

Achievement on Arithmetic in Bilingual Primary Schools: Do the language and  
context matter?

Master thesis

Theme: Cognition

Charles Fayt; 3888789

Supervisor: Drs. M.M.H. Schaars

Second Assessor: Dr. A.J. van Tilborg

Master Educational Sciences, Utrecht University

Date: 10-06-2019

Word count (exclusive abstract, references and appendices): 8096.

### Abstract

Research argued for the development of cognitive and linguistic advantages in bilingual education, but arithmetic performances remained unclear. In less linguistic arithmetic, raw sums, bilinguals' performances should be similar or worse to monolinguals', while bilinguals may perform better, on linguistic arithmetic math problems. Furthermore, the task language could impact performance. The question is: *How do language and context affect student's performance on arithmetic operations in primary bilingual education?* Quasi-Experimental research was conducted involving four auditory verification tasks with two manipulated factors: raw sums versus math problems and Dutch (L1) versus English (L2). Tasks measured accuracy and reaction times. A bilingual experimental group (n=37) and monolingual control group (n=39) of children in grade 2 of primary school were tested. Results suggested that bilinguals performed better in Dutch for raw sums and without differences in reaction times. Furthermore, bilinguals tended to perform less well but faster on math problems. Finally, bilinguals tended to perform better in Dutch than monolinguals, and no performance was less well than monolingual ones. These results confirm previous research where arithmetic performances in the dominant language are better and do not support a bilingual advantage for solving problems. Discussing results, suggestions are made for future research and design of arithmetic classes.

*Keywords:* bilingual education, arithmetic, math problems, linguistic context, Content and Language Integrated Learning

**Achievement on Arithmetic in bilingual primary schools: Do the language and context matter?**

There are more and more bilingual schools (Baker, 2011). These schools are linked to a crucial question: How does bilingual education impact student's cognition and achievement? Bilingualism has often been negatively perceived. In Western societies, a belief claimed that bilingualism led to cognitive troubles because of the apparent mixing of languages bilinguals display (Abdelilah-Bauer, 2015). Nowadays, bilingualism is still mentioned as a risk for student's achievement (Jenniskens, et al. 2018; de Graaff, 03-10-2013; Le Pichon, 2013). However, research argues for a neutral or even positive influence of bilingualism on cognition (Abdelilah-Bauer, 2015). These proposed advantages concern general cognitive and linguistic abilities (Abdelilah-Bauer, 2015; Demont, 2001). Limited attention was spent on school subjects where language is less prominent, such as arithmetic (Demont, 2001). If there is evidence that bilinguals are better in language and higher thinking processes (Abdelilah-Bauer, 2015), there is no clear evidence that completion of bilingual education (BE) enables students to compute arithmetic operations with ease in two languages (Driessen et al., 2016, Demont, 2001). Equal performance in both languages is one of the goals of BE though (Driessen et al., 2016).

Furthermore, BE lacks clear design guidelines (Driessen, et al. 2016). BE can be designed using *content and language integrated learning* (CLIL) as in the Netherlands (Nuffic, n.d.) and Europe (Abdelilah-Bauer, 2015) among other approaches (see Stryker & Leaver, 1997). In CLIL, school subjects are integrated into the process of language acquisition by teaching some subjects in the target language and the rest in the mother language (Nikula; 2016; Mehisto, Marsh, & Frigols, 2008). BE results in *consequential bilingualism*; the first language (L1) was acquired first, and the acquisition of the second

language (L2) started when the child was not older than seven years (Abdelilah-Bauer, 2015). In the Netherlands, children can follow CLIL from the age of 4 or 5. The CLIL guidelines are broad and interpretative. This resulted in various forms of CLIL education across schools (Nikula, 2016).

Theoretically, this research is a step in developing a unified theory of arithmetic in BE where all independently previously tested factors are taken within one research. Practically, this research contributes to the designing of arithmetic classes in BE by investigating the role language plays in it. Language may influence it in two ways: 1. By the language in which the operation is performed and 2. By the extent to which language is incorporated into arithmetic (e.g., *raw sums* like  $2+2$ , where language plays a small role) or an *arithmetic word problem* (referred in the following as *problem*) where language (processing) plays a larger role). So, this study is meant to give insights on performances on raw sums and problems, how this performance is influenced by the first and second language and how this relates to monolinguals' performances. In the following, previous research on the field of BE is discussed.

### **General cognitive advantages**

Research proposed that bilinguals have developed general cognitive advantages impacting their executive functioning. Bilinguals would continuously select one language and suppress one (Abdelilah-Bauer; 2015; Demont, 2001, Zhang, 2018). This cognitive process, inhibition, is proposed as extra-developed in bilinguals, who can transfer this skill to other situations such as school (Durlík, Szewczyk, Muszyński & Wodniecka, 2016). It would favour higher concentration (Durlík, et al., 2016; García, 2011) and lead to faster and more accurate responses in various tasks compared to monolinguals (Costa, Hernandez, & Sebastian-Galles, 2008). Next, bilinguals are assumed to perform better than monolinguals in tasks involving a high level of interaction between task elements (i.e. higher intrinsic

cognitive load) (Blom et al., 2014). This suggests a better use of- or more cognitive resources (Blom et al., 2014). Concluding, bilinguals may outperform monolinguals thanks to extra-developed cognitive abilities.

### **Task-specific advantages**

**Language-related tasks.** Attention is spent on bilinguals' performances on language-related tasks. Indeed, linguistic processing abilities could affect performances on problems. Gollan, Montoya and Werner (2002) followed by, Gollan, Montoya, Fennema-Notestine and Morris (2005) argued for a negative impact of bilingualism on lexical selection and processing. Demont (2001) found, at the contrary, that children in BE demonstrated a higher accuracy and ease on tests measuring linguistic capabilities compared to monolinguals. Laurent and Martinot (2010) found comparable results and investigated the amount of exposure needed. Children in BE only started to outperform (in both languages) their monolingual counterparts on phonological awareness after four years of BE. Nicoladis and Jiang (2018) discovered that even with more limited vocabularies in both languages (compared to monolinguals), children in BE produced stories with as many different words as monolinguals. This suggests a different but more efficient word selection strategy than monolinguals that may also positively impact language processing. Anderson, Chung-Fat-Yim, Bellana, Luk and Bialystok (2018) found in fMRI-data evidence for a different and more efficient language processing method in bilinguals compared to monolinguals, arguably positively impacting the working-memory. Concluding, bilinguals seem better at processing linguistic information when compared to monolinguals, which might give them an advantage in solving problems. In the next section, earlier findings on arithmetic are discussed.

**Arithmetic.** Demont (2001) tested arithmetic abilities for a possible bilingual advantage because she believed that language contributes to arithmetic performances. She expected a better performance since bilingual children (BC) performed better on language

tasks. Demont conducted oral tasks in L1 (French) with 23 BE educated children, asking them to solve raw sums, while at the bilingual school, arithmetic was taught in their L2 (German). The children in BE needed more time, hesitated more, gave more incorrect responses than children in monolingual education. Demont suggested the following: BC did perform less well on this task because the task language (their L1) was not the language used to teach them arithmetic (their L2). However, she hypothesized that the disadvantage would be compensated or even reversed in both languages if items would be problems. BC would benefit from their better linguistic processing abilities for processing information.

Demont (2001)'s idea is against Marsh and Maki (1976), who orally tested BC instructed in L1, in their two languages for raw sums and problems. Findings were a lower accuracy on problems in general enlarged when presented in L2. Marsh and Maki (1976) concluded that the child's dominant language (L1) negatively impacts his arithmetic performance in L2. The computation process is slower because of a translation process (Marsh & Maki, 1976). McClain and Huang (1982) replicated Marsh and Maki (1976) on raw sums only, but in a situation in which math was taught in the L2. They found a difference favouring L2 instead, suggesting that the language of instruction determines the quality of performance on raw sums. This could explain Demont's results since she tested the children in their less L1 for arithmetic while arithmetic was taught in L2. Similarly, research by Gelman and Butterworth (2005) on L2 learners outside of BE had shown that the language used to teach arithmetic (their L1), is the language used to process numbers and solve arithmetic operations when orally asked, independently of the proficiency in L2. As a result, arithmetic operations in the L2 are computed slower due to the translation to L1 preceding the computation of the arithmetic operation, followed by a second translation back to the L2 and that process also increases the chances for mistakes (Gelman & Butterworth, 2005). Children in Demont (2001) were taught arithmetic in their L2, and according to Gelman and

Butterworth (2005), they should instinctively compute arithmetic operations in their L2, resulting in worse and delayed performance in their L1.

Furthermore, differences between monolinguals and bilinguals as in Demont (2001) are not always found. Jenniskens, et al. (2018) did not find differences between the performance of children following BE and children following monolingual education on standardized tests in L1 on solving problems and reading comprehension, which rejects the idea of a bilingual advantage but excludes the vision of BE as detrimental.

Supporting Demont (2001)'s idea, Swanson, Kong and Petcu (2018) proposed that the positive impact of bilingualism on the working memory has a positive impact on the computation of raw sums and problems. Furthermore, the most proficient BC outperformed less proficient children (Swanson, Kong & Pectu, 2018). This result supports a positive influence of bilingualism, but the study lacks the comparison with monolinguals to add a deeper layer for interpretation. Van Rinsveld, Schiltz, Brunner, Landerl, and Ugen (2016) found that trilingual children performed better and faster on raw sums in the language of instruction. However, the difference between the languages can be reduced by incorporating the raw sums in a linguistic context requiring a semantic judgement activating a specific language. They asked BC to judge the truthfulness of a story which formed the context of the problem. All stimuli were presented visually. The presence of a linguistic context only positively impacts the performance in the languages not used for arithmetic instruction. The differences were smaller than expected though, arguably because of the task itself was cognitively more demanding than the control task which only involved raw sums. Van Rinsveld et al., (2016) support the idea of a bilingual advantage on arithmetic as hypothesized by Demont (2001). The experimental task involved next to arithmetic, also semantic and pragmatic reasoning (Van Rinsveld et al., 2016). Another such language-based difference was found by Le Pichon and Kambel (2016). They investigated the impact of the language used in

arithmetic classes in monolingual education, but in a bilingual society. The results were that the language used for education (L2) generated more correct responses for simple arithmetic (less verbal context) while the language not used for education (L1) generated more correct responses for tasks (problems) that relied on linguistic context. Le Pichon and Kambel's interpretation is that problems are more authentic than artificial raw sums. So, for real-life situations as problems, children use their L1 and they use their L2 for raw sums because they have learned to do it in L2. Finally, Canavesio (2013) found that children in BE performed better (accuracy) in all languages involved when compared to children in monolingual education. These findings are against Demont (2001) results of no better performance in both languages compared to monolinguals on raw sums. Canavesio used an auditory verification task where students had to determine if a presented number was a correct response to a raw sum, in various languages. Canavesio saw the role of language in this as enough to suspect a bilingual advantage, which was confirmed and in line with Demont (2001)'s original expectation. Canavesio also recommended investigation of contexts in which language is more prominent such as a problem. However, the absence of differences in reaction times (RT) between conditions is possibly the consequence of additions that were too simple for the participants' level. Canavesio mentioned accounting for automatization as well. Only arithmetic operations that are automatized can generate a meaningful interpretation of RT. Summarizing, the impact of the language used for arithmetic operations computed by students in BE is unclear, as the influence of the embedding these operations into a linguistic context.

### **This research**

In this research, students following BE are referred to as bilinguals and students in monolingual education are referred to as monolinguals, independently of their personal linguistic situation. BE aims similar academic achievement for students in L1 and L2 (Driessen et al., 2016). Concluding from above, there is a gap in the field of arithmetic in



BE and no consensus has been reached on the factors impacting achievement and their precise influence. The role that language plays on performance (i.e. accuracy and RT) remains unclear, as the influence of a linguistic context for arithmetic (problems) and a less linguistic context for arithmetic (raw sums). Finally, the relation to achievement in monolingual education remains. Consequently, this research addressed the question:

*How do language and context affect student's performance on arithmetic operations in primary BE?*, which led to six hypotheses in line with Demont (2001) speculation and complemented with other literature. The hypotheses are organized in three categories for which expectations are formulated from the theory both within the bilingual group as compared to monolinguals:

#### **Raw sums.**

**Hypothesis 1.** BC perform less well and slower on raw sums when the task language is not the instruction language (English) for arithmetic compared to their performance in the class language for arithmetic (Dutch) ( Demont, 2001; Gelman & Butterworth, 2005; Le Pichon & Kambel, 2016; Marsh & Maki, 1976; McClain & Huang; 1982;).

**Hypothesis 2.** BC perform less well and slower on raw sums in the language not used for arithmetic classes (English) in comparison to monolingual children's (MC) in Dutch. (Demont, 2001).

**Hypothesis 3.** BC perform similarly on raw sums, compared to MC in the condition in which the task language is the instruction language for arithmetic (Dutch) (Demont, 2001; Van Rinsveld et al., 2016).

#### **Raw sums against problems.**

**Hypothesis 4.** BC perform better and faster on problems in both languages compared to their performances in both languages on raw sums (Demont, 2001; Swanson, et al., 2018).

**Problems.**

**Hypothesis 5.** BC perform better and faster on problems in both languages than MC (Demont, 2001; Le Pichon & Kambel, 2016; Van Rinsveld, et al., 2016).

**Hypothesis 6.** BC perform better on problems in their dominant language (Dutch) (Le Pichon & Kambel, 2016).

**Method****Research design**

A quasi-experimental quantitative research design was used (Verhoeven, & van Baal, 2011). A manipulation between several conditions was done and results were analyzed within and between experimental groups (Verhoeven & van Baal, 2011). Participants were not assigned randomly to a condition (Verhoeven & van Baal, 2011).

