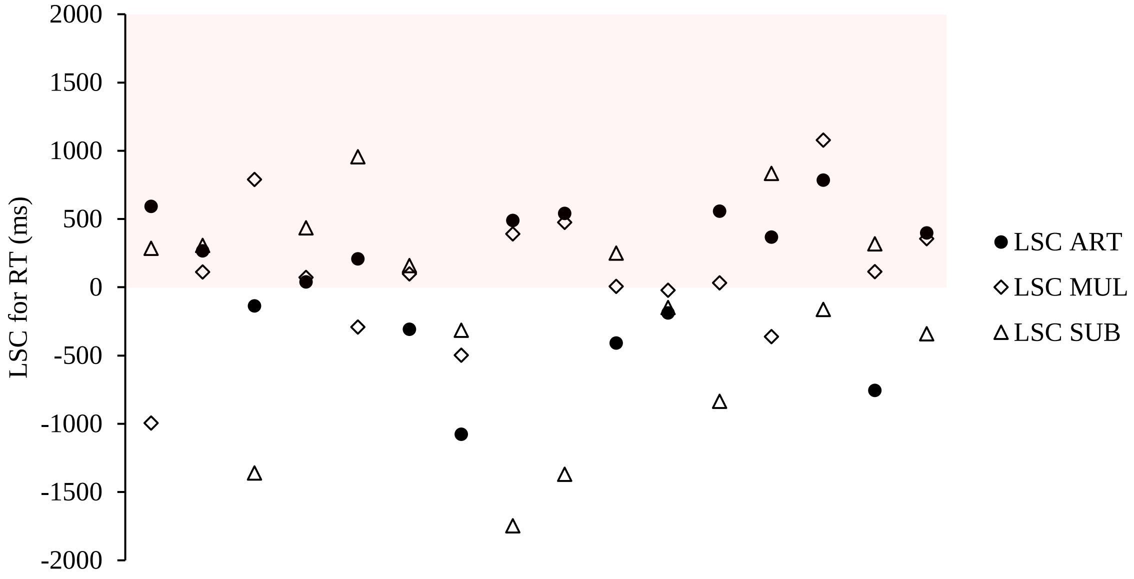
Zaunbauer, A. C. M., Bonerad, E. M., & Möller, J. (2005). Muttersprachliches Leseverständnis immersiv unterrichteter Kinder 1 Dieser Beitrag wurde von DH Rost akzeptiert. *Zeitschrift für Pädagogische Psychologie, 19*(4), 263-265.

Zaunbauer, A. C. M., & Möller, J. (2009). Schulleistungsentwicklung immersiv unterrichteter Grundschüler in den ersten zwei Schuljahren. *Psychologie in Erziehung und Unterricht*, (1), 30-45.