**Participants**

The experimental group consisted of 37 students in grade 2 (around 8 years old) in primary BE, taught in Dutch (L1; also instruction language for arithmetic) and English (L2). They came from two classes at the same school. The control group consisted of 39 students in grade 2 in the Dutch primary monolingual education. They came from two schools with similar populations. The number of participants is equivalent to previous research (Canavesio, 2013; Demont, 2001; Gollan, et al., 2002; Marsh & Maki, 1976; Gollan, et al., 2005; Le Pichon & Kambel; 2016; Nicoladis & Jiang, 2018). The reason for choosing grade 2 is motivated by: 1. schools offering BE in the Netherlands have implemented BE up to grade 2 (Nuffic, n.d.) and 2. four years of exposure are needed before the influence of BE could be tested (Laurent & Martinot, 2010, Abdelilah-Bauer, 2015).


## Materials




Participants took part in two auditory verification tasks. They differ in the extent to which language was incorporated. One condition involved raw sums and the other one, problems. Each operation consisted of two one-digit numbers with a result up to 16 reached by addition, subtraction, multiplication, which is line with what Demont (2001) and Le Pichon and Kambel (2016) did with participants of the same age. The teacher's opinion about the task level of each class was asked three days prior to testing using example items that were similar but not included in the actual tasks to check whether children should have automatized these operations (Canavesio, 2013; Le Pichon & Kambel, 2016, Blom & Unsworth, 2010). Both tasks had a Dutch and English version. Auditory stimuli made it possible to investigate the influence of the task language. The stimuli were recordings of a simultaneous bilingual (English-Dutch) female speaker recorded and edited in Praat, version 6.0.52 (Boersma & Weenink, 2019). Children prefer listening to female voices (Blom & Unsworth, 2010). The tasks were programmed in Zep version 1.16 (Veenker, 2018). The tasks were digital and measured accuracy and RT in milliseconds (ms.) (UiL-OTS, n.d.). Participants had 15 s. to make a choice, otherwise, Zep moved to the next item. The tasks were also grounded in the reaction times paradigm where differences in RT are associated with differences in level of difficulty and heavier cognitive processing (Baayen & Milin, 2010). Faster RT are also associated with the language used for computation (Canavesio, 2013).

**Materials for raw sums.** Canavesio (2013) did similar research. Canavesio (2013)'s raw sums task is here replicated. It is grounded in the paradigm of LeFevre (1998), a methodology to measure arithmetic performance and linguistic interferences (Canavesio, 2013; LeFevre, 1998). The child must decide whether an auditory presented arithmetic operation (e.g.,  $3+2$ ) is associated with the correct auditory target (e.g., 5). The decision is

made by pressing a yes or no button. A trial sequence can be found in Table 1. Canavesio (2013) used 24 trials consisting of 12 trials requiring a yes-answer (four additions, four subtractions, four multiplications) and 12 trials a no-answer (four additions, four subtractions, four multiplications) (see Appendix A for Dutch and Appendix B for English trials). Consequently, the same was done here. The task lasted about three minutes and was preceded by six feedbacked practice items. Furthermore, the experiment should uncover the language in which students process arithmetic operations (Canavesio, 2013). The amount of time between the operation and the target was maintained constant across all trials.

Table 1.


*Sequence of a Trial with Raw Sums (  indicates that the information was only auditory presented with a white screen).*




Step	On display/heard 	Time (in ms)
Fixation cross	+	1000
Operation	2+2	
		
Operation to Target interval		1500
Target	“four”	
		
Respos interval		Max. 15000
Intertrial interval		1500

**Materials for problems.** The task for problems took the design of the previous tasks but involved problems instead. The child was auditory presented with an arithmetic problem like in Le Pichon and Kambel (2016) who tested a comparable group. The problems involved different contexts to ensure that participants paid attention to what was said (Le Pichon & Kambel, 2016; Blom & Unsworth, 2010). A voice announced a possible answer and the child had to decide as fast as possible if that answer was correct or not by pressing on a yes- or no-button. A trial sequence can be found in Table 2. Le Pichon and Kambel (2016) used ten trials because more trials would cause a fatigue effect. Accordingly, twelve trials were used in the

current design to ensure an equal spreading between the kind of operation. Six requested a yes-answer (two additions, two subtractions, two multiplications) and six a no-answer (two additions, two subtractions, two multiplications) (see Appendix C for Dutch and Appendix D for English trials). The amount of time between the operation and the target was maintained constant across all trials. The tasks lasted about four minutes and started with three feedbacked practice trials. The number of items differed from the raw sums task because problems are inherently cognitively heavier and more exhaustive (Le Pichon & Kambel, 2016, Van Rinsveld, et al., 2016).

Table 2.

*Sequence of a Trial with Problems (  indicates that the information was only auditory presented with a white screen).*

Step	On display/heard 	Time (in ms)
Fixation cross	+	1000
Operation 	‘There are two children in the bus. Two more children step into the bus. How many children are now on the bus?’	
Operation to Target interval		1500
Target 	“Four”	
Respon interval		Max. 15000
Intertrial interval		1500

## Procedure

In collaboration with the school, an information letter requesting prior active consent was sent to parents. That letter contained information regarding the purpose of the study, experimental tasks, the way the data was processed and reported. Parents filled in two questions regarding the child's language use.

Children were tested individually, each on their experimental laptop running Windows 10. Zep is robust against hardware differences but require the same exploitation system to allow comparison between data collected with different devices (UiL-OTS, n.d.). Children were instructed in the task language that they would play a game in which they should decide as fast as possible whether the computer correctly solved a trial. Furthermore, they were told that if they did know the answer or did not hear the item correctly, they should wait until to the next trial. They wore headphones. After the feedbacked practice trials, there was room for questions. Then, the child took the actual experiment and did not get any feedback on answers anymore.

Children in BE took the two tasks, but twice (one for each language) and the children in monolingual education participated only in the Dutch tasks. MC were not tested in English because there is clear evidence that monolingual perform less well and slower in their L2. Furthermore, their English may be too limited to understand the contexts of the problems. Finally, this data was not needed to test the research hypotheses. The performance of bilinguals in English is compared to their performance in Dutch, which also became an experimental variable when compared to the performance in Dutch of monolinguals.

Children did the tasks on different moments (at least two hours between tests) to avoid a learning or fatigue effect (Blom & Unsworth, 2010). The trials order was randomly determined by Zep each time an experiment was run. This avoided a learning or sequential

effect within the task, and to avoid the same effects, participants did the tasks in different orders (Blom & Unsworth, 2010).

### **Analysis**

The analysis of accuracy (score on tasks) was done in SPSS version 24 using independent sample t-tests and paired-sample t-tests. As recommended by Baayen and Milin (2010), the analysis of RT was done with a multi-level modelling analysis in R version 3.5.3 (R Core Team, 2019) using CRAN-Packages: lme4 (Bates, Maechler, Bolker, Walker, 2015), LSmeans (Length, 2015), ggplot2 (Wickham, 2016) and influence.ME (Nieuwenhuis, te Grotenhuis & Pelzer, 2012). The guidelines in Winter (2013) for the multilevel modelling analyses recommend the inclusion of factors that are theoretically based or influencing participant's accuracy. Computation of effect size values was done in the tool by Becker (2000) and interpreted using Ellis (2009). Statistical significance was  $p < .05$ , which is usual in social sciences (Vogt & Johnson, 2011).

The six hypotheses display complexity, specificity and not derived or integrated within a single coherent theory (i.e. specific patterns of variables may lead to specific performances but are not yet coherently related). The hypotheses are meant for detection of patterns that may lead to a theory. In such a context, it can be better to test each hypothesis independently (Althouse, 2016; Gelman, Hill and Yajima, 2012; Perneger, 1999; Rothman, 1990, Savitz & Olshan, 1998) Here, multiple t-tests and multilevel modelling analyses were done. This approach has the disadvantage of the multiple comparisons problem (see Miller, 1981). However, in such a context, the chance of type I error would remain low (McDonald, 2014; Rothman, 1990). Furthermore, alternatives may result in a loss of information or specificity leading to overlooking possible influences (Althouse, 2016; Gelman, et al., 2012; Rothman, 1990; Perneger, 1990). To minimize the probability of type I errors, the Benjamini–Hochberg procedure was applied (see Benjamini & Hochberg, 1995) using the tool by McDonald (2015)

with the recommended false discovery rate of 0.25. The results can be seen in Appendix F Table 1 for accuracy and Appendix F Table 2 for RT. Even if the hypotheses were not independent, the Benjamini–Hochberg procedure remains a meaningful addition (Benjamini, & Yekutieli, 2001; McDonald, 2014).

## Results

**Preliminary analyses and descriptive statistics.** One bilingual participant was removed because the participant only pressed on the yes-button in all four tasks. Possible confounding factors were investigated on the accuracy data, and no such factor was found (see Appendix E). Consequently, no such factor was included in the multilevel analyses for RT (Winter, 2013). Table 3 shows the descriptive statistics for each group and for each task in which they were involved, and Table 4 displays the Pearson correlations between tasks from which can be seen that only the Dutch raw sums and the Dutch problems barely correlate ( $r=.259, p=.024$ ).

Table 3.

*Mean (sd.) for Scores on Raw Sums (max. score =24) and Problems (max. score =12) and for RT in ms for Raw Sums and Problems by Bilingual and Monolingual Education.*

	Score Raw sums in Dutch	RT Raw sums in Dutch	Score Raw sums in English	RT raw sums in English	Score Problems in Dutch	RT Problems in Dutch	Score Problems in English	RT Problems in English
Children in bilingual education (n=37)	21.5 (2.7)	3124 (2500)	18.7 (3.8)	3084 (2440)	8.6 (1.9)	2438 (1751)	7.5 (2.1)	2111 (1633)
Children in monolingual education (n=39)	20.1 (3.1)	3140 (2412)			7.7 (2)	2969 (2307)		



Table 4.

*Pearson's Correlations of the Scores between Tasks (number of observations of tasks in Dutch= 74, in English=36).*

Tasks	Raw sums in Dutch	Raw sums in English	Problems in Dutch	Problems in English
Raw sums in Dutch	-	-	-	-
Raw sums in English	.290	-	-	-
Problems in Dutch	.259*	-.261	-	-
Problems in English	-.090	-.006	-.194	-

\* $p < .005$

**Investigation of the assumptions.** The accuracy data did satisfy all assumptions of the tests used, excepted normality according to Shapiro-Wilk tests. The removal of outliers and the log-transformation did not solve the issue. F-tests (Lindman, 1974) and T-tests (Sawilowsky & Blair, 1992) are known for being robust for lack of normality when  $N > 20$ . To guaranty the appropriateness of the conclusions, the outcomes of parametric tests were compared to the non-parametric tests.

The RT data met all assumptions of multi-level modelling analyses, expected normality. Nevertheless, Winter (2013) claimed that normality is the least important assumption and can be overlooked. Following Winter (2013), this problem was solved by log-transformation. Assumptions were also investigated and met with the final model. The inclusion of random intercepts for the items and for the participant is standard (Winter, 2013). The results of this inclusion are only mentioned in the following if they differ from what would normally be expected. In the case of effect with log-values, the analysis was redone using raw data for favoring the interpretation (Winter, 2013).

**Raw sums.** The first hypothesis states that BC would perform less well and slower on raw sums when the task language is not the instruction language for arithmetic (English). A

two-tailed paired-sample t-test suggested that BC performed less well on raw sums in English compared to Dutch (instruction language), ( $t(36)=-4.433, p<.001$ )<sup>1</sup>, with an effect size of Cohen's  $d = 0.84$ , and effect-size correlation  $r=.39$ , suggesting a rather medium effect (Ellis, 2009). The Task language as a fixed effect influencing RT did not improve a model with random intercepts for item and participant ( $\chi^2(1)=.091, p=.763$ ). This suggests that the task language did not influence RT of bilingual students on raw sums in Dutch and in English.

The second hypothesis states that BC would perform less well and slower on raw sums in the language not used for arithmetic instruction (English) compared to MC. A two-tailed independent sample-test suggested that BC's performance in English did not differ from MC's performance in Dutch ( $t(74) = -1.780, p=.08$ )<sup>2</sup>. The type of school (bilingual or monolingual) as a fixed effect influencing RT did not improve a model with random intercepts for item and participant ( $\chi^2(1)=.029, p=.865$ ). This suggests that the Type of school did not influence RT of bilingual's performance on English raw sums compared to monolingual's performance on Dutch raw sums.

The third hypothesis states that BC would perform similarly on raw sums in Dutch (instruction language) compared to MC. An independent sample t-test suggested that MC performed less well on raw sums in Dutch than children following BE ( $t(74)=-2,206, p=.030$ )<sup>3</sup> with an effect size of Cohen's  $d =-.51$ , and effect-size correlation  $r=.24$ , suggesting a rather medium effect (Ellis, 2009). The type of school (bilingual or monolingual) as a fixed effect influencing RT did not improve a model with random intercepts for item and participant ( $\chi^2(1)<.001, p=.984$ ). This suggests that the Type of school did not influence RT of BC and MC on Dutch raw sums.

---

<sup>1</sup> A non-parametric Wilcoxon Signed ranks test does not change anything to these outcomes.

<sup>2</sup> A non-parametric Mann-Whitney test does not change this conclusion..

<sup>3</sup> A non-parametric Mann-withney test does not change this conclusion..

**Raw sums against problems.** The fourth hypothesis states that BC would perform better and faster on problems in both languages compared to their performance in both languages on raw sums. To allow comparison, the scores on twelve of the problem tasks were doubled to get scores on 24. A paired sample t-test suggested that BC performed better on raw sums than on problems in Dutch ( $t(36)=5.817, p<.001$ )<sup>4</sup>with an effect size of Cohen's  $d= 1.33$  and  $r=.55$  suggesting a rather large effect (Ellis, 2009). The difference in accuracy is around 17%. Task (raw sum and problems) as a fixed effect influencing RT improved the model with participant and items as random intercepts ( $\chi^2(1) =7.81, p=.005$ ). This suggests that the type of task influenced the bilingual's RT on the tasks with raw sums and problems in Dutch. Adding task as random slope also improved the model, ( $\chi^2(4) =59.205, p<.001$ ), which optimize its quality (Winter, 2013). The task with raw sums increased RT by about 716 ms  $\pm$  207 (standard errors) compared to RT for problems. The summary of the final model can be found in Table 5.

---

<sup>4</sup> A non-parametric Wilcoxon test points to the same conclusion

Table 5.

*Summary of the linear mixed-effects regression with raw data (conclusion are similar to similar to analysis with log-values). Number of observations = 1308. Estimates are in ms.*

Random	Name	Variance	Std. deviation	N
intercepts				
Participant	(Intercept)	5.350e+05	731.4294	37
	Task (raw sums)	3.984e+05	631.2244	
Item	(Intercept)	1.230e-02	0.1109	36
	Task (raw sums)	4.616e+05	679.4321	

Fixed effect	Name	Estimate	Std. error	T-value
Task	(Intercept)	2439.06	151.37	16.113
	Task raw Sum	715.66	206.61	3.464

The influence of both dimensions of task (raw sum and problems) was further investigated using a LSmeans pairwise test on the log-values, confirming the influence of Tasks on RT ( $t(51.2)=-2.604$ ,  $p=.012$ ) with an effect size of Cohen's  $d=0.32$ , and effect-size correlation  $r=.16$ , suggesting a rather small effect (Ellis, 2009). Summarizing, BC were faster to make a choice for problems than for raw sums in Dutch.

The same way of analyzing was also applied to the data of the English raw sums and problems. A paired sample t-test suggested that bilinguals' performance on English raw sums

was superior to their performance on English problems ( $t(36)=3.935, p<.001$ )<sup>5</sup> with an effect size of Cohen's  $d=.92$  and  $r=.42$  suggesting a rather large effect (Ellis, 2009). The difference in accuracy is again around 17%. Task (raw sum and problems) as a fixed effect influencing RT improved the model with item and participants as random intercepts ( $\chi^2(1)=21.296, p<.001$ ). This suggests that the type of task influenced RT on bilingual's performance on the tasks with English raw sums and problems. Adding random slopes for Task also impacted RT, ( $\chi^2(4)=39.728, p<.001$ ), giving it a higher acceptability (Winter, 2013). The task with raw sums increased RT by about 998 ms.  $\pm$  221 (standard errors) compared to RT for problems. The summary can be found in Table 6.

---

<sup>5</sup> This difference in performance is also found when doing a non-parametric Wilcoxon test.

Table 6.

*Summary of the linear mixed-effects regression with raw data (conclusion are similar to analysis with log-values). Number of observations = 1316. Estimates are in ms.*

Random	Name	Variance	Std. deviation	N
intercepts				
Participant	(Intercept)	2.050e+05	452.7970	37
	Task (raw sums)	8.499e+05	921.9235	
Item	(Intercept)	1.643e-02	0.1282	36
	Task (raw sums)	3.362e+05	579.8498	

Fixed effect	Name	Estimate	Std. error	T-value
Task	(Intercept)	2110.99	116.28	18.15
	Task raw Sum	998.14	221.32	4.51

The influence of both dimensions of Task (raw sum and problems) was further investigated using a LSmeans pairwise test, confirming the influence of task on RT ( $t(44.6) = -4.428, p < .001$ ) with an effect size of Cohen's  $d = 0.47$ , and effect-size correlation  $r = 0.23$ , suggesting a rather small effect (Ellis, 2009). Summarizing, BC were faster to choose a response for problems than for raw sums in English.

**Problems.** The fifth hypothesis states that BC would perform better and faster on problems, in both languages than MC. A two-tailed independent simple t-test suggested that

BC did not perform better on problems in Dutch than MC ( $t(74)=1.844, p=.069$ )<sup>6</sup>. However, the data suggest an outcome that might be called marginally significant (see Pritschet, Powell & Horne, 2016). The Type of school (bilingual or monolingual) as a fixed effect influencing RT is an improvement of the model with participants and item as random intercepts ( $\chi^2(1) = 4.104, p=.043$ ). This suggests that the Type of school influenced RT. MC increased RT with about  $556 \text{ ms} \pm 224$  (standard error). Adding the Type of school as a random slope did not improve the model ( $p=.484$ ), which do not exclude the effect of Type of school (see Winter, 2013). As a result, the previous model with Type of school as fixed effect and participant and item as random intercepts was further used. The influence of both dimensions of Type of school (bilingual vs. monolingual) was further investigated using a LSmeans pairwise test on log-values confirming the influence of Type of school on RT ( $t(75.3)=-2.053, p=.04$ ) with an effect size of Cohen's  $d = -.47$ , and effect-size correlation  $r=.23$ , suggesting a small effect (Ellis, 2009). Summarizing, BC were faster to choose a solution for problems in Dutch than MC. The summary of the final model is in Table 7.

---

<sup>6</sup> A non-parametric Mann-withney test does not change anything to this conclusion.

Table 7.

*Summary of the linear mixed-effects regression with raw data (conclusions are similar to similar to analysis with log-values). Number of observations = 890. Estimates are in ms.*

Random	Name	Variance	Std. deviation	N
intercepts				
Participant	(Intercept)	653673	808.5	76
Item	(Intercept)	79724	282.4	12

Fixed effect	Name	Estimate	Std. error	T-value
Type of school	(Intercept)	2443.0	179.9	13.579
	Monolingual	555.6	224.2	2.478

The same pattern of analyses was also done for the bilingual's performances on English problems, compared to the monolingual's performance on Dutch problems. A two-tailed independent simple t-test suggests that BC did not perform better on problems in English than MC did on problems in Dutch ( $t(74)=.547, p=.586$ )<sup>7</sup>. The Type of school (bilingual or monolingual) as a fixed effect influencing RT improved the model with participants and items as random intercepts ( $\chi^2(1)=12.831, p<.001$ ). This suggests that the Type of school influenced RT of children. MC increased RT with about  $880 \text{ ms} \pm 225$  (standard error). After this, a new model, using log-values, was computed adding Type of

<sup>7</sup> A non-parametric Mann-withney test does not change anything to this conclusion.



school as random slope. This did not improve the model, ( $p=.997$ ), which is not a major concern for the effect of Type of school (see Winter, 2013). So, further investigation was conducted using the previous model with participant and item as random intercepts and the type of school as a fixed effect. The influence of both dimensions of Type of school (bilingual vs. monolingual) were analysed using a LSmeans pairwise test on log-values confirming the influence of type of school on RT ( $t(51.2)=-3.802$ ,  $p<.001$ ) with an effect size of Cohen's  $d=-1.06$ , and effect-size correlation  $r=.47$ , suggesting a rather large effect (Ellis, 2009). Summarizing, BC were faster to make a choice for problems for English problems than MC for Dutch problems. The summary of the model used can be found in Table 8.

Table 8.

*Summary of the linear mixed-effects regression with raw data (conclusions are similar to similar to analysis with log-values). Number of observations = 895. Estimates are in ms.*

Random	Name	Variance	Std. deviation	N
intercepts				
Participant	(Intercept)	475774	689.8	76
Item	(Intercept)	59564	244.1	24

Fixed effect	Name	Estimate	Std. error	T-value
Type of school	(Intercept)	2114.0	160.5	13.172
	Monolingual	880.7	225.1	3.913

The sixth and last hypothesis states that BC would perform better and faster in their dominant language (Dutch) for problems. A two-tailed paired sample t-test suggested that bilinguals performed better on problems in Dutch than on problems in English ( $t(36)=2.113$ ,  $p=.042$ )<sup>8</sup> with a Cohen's  $d=.54$  and  $r=.26$  suggesting a rather medium effect (Ellis, 2009). Item as a random intercept did not improve the model testing RT ( $\chi^2(1)=2.0673, p=.150$ ), arguably because of the small number of items ( $n=12$ ). Consequently, further investigation was done using the model with a random intercept for participants only. Adding the Task (problems in Dutch vs. problems in English) as a fixed effect influencing RT significantly improved the model ( $\chi^2(1)=8.345, p=.004$ ). Adding a random slope for Task also significantly improved the model ( $\chi^2(2)=30.879, p<.001$ ). This suggests that the problems

<sup>8</sup> A non-parametric Wilcoxon Signed Ranks Test lead to the same conclusion.

in English decreased RT by about  $330 \text{ ms} \pm 157$  (standard errors). The model's summary can be found in Table 8. The effect of the task language was further investigated using a LSmeans pairwise test on the log-values suggesting no effect ( $p=.085$ ). However, the same analysis performed with the raw RT suggested an effect ( $t(37.2)= 2.104, p=.042$ ) whereby RT for English problems are lower than for Dutch problems with an effect size of Cohen's  $d= .55$  and  $r=.26$  suggesting a rather medium effect (Ellis, 2009). That is the interpretation that is retained for the rest of this paper. The summary of that model can be found in Table 9. So, BC were faster to make a choice for Dutch than for English problems.

Table 9.

*Summary of the linear mixed-effects regression with raw data (conclusions are similar to similar to analysis with log-values). Number of observations = 877. Estimates are in ms.*

Random	Name	Variance	Std. deviation	N
intercepts				
Participant	(Intercept)	647642	804.8	37
	Problems in English	508289	712.9	24
Fixed effects	Name	Estimate	Std. error	T-value
Task	(Intercept)	2443.50	151.61	16.117
	Problems in English	-330.28	156.95	-2.104

According to the Benjamini-Hochberg procedure, none of the conclusions above should be rejected (see Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001), which lowered the costly risk of type I error due to multiple comparisons.

### Conclusion & Discussion

A brief summary of the outcomes follows. This research looked at the influence of language and the type of education on performance and RT on raw sums, raw sums versus problems and problems alone. The results suggest a tendency of bilinguals to perform better on raw sums and in Dutch (L1 and instruction language) compared to problems, monolinguals and L2. Bilinguals seemed to outperform monolinguals on Dutch raw sums, with a marginally significant effect ( $p=.069$ ) for Dutch problems. RT did not differ within and between groups for raw sums. RT decreased (along with accuracy) for problems with RT on English problems being the lowest. Overall bilinguals never performed less well than monolinguals on any tasks, on any language.

**Raw sums.** The data suggest that children in BE performed less well on raw sums in English than on raw sums in Dutch, but these performances are never less well than what monolinguals did. This in line with previous findings which have suggested that the language in which arithmetic is taught, is the dominant language for computing raw sums and lead to higher performances than in the other language (Gelman & Butterworth, 2005; Le Pichon and Kambel, 2016; Marsh & Maki, 1976; McClain & Huang, 1982). This interpretation is further in line with Demont (2001)'s results that the performance in the not instruction language for arithmetic is worse than in the instruction language in BE. Furthermore, Gelman and Butterworth (2005) did their research with late L2 learners. These results are also in line with Le Pichon and Kambel (2016) who found that computing raw sums is done better in the school's language for arithmetic. The results suggest that even early exposition and intensive use of another language does not impact one's ability to compute arithmetic operation in that

language. The instruction language is the favored language for arithmetic. The results on raw sums in English, which were worse compared to bilinguals in Dutch, also contradict Canavesio (2013) where evidence was found for general better bilinguals' performances (accuracy) compared to monolinguals on raw sums independently of the language. However, the lack of difference in RT would rather suggest even ease across conditions and groups. This is against findings in Marsh & Maki, (1976) and McClain and Huang (1982) which suggested a delay in RT for raw sums because bilinguals have to translate it into their dominant language or language of instruction, compute it and translate it back to the task language. As in Canavesio (2013), no such differences were found in RT between and within groups.

**Raw sums against problems.** The results contradict another part of the idea presented in Demont (2001). She argued that bilinguals should do better on problems than on raw sums because of their better abilities for processing linguistic content. The data here suggests the contrary. Bilinguals were doing systematically better on raw sums than on problems in both languages. This is in line with Marsh & Maki (1976) but against Le Pichon and Kambel (2016) who found that bilinguals performed problems in at least one language than monolinguals. The findings are not in line in line with Van Rinsveld et al. (2016) who found positive effects on performance using problems in the language not used for arithmetic instruction. BC were always faster to answer problems than raw sums, which is line with Demont (2001). Baayen and Milin (2010) would suggest that such a difference suggest that problems are experienced as less cognitively demanding. Due to the decrease in accuracy, this interpretation is not very plausible.

**Problems.** The results on accuracy are against Demont (2001) idea who argued for overall superior bilinguals' performance on problems compared to monolinguals. Indeed, the data suggests that bilinguals did not perform better on problems in Dutch or English than MC

did in Dutch. The absence of differences would rather be in line with Jenniskens, et al. (2018) who did not find differences between monolinguals and bilinguals on solving problems. Arguably, the apparent better bilinguals' linguistic processing skills (Anderson, et al., 2018) may not transfer to problems. Nevertheless, there is a marginally significant effect ( $p=.069$ ) in Dutch which might become significant with a larger group. Due to the large sample needed, it would suggest a rather small effect. However, BC were faster on problems in Dutch and English than MC. Baayen and Milin (2010) would interpret this as that problems are less cognitively demanding for bilinguals, which is then in line with Demont (2001)'s supposition. The difference in speed is also in line with Swanson, et al., (2018) who were mostly researching the situation in the dominant language, which was shared with the monolinguals. However, the accuracy data is against Swanson et al., (2018) who found a better performance of bilinguals. Results regarding problems are fully in line with Van Rinsveld et al. (2016) and Pichon and Kambel (2016). In both studies, the dominant or daily used language led to better performances on problems than in the other language(s) present. Finally, the differences in RT on problems within the bilingual group between English and Dutch, would suggest, according to Canavesio (2013) that children are computing problems in English since the shortest RT are associated with the language of computation and that higher RT are associated with translation procedures. However, this is not plausible because of the outcomes proposed by the accuracy data.