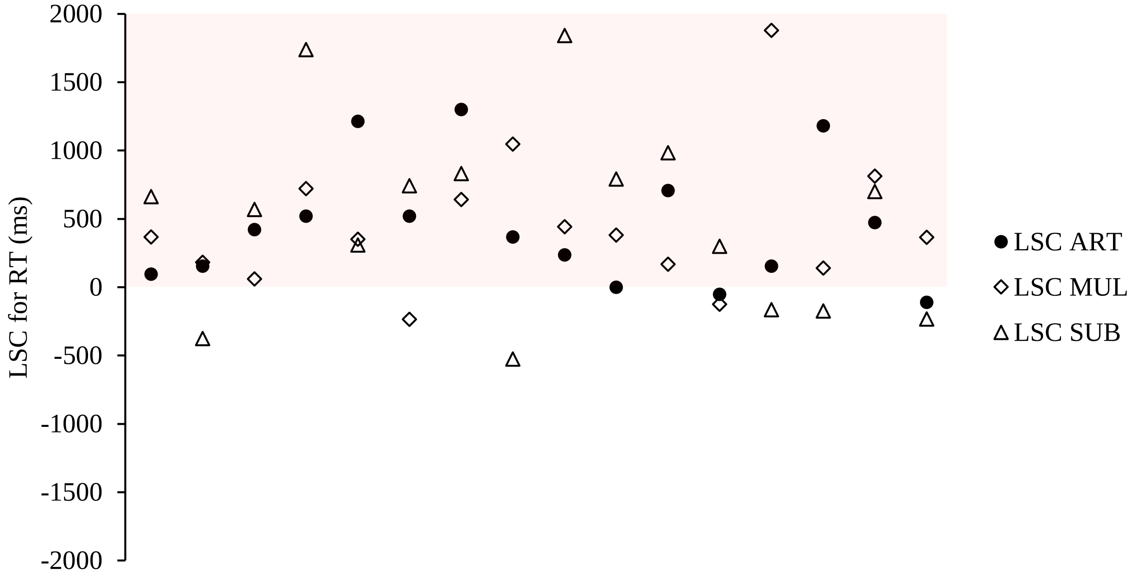
Supplementary Material

LSC for RT in Study 1 for all three tasks, separated for the two test groups. Individual scores in the shaded area represent LSC.