**Overall.** Bilinguals never performed less well than MC. So even their performances on English problems which were the lowest did not differ from the ones of monolinguals on Dutch problems. This conclusion is against Demont (2001) and Canavesio (2013) who found differences between monolinguals and bilinguals. Finally, even if there are effects between conditions that are more or less large, the actual absolute differences are relatively small. For instance, the largest difference in RT is smaller than one second which would be unnoticeable

without software able to measure it. The differences in accuracy are more important but stay under 20%.

**Limitations.** A first limitation is the difficulty of interpretation of RT for problems (i.e. larger ease) compared to accuracy (i.e. more difficulties). This could be explained because children appeared to have favored guessing instead of skipping items of which they did not know the answer. The binary choice in the tasks might have made tempting to guess. Unfortunately, Zep does not offer the possibility for a participant to skip an item (UiL-OTS, n.d.). Zep can only go to the next item after a specific amount of time when no response is given. Guessing lowers RT and decreases accuracy (Baayen & Milin, 2010). So, any conclusion that was derived from RT should be taken with the necessary caution. However, guessing is also a sign of higher level of experienced difficulty (Baayen & Milin, 2010). This would make RT data and its relation the accuracy data more logical and interpretable; namely problems as more difficult for bilinguals than raw sums. English tasks and to some extent Dutch problems seemed to have been impacted by guessing. However, when considering the means on each task, there all above 50 %; the chance level which suggests that children were not only guessing (Baayen & Milin, 2010). The pattern could also be explained by differences in the extent of automatization, which would then have increased RT of the not yet automatized raw sums (Canavesio, 2013) while guessing would have decreased RT on problems.

Another related limitation is that the design lacks depth. The research leaves the question open in which language children compute arithmetic operations, which is in research unclear (Abdelilah-Bauer, 2015). For raw sums, RT suggested that bilinguals can do it in both, but not according to the accuracy data. For problems, bilinguals seemed to not need any translation to L1 since there were the fastest on English problems, but the accuracy data points out to the contrary. Finally, children's reactions while doing the tasks suggested that

they prefer to compute arithmetic operations in Dutch and arguably used Dutch in English tasks. Indeed, participants were sometimes commenting English tasks aloud but only in Dutch and unhappily reacted to an English task.

Another limitation is the choices made when designing items to find an appropriate level of difficulty. Most sums were relatively easy for children in grade 2 to take automatization into account (Canavesio, 2013) and arguably led to the problems around normality. However, too easy sums were problematic in Canavesio (2013) leading to no differences in RT. Furthermore, as Le Pichon and Kambel (2016) mentioned, the operations in raw sums or problems should not be too much challenging for the children because this might include confounding factors.

That bilinguals were doing better on Dutch raw sums than monolinguals forms a limitation. This difference in performance might be due to a general higher arithmetic level of the bilingual students which may have no relation to BE. This suggests caution regarding results involving the control group. However, Canavesio (2013) found overall better bilinguals' performance (in all languages) on raw sums than monolinguals which was interpreted as an effect of BE on arithmetic performances in general.

Finally, this research is mainly made of the testing of different hypotheses derived from literature, and it is the first time that research tries to account for all variables impacting arithmetic in BE. However, this had made the research very complex and in the limited time available, choices had to be made for the analysis, which were questionable but inevitable. There was indeed no single analysis method possible due to the nature of the hypotheses and the collected data and without loss in content or specificity. However, risks for type I errors are limited because most of the results are in line with dominant views in previous research and Benjamini-Hochberg procedure did not impact the results. Unfortunately, this research does not optimally contribute to the understanding of the bigger picture, nor lead to a working



theory. Indeed, the mismatch of the accuracy and RT data makes the developing of a single coherent theory impossible. However, the research points to a better performance in L1 and on raw sums within BE.

**Further research.** A first suggestion for research and a known problem in the literature is the lack of developmental data. Very few studies follow students in bilingual education throughout their development. This might provide interesting insights about whether effects do increase, decrease, appear, disappear with time and determining the impact on these effects when monolinguals become proficient in a second language. Laurent and Martinot (2010) already identified that at least three or four years of BE are needed before any effects can be found but further development is unknown. The marginally significant tendency to perform better on problems in Dutch compared to monolinguals might become significant when children are further in their BE-education.

Abdelilah-Bauer (2015), Demont (2001), and Van Rinsveld et al. (2016) suspected that the number of languages a child speak might enlarge the possible positive effects on cognition; the more languages spoken, the larger the positive effects on cognition would be. It might be interesting to research the same topic in children where a majority also speak a third language to see whether such enlarged advantages exist.

Another suggestion for further research would be to replicate this research at bilingual schools where children get arithmetic in L2, like in primary BE in Belgium, Italy and France (Abdelilah-Bauer, 2015). In the Netherlands, bilingual primary schools have chosen for arithmetic in L1. This questions the influence of the language of instruction on effects that were found or not.

As pointed out earlier, the design lacks depth. It does not provide a very clear image of the events in a participant's head, and the research does not successfully lead to a working theory regarding arithmetic performances in BE. One way to achieve to add depth would be

using the think-aloud protocol methodology, where participants are asked to describe their thinking while doing a task (Jääskeläinen, 2010). This method is very time consuming, though. Further research should focus on developing and testing a unified and coherent theory taking the following parameters into account to predict achievement (with consequent outcomes for accuracy and RT) on arithmetic: monolingual versus bilingual, L1 versus L2 and raw sums versus problems. A first step might be to first focus on the situation within BE before including comparisons with monolinguals.

**Practical implications.** This research offers valuable information for improving BE in the Netherlands. The findings suggest an imperfect transfer of arithmetic skills to L2 when taught in L1. This is against one of the goals of BE, namely that children perform in both languages with even ease (Driessen et al., 2016). So, for instance, the inclusion of arithmetic activities in English next to Dutch might be an alternative. Such a course design, in which the same subject matter is alternatively given in L1 and L2, is another approach concurring with CLIL (see Stryker & Leaver, 1997). Another option would be to teach arithmetic in L2 as it is widely done (Abdelilah-Bauer, 2015). The L1-dominant environment is assumed to favor transfer from L2 to L1 (Abdelilah-Bauer, 2015; Stryker & Leaver, 1997). Furthermore, this research contributes to defeat the belief that BE may negatively impact student's achievement and cognition since bilinguals never performed less well than monolinguals (Abdelilah-Bauer, 2015; de Graff, 03-10-2013; Jenniskens et al., 2018; Le Pichon, 2013).

## References

- Abdelilah-Bauer, B. (2015). *Le défi des enfants bilingues*. Paris, France : La découverte.
- Althouse, A. D. (2016). Adjust for multiple comparisons? It's not that simple. *The Annals of thoracic surgery*, 101(5), 1644-1645. doi: <https://doi.org/10.1016/j.athoracsur.2015.11.024>
- Anderson, J. A., Chung-Fat-Yim, A., Bellana, B., Luk, G., & Bialystok, E. (2018). Language and cognitive control networks in bilinguals and monolinguals. *Neuropsychologia*, 117, 352-363. doi: <https://doi.org/10.1016/j.neuropsychologia.2018.06.023>
- Baker, C. (2011). *Foundations of bilingual education and bilingualism* (Vol. 79). Bristol, UK: Multilingual matters.
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.
- Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12-28. doi: <https://doi.org/10.21500/20112084.807>
- Becker, L.A. (2000). Effect size calculators [Apparatus]. Colorado Springs, USA: University of Colorado. Retrieved from <https://www.uccs.edu/lbecker/>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 57, 289-300. doi: <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4), 1165-1188. doi: 10.1214/aos/1013699998

Blom, E., Kuntay, A. C., Messer, M., Verhagen, J., & Leseman, P. (2014). The benefits of being bilingual: Working memory in bilingual Turkish Dutch child. *Journal of Experimental Child Psychology*, 128, 105-119. doi:

<https://doi.org/10.1016/j.jecp.2014.06.007>

Blom, E., & Unsworth, S. (Eds.). (2010). *Experimental methods in language acquisition research* (Vol. 27). Amsterdam, the Netherlands: John Benjamins Publishing.

Boersma, P.P.G & Weenink, D.J.M. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.52. Retrieved from <http://www.praat.org/>

Canavesio, M. L. (2013). *Bilingual Education in the Primary School: Curriculum Study and Experimental Research on Language of Acquisition Effects in the Arithmetic Facts*. (Doctoral dissertation). University of Trento, Trento, Italy.

Costa, A., Hernandez, M., & Sebastian-Galles, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, 106, 59–86. doi :

<https://doi.org/10.1016/j.cognition.2006.12.013>

Demont, E. (2001). Contribution de l'apprentissage précoce d'une deuxième langue au développement de la conscience linguistique et à l'apprentissage de la lecture. *International journal of psychology*, 36(4), 274-285. doi:

<https://doi.org/10.1080/00207590042000137>

Ellis, P. D. (2009). Thresholds for Interpreting Effect Sizes. Retrieved from

[http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds\\_for\\_interpreting\\_effect\\_sizes2.html](http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds_for_interpreting_effect_sizes2.html)

de Graaff, R. (03-10-2013). *Taal om te leren: Didactiek en opbrengsten van tweetalig onderwijs*. Utrecht, the Netherlands: Universiteit Utrecht.

- Driessen, G., Krikhaar, E., de Graaff, H.C.J., Unsworth, S., Leest, B., Coppens, K. & Wierenga, J. (2016). *Evaluatie pilot Tweetalig Primair Onderwijs - Startmeting schooljaar 2014/15* (publieksversie). Nijmegen, The Netherlands: ITS Nijmegen.
- Durlak, J., Szewczyk, J., Muszyński, M., & Wodniecka, Z. (2016). Interference and Inhibition in Bilingual Language Comprehension: Evidence from Polish-English Interlingual Homographs. *PloS one*, 11(3), e0151430. doi: <https://doi.org/10.1371/journal.pone.0151430>
- García, O. (2011). *Bilingual education in the 21st century: A global perspective*. Hoboken, USA: John Wiley & Sons.
- Gelman, R., & Butterworth, B. (2005). Number and language: how are they related?. *Trends in cognitive sciences*, 9(1), 6-10. doi: <https://doi.org/10.1016/j.tics.2004.11.004>
- Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189-211. doi: <https://doi.org/10.1080/19345747.2011.618213>
- Gollan, T.H., Montoya, R.I., & Werner (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology*, 16, 562-576. doi: <https://doi.org/10.1016/j.tics.2004.11.004>
- Gollan, T.H., Montoya, R.I., Fennema-Notestine, C., & Morris, S.K. (2005). Bilingualism affects picture naming but not picture classification. *Memory and Cognition*, 33, 1220-1234. doi: <http://dx.doi.org/10.1037/0894-4105.16.4.562>
- Jääskeläinen, R. (2010). Think-aloud protocol. In Y. Gambier & L. van Doorslaer, (Eds.) *Handbook of translation studies* (371-374), Amsterdam, the Netherlands: John Benjamin.

- Jenniskens, T., Leest, B., Wolbers, M., Krikhaar, E., Teunissen, C., de Graaff, H., Unsworth, S. & Coppens, K. (2018). *Evaluatie pilot Tweektalig Primair Onderwijs - Vervolgmeting schooljaar 2016/17 - publieksversie*. Nijmegen, the Netherlands: KBA Nijmegen.
- Length, R. (2016). Least-Squares Means: The R Package LSmeans. *Journal of Statistical Software*, 69(1), 1-33. doi: [doi:10.18637/jss.v069.i01](https://doi.org/10.18637/jss.v069.i01)
- Laurent, A., & Martinot, C. (2010). Bilingualism and phonological awareness: the case of bilingual (French–Occitan) children. *Reading and Writing*, 23(3-4), 435-452. doi: <https://doi.org/10.1007/s11145-009-9209-3>
- Le Pichon, E.M.M. & Kambel, E-R. (2016). Challenges of mathematics education in a multilingual post-colonial context - The case of Suriname. In Z. Babaci-Wilhite (Eds.), *Human rights in language and STEM education* (pp. 221-240). Rotterdam, the Netherlands: Sense Publishers.
- Le Pichon, E.M.M. (2013). Handling plurilingualism in kindergarten and primary school. In W., Griebel; R., Heinisch; C., Kieferle; E. Röbe & A., Seifert, (Eds.), *Transition to School and Multilingualism – A Curriculum for Educational Professionals*. Hamburg, Germany: Verlag Dr. Kovac.
- Lindman, H. R. (1974). *Analysis of Variance in complex experimental design*. New York, USA: W.H. Freeman and Co.
- Marsh, L. G., & Maki, R. H. (1976). Efficiency of arithmetic operations in bilinguals as a function of language. *Memory & Cognition*, 4(4), 459-464. doi: <https://doi.org/10.3758/BF03213203>
- McClain, L., & Huang, J. Y. S. (1982). Speed of simple arithmetic in bilinguals. *Memory & Cognition*, 10(6), 591-596. doi: <https://doi.org/10.3758/BF03202441>

- McDonald, J.H. (2014). *Handbook of Biological Statistics* (3rd ed.). Baltimore, USA: Sparky House Publishing.
- McDonald, J.H. (2015). Multiple comparisons -Spreadsheet [Apparatus]. Retrieved from <http://www.biostathandbook.com/multiplecomparisons.html>
- Mehisto, P., Marsh, D., & Frigols, M. J. (2008). *Uncovering CLIL content and language integrated learning in bilingual and multilingual education*. London, United-Kingdom: Macmillan.
- Miller, R.G. (1981). *Simultaneous Statistical Inference* (2nd ed.). New-York, USA Springer Verlag.
- Nicoladis, E., & Jiang, Z. (2018). Language and cognitive predictors of lexical selection in storytelling for monolingual and sequential bilingual children. *Journal of Cognition and Development, 19*(4), 413-430. doi: <https://doi.org/10.1080/15248372.2018.1483370>
- Nieuwenhuis, R., te Grotenhuis, M. & Pelzer, B. (2012). influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *R Journal, 4*(2), 38-47.
- Nikula T. (2016). CLIL: A European Approach to Bilingual Education. In: N. Van Deusen-Scholl & S. May (eds) *Second and Foreign Language Education. Encyclopedia of Language and Education* (3rd ed.). Cham, Switzerland: Springer.
- Nuffic. (n.d). Tweetalige basisscholen. Retrieved from <https://www.nuffic.nl/onderwerpen/tweetalige-basisscholen/>
- R Core Team (2019). R: A language and environment for statistical computing. Vienna: Austria, R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL <http://www.R-project.org/> (version 3.5.3).

- Perneger, T. V. (1999). Adjusting for multiple testing in studies is less important than other concerns. *Bmj*, *318*(7193), 1288.
- Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*(7), 1036-1042. <https://doi.org/10.1177/0956797616645672>
- Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* *1*(1), 43-46. doi: <https://doi.org/10.1136/bmj.318.7193.1288a>
- Savitz, D. A., & Olshan, A. F. (1998). Describing data requires no adjustment for multiple comparisons: A reply from Savitz and Olshan. *American Journal of Epidemiology*, *147*(9), 813-814. doi: 10.1093/oxfordjournals.aje.a009532
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological bulletin*, *111*(2), 352. doi: <http://dx.doi.org/10.1037/0033-2909.111.2.352>
- Swanson, H. L., Kong, J., & Petcu, S. (2018). Math Difficulties and Working Memory Growth in English Language Learner Children: Does Bilingual Proficiency Play a Significant Role?. *Language, speech, and hearing services in schools*, *49*(3), 379-394. doi: [https://doi.org/10.1044/2018\\_LSHSS-17-0098](https://doi.org/10.1044/2018_LSHSS-17-0098)
- Stryker, S. B., & Leaver, B. L. (Eds.). (1997). *Content-based instruction in foreign language education: Models and methods*. Georgetown, Washington D.C.: Georgetown University Press.
- UiL-OTS, (n.d). The Zep programming language. Retrieved from <https://uilots-labs.wp.hum.uu.nl/software/zep/>



Van Rinsveld, A., Schiltz, C., Brunner, M., Landerl, K., & Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter?. *Learning and instruction*, 42, 72-82. doi:

<https://doi.org/10.1016/j.learninstruc.2016.01.003>

Veenker, T.J.G. (2018). The Zep Experiment Control Application (Version 1.16) [Windows 10]. Utrecht Institute of Linguistics OTS, Utrecht University. Available from

<http://www.hum.uu.nl/uilots/lab/zep/> .

Verhoeven, P. S., & van Baal, A. (2011). *Doing research: The hows and whys of applied research*. The Hague, the Netherlands: Eleven International Publishing.

Vogt, W. P., & Johnson, B. (2011). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. New York, USA: Sage publication.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New-York, USA: Springer-Verlag.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499*.

## Appendix A.

Materials and trials used for raw sums in Dutch.

Operation	Target	Correctness
2+2	4	Correct
3+4	7	Correct
9+3	12	Correct
7+6	13	Correct
9+1	3	Incorrect
1+4	7	Incorrect
9+5	15	Incorrect
8+7	11	Incorrect
9-2	7	Correct
6-4	2	Correct
9-6	3	Correct
8-7	1	Correct
8-2	10	Incorrect
7-1	3	Incorrect
8-4	6	Incorrect
3-2	8	Incorrect
4x1	4	Correct
5x1	5	Correct
3x3	9	Correct
2x7	14	Correct
2x1	4	Incorrect
5x2	7	Incorrect
4x4	11	Incorrect
2x6	15	Incorrect

## Appendix B.

Materials and trials used for the raw sums in English.

Operation	Target	Correctness
1+2	3	Correct
3+5	8	Correct
9+2	11	Correct
6+8	14	Correct
9+4	12	Incorrect
1+7	4	Incorrect
2+4	3	Incorrect
5+8	12	Incorrect
9-1	8	Correct
6-3	3	Correct
9-5	4	Correct
8-6	2	Correct
8-3	10	Incorrect
7-2	5	Incorrect
8-5	6	Incorrect
4-2	7	Incorrect
2x3	6	Correct
3x5	15	Correct
3x4	12	Correct
3x1	3	Correct
2x2	6	Incorrect
1x1	3	Incorrect
2x8	15	Incorrect
2x4	10	Incorrect

## Appendix C.

## Materials and trials used for problems in Dutch.

Problem	Target	Correctness
Er zijn 9 auto's op het parkeerterrein. Er komen nog 4 auto's bij. Hoeveel auto's zijn er nu op het parkeerterrein?	13	Correct
Mama geeft me 2 boterhammen voor de lunch. Ik vraag er nog 3 bij en mama maakt en geeft ze. Hoeveel boterhammen heb ik nu?	5	Correct
Op de luchthaven staan er 9 vliegtuigen. Er zijn er nog 3 bij geland. Hoeveel vliegtuigen zijn er nu op de luchthaven?	14	Incorrect
Ik heb 2 mooie goudvissen in mijn aquarium. Ik koop er nog 6 bij. Hoeveel vissen heb ik nu?	9	Incorrect
De supermarkt heeft nog 9 repen melkchocolade. Mijn vriend koopt er 3. Hoeveel repen melkchocolade kan ik nog kopen?	6	Correct
Een student had 4 opdrachten voor morgen gekregen. Nu heeft hij er al twee af. Hoeveel opdrachten moet hij nog maken?	2	Correct
Een leraar had 9 boeken. Hij is er 5 kwijt. Hoeveel boeken heeft de leraar nu?	6	Incorrect
De banketbakker had 9 appeltaarten gemaakt. Hij heeft er vandaag 8 verkocht. Hoeveel appeltaarten zijn er over?	3	Incorrect
De moeder van Anna geeft haar elke dag 2 gevulde koeken als tussendoortjes. Dit duurt 6 dagen lang. Hoeveel gevulde koeken heeft Anna na die tijd gehad?	12	Correct
Peter wint 3 superheldenplaatjes tijdens het spelen per dag. Dit gebeurt 3 dagen achter elkaar. Hoeveel plaatjes heeft hij in totaal gewonnen?	9	Correct
De appels worden per 2 verkocht in een zak. Ik koop 7 zakjes. Hoeveel appels heb ik nu?	11	Incorrect
Op het station liggen er 3 sporen, elk voor één trein bedoeld. Hoeveel treinen kunnen in het station?	1	Incorrect

## Appendix D.

## Materials and trials used for problems in English.

Problem	Target	Correctness
There are 8 busses at the bus station. 6 busses have just arrived at the station now. How many busses are there at the bus station?	14	Correct
Mom gave me 1 cookie for lunch. At school, the teacher gave me 2 more cookies. How many cookies do I have?	3	Correct
Peter had 9 toys. For his birthday, he got 2 new toys. How many toys does Peter have now?	10	Incorrect
My parents have 2 cats at home. They decided to adopt 4 new cats. How many cats do we have at home?	8	Incorrect
There are 9 waffles in the bakery. John takes 4 of them. How many waffles are left for William?	5	Correct
I had 7 chocolate bars. I gave 5 to my friends. How many do I have now?	2	Correct
The teacher had 6 pencils. He lost one of them. How many pencils do the teacher have?	4	Incorrect
I have made a bouquet of 8 flowers for mom. One fell on the ground. How many flowers are left in the bouquet?	6	Incorrect
I got 2 euros from my parents every day. It has lasted for 5 days. How many euros do I have now?	10	Correct
Anna won 2 Princesses cards. She did so for 2 days. How many Princesses cards did Anna win?	4	Correct
In a holiday house, there are 2 bedrooms. Each room has 4 single beds. How many people can sleep there?	6	Incorrect
3 friends like to play together. They all have 4 board games each. How many board games can they choose of?	10	Incorrect

## Appendix E.

## Investigation of possible cofounding factors.

Factor	Results <sup>9</sup>
Gender (n female= 29, n male= 48)	No differences according to an independent sample t-test (raw sum Dutch: $p=805$ , raw sum English: $p=.331$ , problems in Dutch: $p=.160$ , problems in English: $p=.670$ )
Teacher (n=14)	For English tasks: no differences between teachers according to a one-way ANOVA (raw sum: $p=174$ , problems: $p=.567$ ). Possible effect on Dutch tasks according to one-way ANOVA ( $F(3,72)= 2,81$ , $p=.045$ ), but not confirmed by a post-hoc Scheffé.
Influence of speaking a third language (other than Dutch or English) (n=15)	No evidence according to independent sample t-tests (raw sum Dutch: $p=.084$ , raw sum English: $p=.085$ , problems in Dutch: $p=.469$ , problems in English: $p=.477$ ).
Daily speaking a different language next to Dutch (n=19)	No evidence according to independent-sample t-tests (raw sum Dutch: $p=.690$ , raw sum English: $p=.358$ , problems in Dutch: $p=.497$ , problems in English: $p=.081$ ).

---

<sup>9</sup> Non prametric equivalent points to the same outcomes.

## Appendix F.

Benjamini-Hochberg procedure for multiple testing<sup>10</sup>.

Table 1.

*Benjamini-Hochberg procedure for multiple testing on the Accuracy data. False discovery rate = .25.*

Hypothesis	Original <i>p</i> -values	Benjamini-Hochberg significance	Benjamini-Hochberg <i>p</i> -value
Hypothesis 1	<.001	significant	0,002666667
Hypothesis 4 (Dutch)	<.001	significant	0,002666667
Hypothesis 4 (English)	<.001	significant	0,002666667
Hypothesis 3	.03	significant	0,06
Hypothesis 6	.042	significant	0,0672
Hypothesis 5 (Dutch)	.069	significant	0,091428571
Hypothesis 2	.08	significant	0,091428571
Hypothesis 5 (English)	.586	not significant	0,586

Table 2.

*Benjamini-Hochberg procedure for multiple testing on the Reaction times data. False discovery rate = .25. (original *p*-values of final models).*

Hypothesis	Original <i>p</i> -values	Benjamini-Hochberg significance	Benjamini-Hochberg <i>p</i> -value
Hypothesis 4 (English)	<.001	significant	0,004
Hypothesis 5 (English)	<.001	significant	0,004
Hypothesis 4 (Dutch)	.012	significant	0,032
Hypothesis 5 (Dutch)	.04	significant	0,0672
Hypothesis 6	.042	significant	0,0672
Hypothesis 1	.763	not significant	0,984
Hypothesis 2	.865	not significant	0,984
Hypothesis 3	.984	not significant	0,984

<sup>10</sup> In the case of dependency between hypotheses, suggestion for type II errors might be better ignored and the outcomes must be used to unmask possible type I errors (Benjamini, & Yekutieli, 2001; McDonald, 2014).

## Appendix G.

## FETC-Form.

**APPLICATION FORM FOR THE ASSESSMENT OF A RESEARCH PROTOCOL BY THE FACULTY ETHICS REVIEW BOARD (FERB) OF THE FACULTY OF SOCIAL AND BEHAVIOURAL SCIENCES****General guidelines for the use of this form**

1. This form can be used for a single research project or a series of related studies (hereinafter referred to as: "research programme"). Researchers are encouraged to apply for the assessment of a research programme if their proposal covers multiple studies with related content, identical procedures (methods and instruments) and contains informed consent forms and participant information, with a similar population. For studies by students, the FERB recommends submitting, in advance, a research programme under which protocol multiple student projects can be conducted so that their execution will not be delayed by the review procedure. The application of such a research programme must include a proper description by the researcher(s) of the programme as a whole in terms of the maximum burden on the participants (e.g. maximum duration, strain/efforts, types of stimuli, strength and frequency, etc.). If it is impossible to describe all the studies within the research programme, it should, in any case, include a description of the most invasive study known so far.
2. Solely the first responsible senior researcher(s) (from post-doctoral level onwards) may submit a protocol.
3. Any approval by the FERB is valid for 5 years or until the information to be provided in the application form below is modified to such an extent that the study becomes more invasive. For a research programme, the term of validity is 2 years and any extension is subject to approval. The researcher(s) and staff below commit themselves to treating the participants in accordance with the principles of the Declaration of Helsinki and the Dutch Code of Conduct for Scientific Practices as determined by the VSNU Association of Universities in the Netherlands (which can both be downloaded from the FERB site on the Intranet<sup>11</sup>) and guarantee that the participants (whether decisionally competent or incompetent and/or in a dependent relationship vis-a-vis the researcher or not) may at all times terminate their participation without any further consequences.
4. The researcher(s) commit themselves to maximising the quality of the study, the statistical analysis and the reports, and to respect the specific regulations and legislation pertaining to the specific methods.
5. The procedure will run more smoothly if the FERB receives all the relevant documents, such as questionnaires and other measurement instruments as well as literature and other sources on studies using similar methods which were found to be ethically acceptable and that testify to the fact that this procedure has no harmful consequences. Examples of studies where the latter will always be an issue are studies into bullying behaviour, sexuality, and parent-child relationships. The FERB asks the researcher(s) to be as specific as possible when they answer the relevant questions while limiting their answers to 500 words maximum per question. It is helpful to the FERB if the answers are brief and to the point.
6. **Our FAQ document that can be accessed through the Intranet provides background information with regards to any questions.**
7. The researcher(s) declare to have described the study truthfully and with a particular focus on its ethical aspects.

---

<sup>11</sup> See: <https://intranet.uu.nl/facultaire-ethische-toetsingscommissie-fetc>



Signed for approval<sup>12</sup>:

Date:

---

<sup>12</sup> The senior researcher (holding at least a doctoral degree) should sign here.

**A. GENERAL INFORMATION/PERSONAL DETAILS**

1.