Test order A

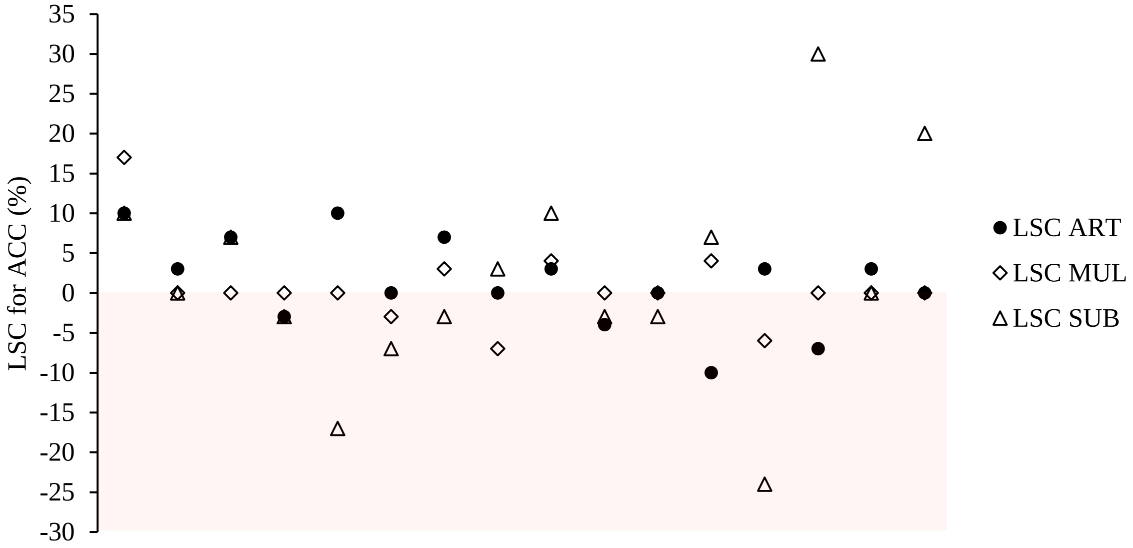


Test order B

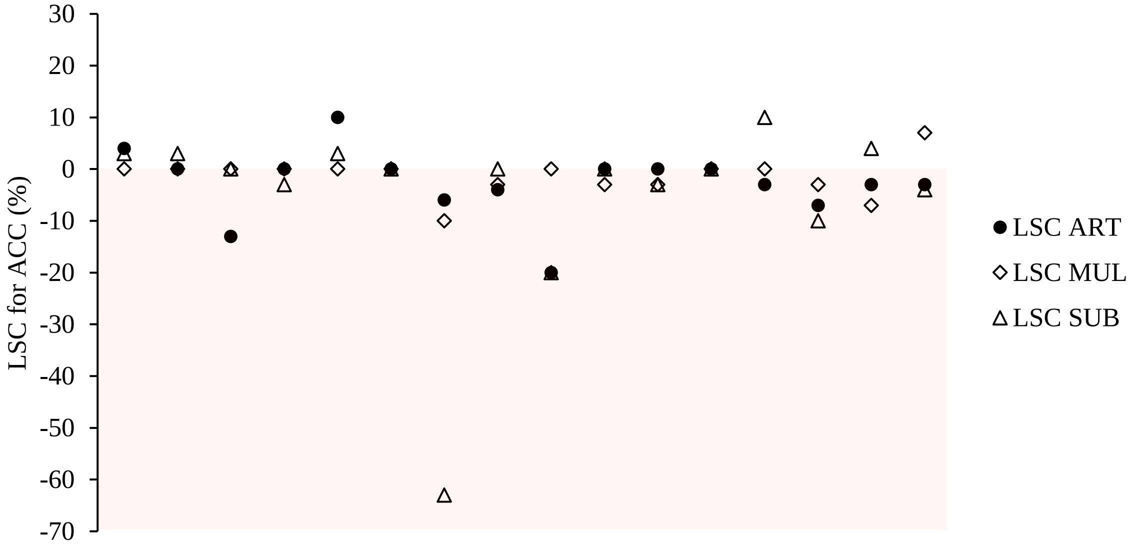


LSC for ACC in Study 1 for all three tasks, separated for the two test groups. Individual scores in the shaded area represent LSC.

Test order A

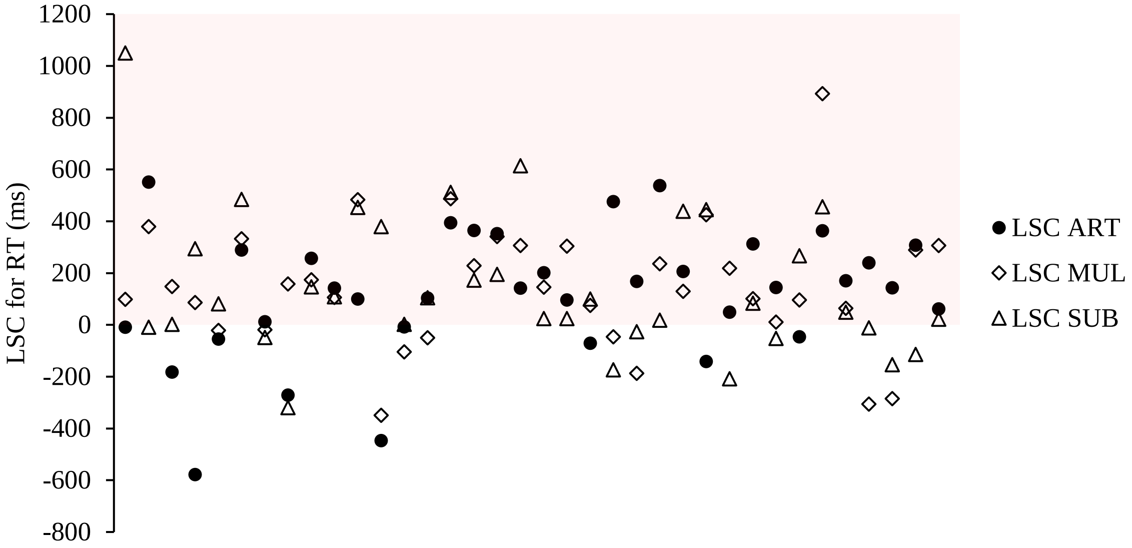


Test order B

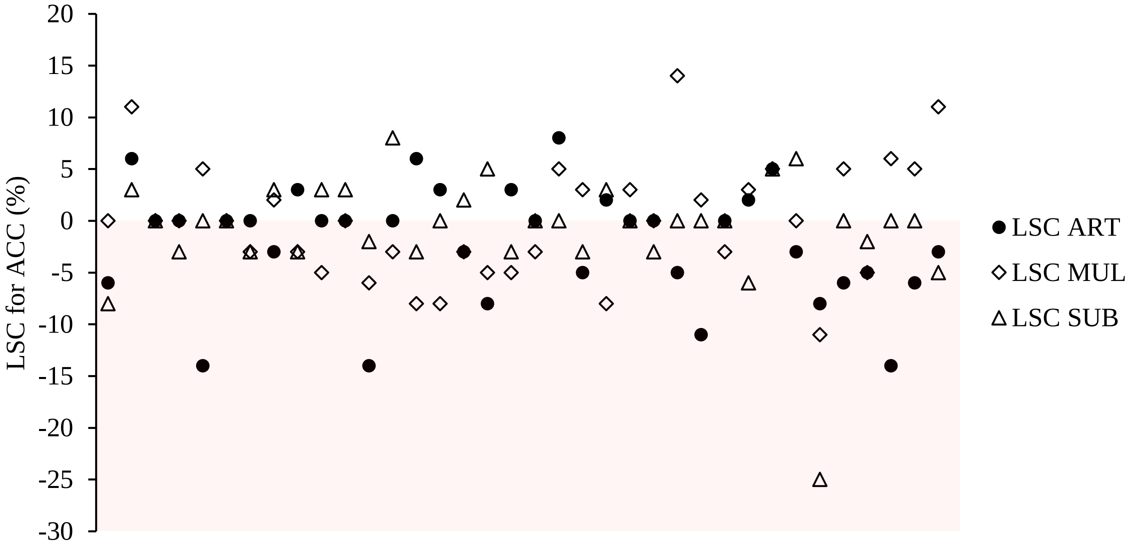


LSC for RT and ACC in Study 2 for all three tasks. Individual scores in the shaded area represent LSC.