- a. a. Name(s), position(s) and department(s) of the responsible researcher(s):

Moniek.M.J. Schaars MA, docent, department of education

- b. Name(s), position(s) and department(s) of the executive researcher(s):

Charles.J.E. Fayt, student, Department of Education

2. Title of the study or research programme - Does it concern a single study or a research programme? Does it concern a study for the final thesis in a bachelor's or master's degree course?:

Achievement on Arithmetic in bilingual primary schools: Do the language and context matter?

This research is the final thesis of the master's degree in Educational Sciences and as such a single study.

3. Type of study (with a brief rationale):

- quasi-experimental

Participants were involved in four (experimental group) and two (control group) tasks. The tasks will last about 5 minutes and consist of auditory verification tasks done on a computer. To answer the research, the comparison of the data between these experiments (conditions) is required. Participants are not randomly assigned to conditions and are a convenient sample.

4. Grant provider:

None

5. Intended start and end date for the study:

Nov- 2018 till June 2019

6. Research area/discipline:

Educational Sciences, Cognition

7. For some (larger) projects it is advisable to appoint an independent contact or expert whom participants can contact in case of questions and/or complaints. Has an independent expert been appointed for this study?<sup>13</sup>

no

8. Does the study concern a multi-centre project, e.g. in collaboration with other universities, a GGZ mental health care institution, a university medical centre? Where exactly will the study be conducted? By which institute(s) are the executive researcher(s) employed?:

It is not a multi-centre project. The testing will be done at a bilingual primary school and two monolingual primary schools.

8. Is the study related to a prior research project that has been assessed by a recognised Medical Ethics Review Board (MERB) or FERB?

No

If so, which? Please state the file number:

---

<sup>13</sup> This contact may, in principle, also be a researcher (within the same department, or not) who is able to respond to the question or complaint in detail. Independent is to say: not involved in the study themselves. The FERB upholds that an independent contact is not obligatory, but will be necessary when the study is more invasive.

## B. SUMMARY OF THE BACKGROUND AND METHODS

### *Background*

1. What is the study's theoretical and practical relevance? (500 words max.):

Bilingual (primary) schools are developing in Europe and the Netherlands and plenty of work has still to be done in order to fully understand how bilingual education impacts student's achievement. Previous research has paid attention to the development of general cognitive and linguistic advantages in bilinguals compared to monolinguals. In the context of bilingual education, arithmetic performance has remained understudied while bilingual education strives to achieve computation of math operations in both languages with even ease. Divergent ideas have been proposed and this study would contribute by investigating the role of language in the performance of students following bilingual education. The practical relevance is that it gives bilingual schools insight in how to design arithmetic classes. The theoretical relevance is that the study was meant to prove insights to reach a global and unified theory on arithmetic in bilingual education (as part of future research)

2. What is the study's objective/central question?:

How do language and context affect student's performance on arithmetic operations in bilingual education?

3. What are the hypothesis/hypotheses and expectation(s)?:

The hypotheses are organized in three categories for which expectations are formulated from the theory both within the bilingual group as compared to monolinguals:

Raw sums.

Hypothesis 1. BC perform less well and slower on raw sum when the task language is not the class language (English) for arithmetic compared to their performance in the class language for arithmetic (Dutch) (Marsh & Maki, 1976; McClain & Huang; 1992, Gelman & Butterworth, 2005; Le Pichon & Kambel, 2016).

Hypothesis 2. BC perform less well and slower on raw sums in the language not used for arithmetic classes (English) in comparison to monolingual children's (MC) performances tested in their dominant language (Dutch) (Demont, 2001).

Hypothesis 3. BC perform similarly on raw sums, compared to MC in the condition in which the task language is the class language for arithmetic (Dutch) (Demont, 2001; van Rinsveld, et al., 2016).

Raw sums against problems.

Hypothesis 4. BC perform better and faster on problems in both languages compared to their performances in both languages on raw sums (Demont, 2001, ; Swanson, et al., 2018).

Problems.

Hypothesis 5. BC perform better and faster on problems in both languages than MC (Demont, 2001; Le Pichon & Kambel, 2016; Van Rinsveld, et al., 2016).

Hypothesis 6. BC perform better on problems in their dominant language (Dutch) (Le Pichon & Kambel, 2016).

#### *Design/procedure/invasiveness*

4. What is the study's design and procedure? (500 words max.):

The research has a quasi-experimental design.

Before the experiment took place, a letter was sent with the collaboration of the school to all parents of potential participating children. That letter contained information regarding the purpose of the study, the activities their child would be involved to during the experiment, the way the data would be processed, reported and ultimately deleted. Next, to that, parents will find a few questions regarding the child language use (whether the child spoke another language than Dutch or English and whether a language other than Dutch was spoken on daily basis). Parent's written consent was requested prior to the actual testing.

Children were in groups of five, each on a personal laptop while wearing headphones They were instructed (in the language the experiment is taking place; English or Dutch) that they will have to play a game in which they will have to decide as fast as possible whether the computer solved correctly arithmetic operations. Furthermore, it was told to them that In case they did not know or hear, not responding was not a problem. The experiment started with a trial session consisting, then there was room for question, then the child did the actual experiment. Children in the bilingual education took part in the two experiments, but twice (one for each language and condition) and the children in monolingual education took part only in Dutch tasks. Test were spread on different moments (at least two hours between to tests) to avoid learning and fatigue effects. The order of the trials within task was shuffled each time and children did the tasks in different orders.

Invasiveness: None

5.

- a. Which measurement instruments, stimuli and/or manipulations will be used?<sup>14</sup>

Measurement instruments were auditory verification tasks on a computer involving auditory stimuli programmed in Zep. The child heard an arithmetic sum (e.g. 2+2) or a problem (e.g. a problem) and had to decide as fast as possible whether the answer proposed by the computer was correct or not. To do so, they could press on a yes or no button. Tasks measured both accuracy and Reaction Time (RT)

- b. What does the study's burden on the participants comprise in terms of time, frequency and strain/efforts?:

The experimental group was tested four times. Each experiment lasted for about five minutes. Each participant is then tested for 20 minutes, but spread on four different moments.

The control group is tested twice. Each experiment lasted for about five minutes. Each participant was then tested for 10 minutes, but spread on two different moment.

- c. Will the participants be subjected to interventions or a certain manner of conduct that cannot be considered as part of a normal lifestyle?:

No

- d. Will unobtrusive methods be used (e.g. data collection of uninformed subjects by means of observations or video recordings)?:

No

- e. Will the study involve any deception? If so, will there be an adequate debriefing and will the deception hold any potential risks?:

No, the child's performance will not be displayed to them, nor to their teacher, nor to their parents. The computer system displayed after the practice session not whether a given answer was correct or not. Children were thanked even warmly for their participation and all children in the class got a little presents: A Utrecht University Pen and a mini Connect4.

---

<sup>14</sup> Examples: invasive questionnaires; interviews; physical/psychological examination, inducing stress, pressure to overstep important standards and values; inducing false memories; exposure to aversive materials like a unpleasant film, video clip, photos or electrical stimulus; long-term of very frequent questioning; ambulatory measurements, participation in an intervention, evoking unpleasant psychological or physical symptoms in an experiment, denial, diet, blood sampling, fMRI, TMS, ECG, administering stimuli, showing pictures, etc. In case of the use of a device (apparatus) or administration of a substance, please enclose the CE marking brochure for the relevant apparatus or substance, if possible.

6. Will the participants be tested beforehand as to their health condition or according to certain disorders? Are there any inclusion and/or exclusion criteria or specific conditions to be met in order for a participant to take part in this study?:

No

7. Risks for the participants -

- a. Which risks does the study hold for its participants?  
The study did not involve any specific risk for participants.
  
- b. To what extent are the risks and objections limited? Are the risks run by the participants similar to those in daily life?  
Yes

9. How does the burden on the participants compare to the study's potential scientific contribution (theory formation, practical usability)?:

The results of the study shed light on the contradicting results earlier found on this topic and may help in designing a unified and coherent theory of arithmetic at bilingual schools. Practical implication were drawn for the study regarding the teaching and assessing of arithmetic at bilingual schools. So, the burden on the participant (verification tasks lasting at most 20 minutes in total) seems fair regarding the value of the outcome.

10. Will a method be used that may, by coincidence, lead to a finding of which the participant should be informed?<sup>15</sup> If so, what actions will be taken in the case of a coincidental finding?:

No

#### *Analysis/power*

11. How will the researchers analyse the data? Which statistical analyses will be used?:

Hypothesis (tasks)	Analysis	Independent variables	Dependent variables
1 (raw sums)	Accuracy: A Paired Sample T-test. RT: Multi-level modeling analyses	Task language	Accuracy and RT
2 (raw sums)	Accuracy: Independent sample t-test RT: Multi-level modeling analyses	Kind of education	Accuracy and RT
3 (raw sums)	Accuracy: Independent sample-test. RT: Multi-level modeling analyses	Kind of education	Accuracy and RT
4 (raw sums vs problems)	Accuracy: Paired-sample t-test. RT: Multi-level modeling analyses	Task (raw sums vs. problems) Task language	Accuracy and RT
5 (problems)	Accuracy: Independent sample t-test. RT: Multi-level modeling analyses	Kind of education (bilingual vs. monolingual) Task language	Accuracy and RT
Hypothesis 6 (problems)	Accuracy: Paired-sample t-test. RT: Multi-level modeling analyses	Task language	Accuracy and RT

11. What is the number of participants? Provide a power analysis and/or motivation for the number of participants. The current convention is a power of 0.80. If the study deviates from this power, the FERB would like you to justify why this is necessary:

<sup>15</sup> For instance: dementia, dyslexia, giftedness, depression, extremely low heartbeat in an ECG, etc. If coincidental findings may be found, this should be included in the informed consent, including a description of the actions that will be taken in such an event.



The total number of participants was 76 students spread as 39 students at a monolingual school and 37 at a bilingual school. For research in the social sciences, this a relatively small sample. This had to do with practical reasons that are detailed hereafter.

Bilingual schools in the Netherlands are developing. There are relative small numbers of bilingual school (around 20 across the whole country) which limit the number of participants available. Testing participants over many schools would drastically higher the change of cofounding results. As a results two classes at the same bilingual school were tested and two groups at a monolingual schools in order to get two groups of similar size.

Furthermore, the sample size is in line with many studies cited in the theoretical framework which did find results. This suggests that this sample size would be enough.

**C. PARTICIPANTS, RECRUITMENT AND INFORMED CONSENT PROCEDURE**

1. The nature of the research population (please tick):

1. General population without complaints/symptoms

2. Age category of the participants (please tick):

- 12 years or younger

2. Does the study require a specific target group? If so, justify why the study cannot be conducted without the participation of this group (e.g. minors):

To know more on bilingual primary education and how this impacts student's achievement, it is needed to test minors following bilingual primary education.

4. Recruitment of participants -

- a. How will the participants be recruited?

By contacting schools. So, the school board will give written consent as the parents (written) and the child (orally)

- b. How much time will the prospective participants have to decide as to whether they will indeed participate in the study?

School board got as much time as they wanted to react (they were approached few months or in the latest cases few weeks before testing), parents will be handed a letter few weeks before testing and requested to return the informed consent prior to the first day of testing, and finally, children were asked if they are willing to participate before the testing start (on the same day).

5. Does the study involve informed consent or mutual consent? Clarify the design of the consent procedure (who gives permission, when and how). Does the study involve active consent or passive consent? If no informed consent will be sought, please clarify the reason:

School board: written in an email and by active informed consent.

Parents: written by active informed consent.

Children: before the testing starts by saying yes to the question: *do you want to participate?*

6. Are the participants fully free to participate and terminate their participation whenever they want and without stating their grounds for doing so?:

This was mentioned in the informed consents for the school board and parents and was explicitly mentioned to the child before the experiment started

7. Will the participants be in a dependent relationship with the researcher?:

No

8. Compensation

- a. Will the participants be compensated for their efforts? If so, what is included in this recompense (financial reimbursement, travelling expenses, otherwise). What is the amount?

Each child will be handed a mini connect4. game and a Utrecht University pen

- b. Will this compensation depend on certain conditions, such as the completion of the study?

No every child in the class got the compensation.

**D. PRIVACY AND INFORMATION**

1.

- a. Will the study adhere to the requirements for anonymity and privacy, as referred to in the Faculty Protocol for Data Storage<sup>16</sup>?:
- anonymous processing and confidential storage of data (i.e. storage of raw data separate from identifiable data): yes
  - the participants' rights to inspect their own data: yes
  - access to the data for all the researchers involved in the project: yes

If not, please clarify.

- b. Has a Data Management Plan been designed?

All anonymized data files (with only participants 'codes) and the informed consents were stored in the YODA environment of the Utrecht University. After completion of the thesis and after the amount of time the data must be conserved for achieve purposes, the data will be deleted. The data file containing the connection between participant's name and code has been saved to another safe environment and is not accessible to any outsiders.

2.

- a. Will the participant be offered the opportunity to receive the results (whether or not at the group level)?:

Yes, at the group level.

- b. Will the results of the study be fed back to persons other than the participants (e.g. teachers, parents)?:

Yes, parents, teachers, school boards in the form of the sending of the research report and an extend summary in Dutch.

If so, will this feedback be provided at the group or at the individual level?

---

<sup>16</sup> This can be found on the Intranet: <https://intranet.uu.nl/wetenschappelijke-integriteit-facultair-protocol-dataopslag>

Group level.

3.

- a. Will the data be stored on the faculty's data server?

Yes

- b. Will the data that can be traced back to the individual be stored separately on the other faculty server available for this specific purpose?

Yes

**E. ADDITIONAL INFORMATION**

Optional.

**F. FORMS TO BE ENCLOSED (CHECKLIST)**

- Text (advert) for the recruitment of participants
- Information letter for participant
- Informed consent form for participants
- Written or oral feedback information (debriefing text)
- (Descriptions of) questionnaires
- (Descriptions of) measurement instruments/stimuli/manipulations
- Literature/references

Signature(s):<sup>17</sup>

Date and place:

Name, position:

---

<sup>17</sup> The senior researcher (holding at least a doctoral degree) should sign here.