LSC for RT

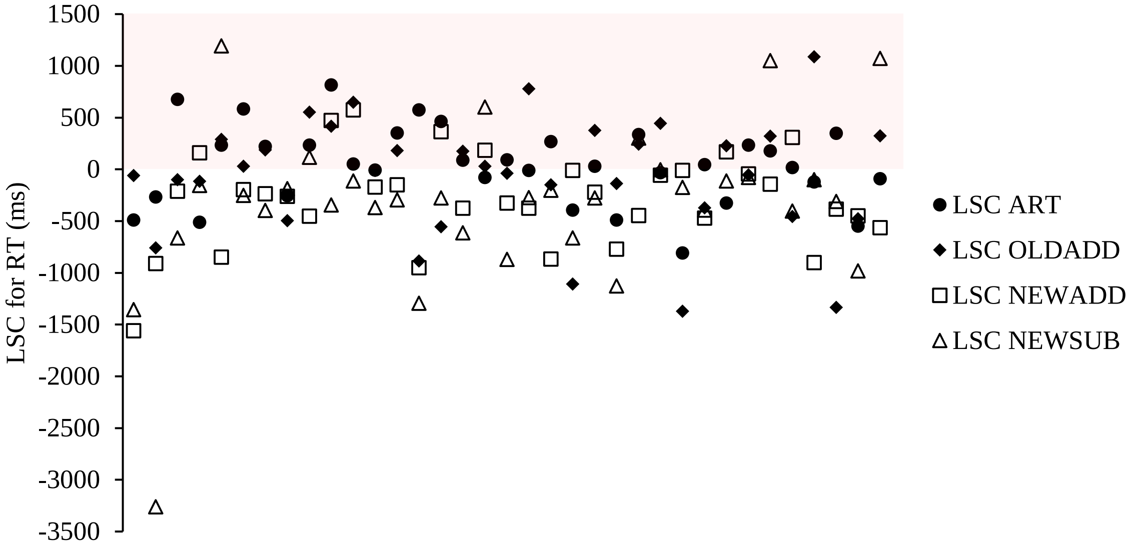


LSC for ACC

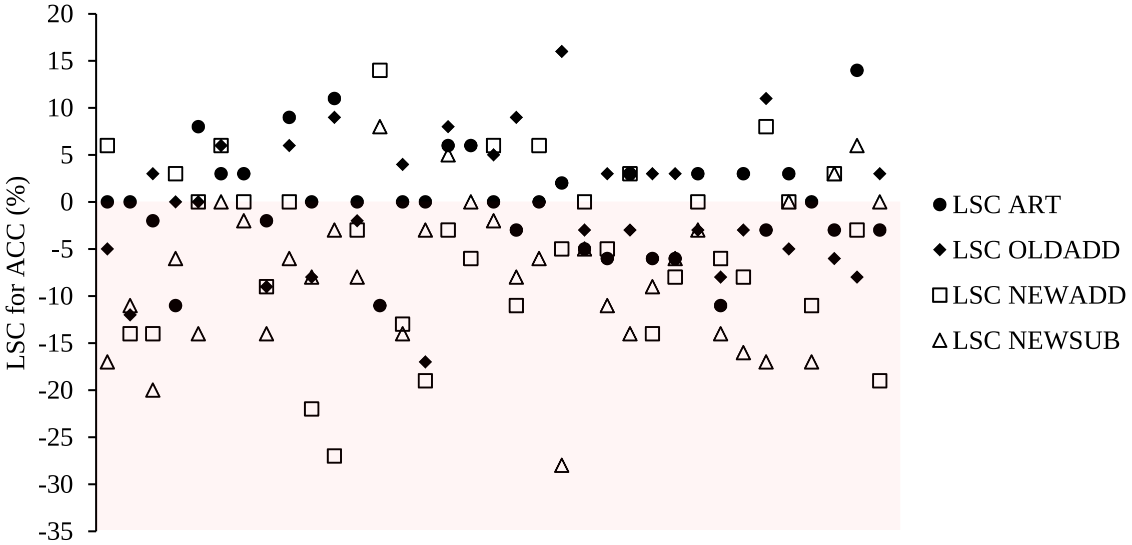


LSC for RT and ACC in Study 3 for all four tasks. Individual scores in the shaded area represent LSC.