**Appendices****Appendix I.****Recruit letter for school broad****A. BILINGUAL****Universiteit Utrecht***Charles Fayt**Henegouwen 98**3524 RB Utrecht**Tel: 0631271721**Email: [c.j.e.fayt@uu.nl](mailto:c.j.e.fayt@uu.nl)*

Utrecht, januari 2019

Betreft: onderzoek naar invloed van taal op rekenvermogen

Aanleiding: Masterthesisonderzoek, *Educational Sciences*

Instelling: Universiteit Utrecht

Geachte directie,

Mijn naam is Charles Fayt en ik volg met veel plezier en enthousiasme de master *Educational Sciences* aan de Universiteit. In het kader van mijn afstudeeronderzoek neem ik de vrijheid om met u contact op te nemen.

Tweetalig onderwijs is in volop ontwikkeling in Nederland en uw school heeft er al voor gekozen om een tweetalig curriculum aan te bieden. Zoals u mogelijk weet, heeft onderzoek al meermaals aangetoond dat kinderen die tweetalig onderwijs volgen voordelen zouden ontwikkelen op het gebied van cognitie en taal. Er blijft desondanks veel werk te verrichten om tweetalig onderwijs zo goed mogelijk vorm te geven. Een gebied dat tot op heden redelijk ongestudeerd bleef, is wiskunde. Een klein aantal onderzoeken heeft geen eenduidige conclusie met zich gebracht maar een analyse van de literatuur suggereert dat kinderen in het tweetalig onderwijs beter zouden moeten presteren op verhaaltjessommen. Ook bestaat er onduidelijkheid over hoe de taal van de opdracht rekenen kan beïnvloeden, al is rekenvermogen aan zich geen talige activiteit bij uitstek. Een goed beeld van dit onderwerp is noodzakelijk voor het optimaliseren van tweetalig primair onderwijs in Nederland en Europa.

Graag will ik u uitnodigen om deel te nemen aan dit onderzoek. Deelname van uw school zal inhouden dat ik in maart 2019 een viertal digitale taken afnemen bij de leerlingen die nu in groep 4 zitten. Elk experiment duurt ongeveer 5 minuten en worden elk op een ander moment afgenomen. Benodigd materiaal wordt meegebracht. Het ter beschikking stellen van een ruimte wordt gewaardeerd maar is niet noodzakelijk. Uiteraard worden ouders geïnformeerd over het onderzoek en worden gevraagd om akkoord te gaan. Na afloop zal aan uw school een rapportage worden toegezonden. Zo krijgt u inzicht in hoe taal en een talige context de prestatie van leerlingen kunnen beïnvloeden. Zo kunt u uw onderwijs verder nog optimaliseren. Uw deelname en de resultaten van het onderzoek zouden ook gebruikt kunnen worden om uw school van een unieke positie te voor zien. Dergelijk onderzoek is nooit eerder gedaan en het is aannemelijk dat dit onderzoek positief zou kunnen bijdragen

aan uw imago. Uiteindelijk komt u dan heel dichtbij te staan van hoe tweetalig onderwijs wordt onderzocht en vormgegevens. Uw deelname biedt ons de kans om het onderwijs van morgen beter en efficiënter te maken.

Ik hoop met heel mijn hart dat uw school bereid is deel te nemen aan dit onderzoek. Volgende week zal ik hierover telefonisch contact met u opnemen. Dan zal ik ook graag uw verdere vragen en opmerkingen bij het onderzoek bespreken.

Met vriendelijke groet,

Charles Fayt BA ; begeleid door Drs. Moniek Schaars



**B. MONOLINGUAL****Universiteit Utrecht***Charles Fayt**Henegouwen 98**3524 RB Utrecht**Tel: 0631271721**Email: [c.j.e.fayt@uu.nl](mailto:c.j.e.fayt@uu.nl)*

Utrecht, januari 2019

Betreft: onderzoek naar invloed van taal op rekenvermogen

Aanleiding: Masterthesisonderzoek, *Educational Sciences*

Instelling: Universiteit Utrecht

Geachte directie,

Mijn naam is Charles Fayt en ik volg met veel plezier en enthousiasme de master *Educational Sciences* aan de Universiteit. In het kader van mijn afstudeeronderzoek neem ik de vrijheid om met u contact op te nemen.

Opdrachten bij wiskunde kunnen aangeboden worden op twee manieren. De eerste is een kale sommen, waarbij de leerling antwoord moet geven op de vraag; wat is  $2+2$ ?. De tweede is een verhaaltjesom waarbij de leerling uit een (talige) context de nodige informatie eruit moet halen om antwoord te kunnen geven op een vraag zoals: hoeveel kinderen zitten er nu in de bus? Beide vormen ogen hetzelfde te meten maar er is onduidelijkheid over de best werkende vorm. Ook mist er onderzoek in relatie met de vorm onderwijs, zoals tweetalige scholen of andere scholen die een bepaalde onderwijsvorm hanteren. Een goed beeld van dit onderwerp is noodzakelijk, zowel binnen leerling als tussen schoolvormen voor het optimaliseren van het primair onderwijs in Nederland en Europa.

Graag will ik u uitnodigen om deel te nemen aan dit onderzoek. Deelname van uw school zal inhouden dat ik in maart 2019 een tweetal digitale taken afnemen bij de leerlingen die nu in groep 4 zitten. Elk experiment duurt ongeveer 5 minuten en worden elk op een ander moment afgenomen. Benodigd materiaal wordt meegebracht. Het ter beschikking stellen van een ruimte wordt gewaardeerd maar is niet noodzakelijk. Uiteraard worden ouders geïnformeerd over het onderzoek en worden gevraagd om akkoord te gaan. Na afloop zal aan uw school een rapportage worden toegezonden. Zo krijgt u inzicht in hoe verschillende contexten voor wiskunde een invloed kunnen oefenen en hoe deze data zich verhoudt tot andere onderwijsvormen. Dergelijk onderzoek is vernieuwend en het is aannemelijk dat dit onderzoek positief zou kunnen bijdragen aan uw imago. Uiteindelijk komt u dan heel dichtbij te staan van hoe onderwijs wordt onderzocht en vormgegeven. Uw deelname biedt ons de kans om het onderwijs van morgen beter en efficiënter te maken. Ook belangrijk om te weten is dat de

resultaten niet de mogelijkheid geven om enig oordeel te doen over de kwaliteit van het onderwijs op uw school. Het onderzoek is daar ook niet voor bedoeld.

Ik hoop met heel mijn hart dat uw school bereid is deel te nemen aan dit onderzoek. Volgende week zal ik hierover telefonisch contact met u opnemen. Dan zal ik ook graag uw verdere vragen en opmerkingen bij het onderzoek bespreken.

Met vriendelijke groet,

Charles Fayt BA ; begeleid door Drs. Moniek Schaars

## Appendix 2.

### Active informed consent

Beste ouder/ verzorger,

#### Wie ben ik?

Mijn naam is Charles Fayt, 24 jaar en ik ben student aan de Universiteit Utrecht waar ik met veel plezier de master Educational Sciences volgt. Het onderzoek wordt uitgevoerd in het kader van mijn masterscriptie. Daarom heb ik de school van uw kind benaderd.

#### Wat is het doel van het onderzoek?

Het onderzoek haalt de mogelijke invloed van taal (Engels of Nederlands) en van een talige context (kale sommen vs. rekensommen) op de rekenvermogen van de leerlingen.

#### Wat houdt het onderzoek in?

Deelnemers doen twee/vier taken op een computer. Leerlingen horen een stem die een kale som (2+2) of een verhaaltjessom voorleest. Vervolgens stel de stem een mogelijk antwoord. Aan de leerling is de taak om te bepalen of het antwoord goed of slecht is door zo snel mogelijk te drukken op een ja of op een nee knop. Elk experiment duurt vijf minuten verspreid of twee testmomenten.

#### Privacy en vertrouwelijkheid

Alle gegevens worden vertrouwelijk behandeld en anoniem verwerkt. Het onderzoek is niet bedoeld en niet in staat om een kwalitatief oordeel te doen over het algemene rekenvermogen van uw kind. De docent krijgt de antwoorden van de individuele leerlingen niet te zien. De gegevens worden alleen voor opleidings- en onderzoeksdoeleinden gebruikt. Leerlingen kunnen zelf ook aangeven of ze wel of niet mee willen doen net voor het begin van het onderzoek. Die vraag zal namelijk aan hem gesteld worden.

#### Mogelijkheid tot vragen, informatie en toestemming

Als u nog vragen heeft over het onderzoek of als u op de hoogte gehouden wilt worden over dit onderzoek, stuur dan een mail aan: Charles Fayt; [c.j.e.fayt@uu.nl](mailto:c.j.e.fayt@uu.nl) . Voor verdere vragen over de cursus en opdracht die ik maak, kunt u contact opnemen met: Moniek Schars; [m.m.h.schaars@uu.nl](mailto:m.m.h.schaars@uu.nl).

We verzoeken u vriendelijk om het onderstaande strookje in te vullen zowel als u bezwaar of geen bezwaren heeft omtrent de deelname van uw kind. U kunt het meegeven aan uw kind. Het strookje kan **uiterlijk** **[DATUM]** ingeleverd worden bij de mentor van uw kind.₂

U mag het strookje ook digitaal invullen, als er maar een handtekening op staat (bv. inscannen/foto maken met telefoon). **Stuurt u uw strookje dan naar:** [c.j.e.fayt@uu.nl](mailto:c.j.e.fayt@uu.nl)

We danken u,

Met vriendelijke groet,

MA, Moniek Schaars, Charles Fayt

✂-----

Ik vind het **goed** - **niet goed** (omcirkel wat van toepassing is) dat mijn kind meedoet aan dit onderzoek

Uw naam:

Datum:

Naam kind:

Handtekening:

Klas kind:

School:

Mijn kind kan zich uiten in een andere taal dan Nederlands of Engels **JA – NEE** (omcirkel wat van toepassing is)


Mijn kind spreekt buiten de school een andere taal dan Nederlands **JA – NEE** (omcirkel wat van toepassing is)




**Appendix 3.****Descriptions of the materials.**

Participants took part in two auditory verification tasks. They differ in the extent to which language was incorporated. One condition involved raw sums and the other one, problems. Each operation consisted of two one-digit numbers with a result up to 16 reached by addition, subtraction, multiplication, which is line with what Demont (2001) and Le Pichon and Kambel (2016) did with participants of the same age. The teacher's opinion about the task level of each class was asked three days prior to testing using example items that were similar but not included in the actual tasks to check whether children should have automatized these operations (Canavesio, 2013; Le Pichon & Kambel, 2016, Blom & Unsworth, 2010). Both tasks had a Dutch and English version. Auditory stimuli made it possible to investigate the influence of the task language. The stimuli were recordings of a simultaneous bilingual (English-Dutch) female speaker recorded and edited in Praat, version 6.0.52 (Boersma & Weenink, 2019). Children prefer listening to female voices (Blom & Unsworth, 2010). The tasks were programmed in Zep version 1.16 (Veenker, 2018). The tasks were digital and measured accuracy and RT in milliseconds (ms.) (UiL-OTS, n.d.). Participants had 15 s. to make a choice, otherwise, Zep moved to the next item. The tasks were also grounded in the reaction times paradigm where differences in RT are associated with differences in level of difficulty and heavier cognitive processing (Baayen & Milin, 2010). Faster RT are also associated with the language used for computation (Canavesio, 2013).

**Materials for raw sums.** Canavesio (2013) did similar research and experiment. Canavesio (2013)'s raw sums task replicated here. It is grounded in the paradigm of LeFevre (1998), a methodology to measure arithmetic performance and linguistic interferences (Canavesio, 2013; LeFevre, 1998). The child must decide whether an auditory presented arithmetic operation (e.g.,  $3+2$ ) is associated with the correct auditory target (e.g., 5). The decision is made by pressing a yes or no button as fast as possible. A trial sequence can be found in Table 1. Canavesio (2013) used 24 trials consisting of 12 trials requiring a yes-answer (four additions, four subtractions, four multiplications) and 12 trials a no-answer (four additions, four subtractions, four multiplications) (see Appendix A for Dutch and Appendix B for English trials). Consequently, the same was done here. The task lasted about three minutes and was preceded by six feedbacked practice items. Furthermore, the experiment should uncover the language in which students process arithmetic operations (Canavesio, 2013). The amount of time between the operation and the target was maintained constant across all trials.


Table 1.




Sequence of a Trial with Raw Sums (  indicates that the information was only auditory presented with a white screen).

Step	On display/heard 	Time (in ms)
Fixation cross	+	1000
Operation	2+2	
		
Operation to Target interval		1500
Target	“four”	
		
Respons interval		Max. 15000
Intertrial interval		1500

**Materials for problems.** Experiment problems took the design of experiment 1 but involved problems instead. The child was auditory presented with an arithmetic problem like in Le Pichon and Kambel (2016) who tested a comparable group. The problems involved different contexts to ensure that participants paid attention to what was said (Le Pichon & Kambel, 2016; Blom & Unsworth, 2010). A voice announced a possible answer and the child had to decide as fast as possible if that answer was correct or not by pressing on a yes- or no-button. A trial sequence can be found in Table 2. Le Pichon and Kambel (2016) used ten trials because more trials would cause a fatigue effect. Accordingly, twelve trials were used in the current design to ensure an equal spreading between the kind of operation. Six requested a yes-answer (two additions, two subtractions, two multiplications) and six a no-answer (two additions, two subtractions, two multiplications) (see Appendix C for Dutch and Appendix D for English trials). The amount of time between the operation and the target was maintained constant across all trials. The experiment lasted about four minutes and started with three feedbacked practice trials. The number of items differed from the raw sums task because problems are inherently cognitively heavier and more exhaustive (Le Pichon & Kambel, 2016, Van Rinsveld, et al., 2016).