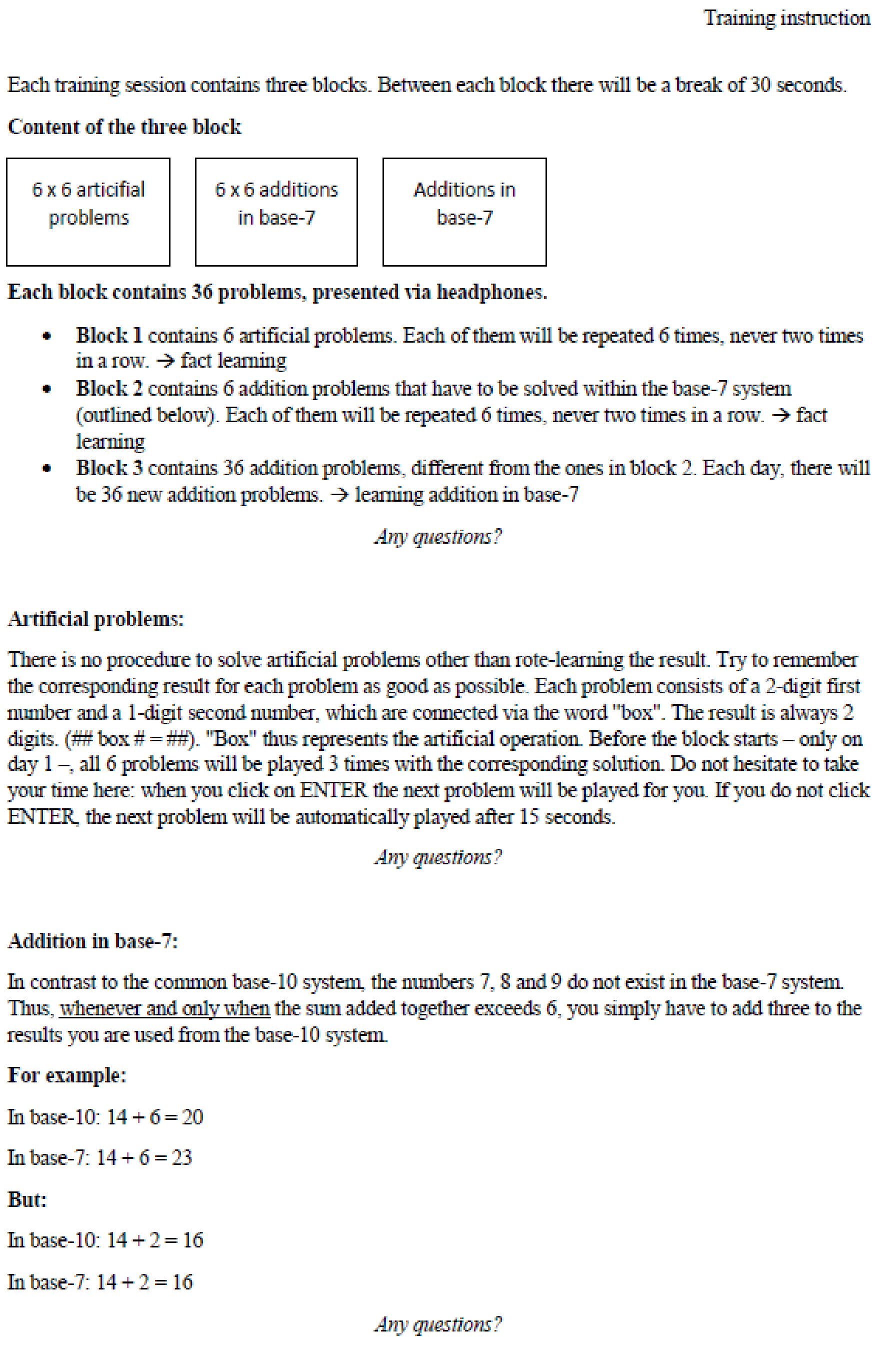
LSC for RT



LSC for ACC



Instructional guide for Study 3



Training instruction

Each training session contains three blocks. Between each block there will be a break of 30 seconds.

Content of the three block

6 x 6 articifial problems	6 x 6 additions in base-7	Additions in base-7

Each block contains 36 problems, presented via headphones.

- **Block 1** contains 6 artificial problems. Each of them will be repeated 6 times, never two times in a row. → fact learning
- **Block 2** contains 6 addition problems that have to be solved within the base-7 system (outlined below). Each of them will be repeated 6 times, never two times in a row. → fact learning
- **Block 3** contains 36 addition problems, different from the ones in block 2. Each day, there will be 36 new addition problems. → learning addition in base-7

Any questions?

Artificial problems:

There is no procedure to solve artificial problems other than rote-learning the result. Try to remember the corresponding result for each problem as good as possible. Each problem consists of a 2-digit first number and a 1-digit second number, which are connected via the word "box". The result is always 2 digits. (## box # = ##). "Box" thus represents the artificial operation. Before the block starts – only on day 1 –, all 6 problems will be played 3 times with the corresponding solution. Do not hesitate to take your time here: when you click on ENTER the next problem will be played for you. If you do not click ENTER, the next problem will be automatically played after 15 seconds.

Any questions?

Addition in base-7:

In contrast to the common base-10 system, the numbers 7, 8 and 9 do not exist in the base-7 system. Thus, whenever and only when the sum added together exceeds 6, you simply have to add three to the results you are used from the base-10 system.

For example:

In base-10: 14 + 6 = 20

In base-7: 14 + 6 = 23

But:

In base-10: 14 + 2 = 16

In base-7: 14 + 2 = 16

Any questions?