Table 2.

Sequence of a Trial with Problems (  indicates that the information was only auditory presented with a white screen).

Step	On display/heard 	Time (in ms)
Fixation cross	+	1000
Operation 	‘There are two children in the bus. Two more children step into the bus. How many children are now on the bus?’	
Operation to Target interval		1500
Target 	“Four”	
Respons interval		Max. 15000
Intertrial interval		1500

**Appendix 4.****Materials (trials) per task**

A. Materials and trials used for raw sums in Dutch.  
Table 1.

*Materials and trials used for raw sums in Dutch.*

Operation	Target	Correctness
2+2	4	Correct
3+4	7	Correct
9+3	12	Correct
7+6	13	Correct
9+1	3	Incorrect
1+4	7	Incorrect
9+5	15	Incorrect
8+7	11	Incorrect
9-2	7	Correct
6-4	2	Correct
9-6	3	Correct
8-7	1	Correct
8-2	10	Incorrect
7-1	3	Incorrect
8-4	6	Incorrect
3-2	8	Incorrect
4x1	4	Correct
5x1	5	Correct
3x3	9	Correct
2x7	14	Correct
2x1	4	Incorrect
5x2	7	Incorrect
4x4	11	Incorrect
2x6	15	Incorrect



B. Materials and trials used for the raw sums in English.

Table 1.

*Materials and trials used for the raw sums in English.*

Operation	Target	Correctness
1+2	3	Correct
3+5	8	Correct
9+2	11	Correct
6+8	14	Correct
9+4	12	Incorrect
1+7	4	Incorrect
2+4	3	Incorrect
5+8	12	Incorrect
9-1	8	Correct
6-3	3	Correct
9-5	4	Correct
8-6	2	Correct
8-3	10	Incorrect
7-2	5	Incorrect
8-5	6	Incorrect
4-2	7	Incorrect
2x3	6	Correct
3x5	15	Correct
3x4	12	Correct
3x1	3	Correct
2x2	6	Incorrect
1x1	3	Incorrect
2x8	15	Incorrect
2x4	10	Incorrect

## C. Materials and trials used for problems in Dutch.

Table 1.

*Materials and trials used for problems in Dutch.*

Problem	Target	Correctness
Er zijn 9 auto's op het parkeerterrein. Er komen nog 4 auto's bij. Hoeveel auto's zijn er nu op het parkeerterrein?	13	Correct
Mama geeft me 2 boterhammen voor de lunch. Ik vraag er nog 3 bij en mama maakt en geeft ze. Hoeveel boterhammen heb ik nu?	5	Correct
Op de luchthaven staan er 9 vliegtuigen. Er zijn er nog 3 bij geland. Hoeveel vliegtuigen zijn er nu op de luchthaven?	14	Incorrect
Ik heb 2 mooie goudvissen in mijn aquarium. Ik koop er nog 6 bij. Hoeveel vissen heb ik nu?	9	Incorrect
De supermarkt heeft nog 9 repen melkchocolade. Mijn vriend koopt er 3. Hoeveel repen melkchocolade kan ik nog kopen?	6	Correct
Een student had 4 opdrachten voor morgen gekregen. Nu heeft hij er al twee af. Hoeveel opdrachten moet hij nog maken?	2	Correct
Een leraar had 9 boeken. Hij is er 5 kwijt. Hoeveel boeken heeft de leraar nu?	6	Incorrect
De banketbakker had 9 appeltaarten gemaakt. Hij heeft er vandaag 8 verkocht. Hoeveel appeltaarten zijn er over?	3	Incorrect
De moeder van Anna geeft haar elke dag 2 gevulde koeken als tussendoortjes. Dit duurt 6 dagen lang. Hoeveel gevulde koeken heeft Anna na die tijd gehad?	12	Correct
Peter wint 3 superheldenplaatjes tijdens het spelen per dag. Dit gebeurt 3 dagen achter elkaar. Hoeveel plaatjes heeft hij in totaal gewonnen?	9	Correct
De appels worden per 2 verkocht in een zak. Ik koop 7 zakjes. Hoeveel appels heb ik nu?	11	Incorrect
Op het station liggen er 3 sporen, elk voor één trein bedoeld. Hoeveel treinen kunnen in het station?	1	Incorrect

## D. Materials and trials used for problems in English.

Table 2.

*Materials and trials used for problems in English.*

Problem	Target	Correctness
There are 8 busses at the bus station. 6 busses have just arrived at the station now. How many busses are there at the bus station?	14	Correct
Mom gave me 1 cookie for lunch. At school, the teacher gave me 2 more cookies. How many cookies do I have?	3	Correct
Peter had 9 toys. For his birthday, he got 2 new toys. How many toys does Peter have now?	10	Incorrect
My parents have 2 cats at home. They decided to adopt 4 new cats. How many cats do we have at home?	8	Incorrect
There are 9 waffles in the bakery. John takes 4 of them. How many waffles are left for William?	5	Correct
I had 7 chocolate bars. I gave 5 to my friends. How many do I have now?	2	Correct
The teacher had 6 pencils. He lost one of them. How many pencils do the teacher have?	4	Incorrect
I have made a bouquet of 8 flowers for mom. One fell on the ground. How many flowers are left in the bouquet?	6	Incorrect
I got 2 euros from my parents every day. It has lasted for 5 days. How many euros do I have now?	10	Correct
Anna won 2 Princesses cards. She did so for 2 days. How many Princesses cards did Anna win?	4	Correct
In a holiday house, there are 2 bedrooms. Each room has 4 single beds. How many people can sleep there?	6	Incorrect
3 friends like to play together. They all have 4 board games each. How many board games can they choose of?	10	Incorrect

**Appendix 5.****Reference List**

Abdelilah-Bauer, B. (2015). *Le défi des enfants bilingues*. Paris, France : La découverte.

Althouse, A. D. (2016). Adjust for multiple comparisons? It's not that simple. *The Annals of thoracic surgery*, 101(5), 1644-1645. doi:

<https://doi.org/10.1016/j.athoracsur.2015.11.024>

Anderson, J. A., Chung-Fat-Yim, A., Bellana, B., Luk, G., & Bialystok, E. (2018). Language and cognitive control networks in bilinguals and monolinguals. *Neuropsychologia*, 117, 352-363. doi: <https://doi.org/10.1016/j.neuropsychologia.2018.06.023>

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

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12-28. doi: <https://doi.org/10.21500/20112084.807>

Becker, L.A. (2000). Effect size calculators [Apparatus]. Colorado Springs, USA: University of Colorado. Retrieved from <https://www.uccs.edu/lbecker/>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 57, 289-300. doi: <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4), 1165-1188. doi:

10.1214/aos/1013699998

Blom, E., Kuntay, A. C., Messer, M., Verhagen, J., & Leseman, P. (2014). The benefits of being bilingual: Working memory in bilingual Turkish Dutch child. *Journal of Experimental Child Psychology*, 128, 105-119. doi: <https://doi.org/10.1016/j.jecp.2014.06.007>

Blom, E., & Unsworth, S. (Eds.). (2010). *Experimental methods in language acquisition research* (Vol. 27). Amsterdam, the Netherlands: John Benjamins Publishing.

Boersma, P.P.G & Weenink, D.J.M. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.52. Retrieved from <http://www.praat.org/>

Canavesio, M. L. (2013). *Bilingual Education in the Primary School: Curriculum Study and Experimental Research on Language of Acquisition Effects in the Arithmetic Facts*. (Doctoral dissertation). University of Trento, Trento, Italy.

Costa, A., Hernandez, M., & Sebastian-Galles, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, 106, 59–86. doi :

<https://doi.org/10.1016/j.cognition.2006.12.013>

Demont, E. (2001). Contribution de l'apprentissage précoce d'une deuxième langue au développement de la conscience linguistique et à l'apprentissage de la lecture. *International journal of psychology*, 36(4), 274-285. doi:

<https://doi.org/10.1080/00207590042000137>

Ellis, P. D. (2009). Thresholds for Interpreting Effect Sizes. Retrieved from

[http://www.polyu.edu.hk/mm/effectsizefags/thresholds\\_for\\_interpreting\\_effect\\_sizes\\_2.html](http://www.polyu.edu.hk/mm/effectsizefags/thresholds_for_interpreting_effect_sizes_2.html)

de Graaff, R. (03-10-2013). [\*Taal om te leren: Didactiek en opbrengsten van tweetalig\*](#)

[\*onderwijs\*](#). Utrecht, the Netherlands: Universiteit Utrecht.

Driessen, G., Krikhaar, E., de Graaff, H.C.J., Unsworth, S., Leest, B., Coppens, K. & Wierenga,

J. (2016). [\*Evaluatie pilot Tweetalig Primair Onderwijs - Startmeting schooljaar 2014/15\*](#)

[\*\(publieksversie\)\*](#). Nijmegen, The Netherlands: ITS Nijmegen.

Durlik, J., Szewczyk, J., Muszyński, M., & Wodniecka, Z. (2016). Interference and Inhibition in

Bilingual Language Comprehension: Evidence from Polish-English Interlingual

Homographs. *PloS one*, 11(3), e0151430. doi:

<https://doi.org/10.1371/journal.pone.0151430>

García, O. (2011). *Bilingual education in the 21st century: A global perspective*. Hoboken,

USA: John Wiley & Sons.

Gelman, R., & Butterworth, B. (2005). Number and language: how are they related?. *Trends*

*in cognitive sciences*, 9(1), 6-10. doi: <https://doi.org/10.1016/j.tics.2004.11.004>

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple

comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189-211. doi:

<https://doi.org/10.1080/19345747.2011.618213>

Gollan, T.H., Montoya, R.I., & Werner (2002). Semantic and letter fluency in Spanish-English

bilinguals. *Neuropsychology*, 16, 562-576. doi:

<https://doi.org/10.1016/j.tics.2004.11.004>

Gollan, T.H., Montoya, R.I., Fennema-Notestine, C., & Morris, S.K. (2005). Bilingualism affects picture naming but not picture classification. *Memory and Cognition*, 33,1220-1234.

doi: <http://dx.doi.org/10.1037/0894-4105.16.4.562>

Jääskeläinen, R. (2010). Think-aloud protocol. In Y. Gambier & L. van Doorslaer, (Eds.) *Handbook of translation studies* (371-374), Amsterdam, the Netherlands: John Benjamin.

Jenniskens, T., Leest, B., Wolbers, M., Krikhaar, E., Teunissen, C., de Graaff, H., Unsworth, S. & Coppens, K. (2018). [Evaluatie pilot Tweetalig Primair Onderwijs - Vervolgmeting schooljaar 2016/17 - publieksversie](#). Nijmegen, the Netherlands: KBA Nijmegen.

Length, R. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software*, 69(1), 1-33. doi: [doi:10.18637/jss.v069.i01](https://doi.org/10.18637/jss.v069.i01)

Laurent, A., & Martinot, C. (2010). Bilingualism and phonological awareness: the case of bilingual (French–Occitan) children. *Reading and Writing*, 23(3-4), 435-452. doi: <https://doi.org/10.1007/s11145-009-9209-3>

Le Pichon, E.M.M. & Kambel, E-R. (2016). [Challenges of mathematics education in a multilingual post-colonial context - The case of Suriname](#). In Z. Babaci-Wilhite (Eds.), *Human rights in language and STEM education* (pp. 221-240). Rotterdam, the Netherlands: Sense Publishers.

Le Pichon, E.M.M. (2013). Handling plurilingualism in kindergarten and primary school. In W., Griebel; R., Heinisch; C., Kieferle; E. Röbe & A., Seifert, (Eds.), *Transition to School and Multilingualism – A Curriculum for Educational Professionals*. Hamburg, Germany: Verlag Dr. Kovac.

Lindman, H. R. (1974). *Analysis of Variance in complex experimental design*. New York, USA:

W.H. Freeman and Co.

Marsh, L. G., & Maki, R. H. (1976). Efficiency of arithmetic operations in bilinguals as a

function of language. *Memory & Cognition*, 4(4), 459-464. doi:

<https://doi.org/10.3758/BF03213203>

McClain, L., & Huang, J. Y. S. (1982). Speed of simple arithmetic in bilinguals. *Memory &*

*Cognition*, 10(6), 591-596. doi: <https://doi.org/10.3758/BF03202441>

McDonald, J.H. (2014). *Handbook of Biological Statistics* (3rd ed.). Baltimore, USA: Sparky

House Publishing.

McDonald, J.H. (2015). Multiple comparisons -Spreadsheet [Apparatus]. Retrieved from

<http://www.biostathandbook.com/multiplecomparisons.html>

Mehisto, P., Marsh, D., & Frigols, M. J. (2008). *Uncovering CLIL content and language*

*integrated learning in bilingual and multilingual education*. London, United-Kingdom:

Macmillan.

Miller, R.G. (1981). *Simultaneous Statistical Inference* (2nd ed.). New-York, USA Springer

Verlag.

Nicoladis, E., & Jiang, Z. (2018). Language and cognitive predictors of lexical selection in

storytelling for monolingual and sequential bilingual children. *Journal of Cognition and*

*Development*, 19(4), 413-430. doi: <https://doi.org/10.1080/15248372.2018.1483370>

Nieuwenhuis, R., te Grotenhuis, M. & Pelzer, B. (2012). influence.ME: Tools for Detecting

Influential Data in Mixed Effects Models. *R Journal*, 4(2), 38-47.



Nikula T. (2016). CLIL: A European Approach to Bilingual Education. In: N. Van Deusen-Scholl & S. May (eds) *Second and Foreign Language Education. Encyclopedia of Language and Education* (3rd ed.). Cham, Switzerland: Springer.

Nuffic. (n.d). Tweetalige basisscholen. Retrieved from

<https://www.nuffic.nl/onderwerpen/tweetalige-basisscholen/>

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna:

Austria, R Foundation for *Statistical Computing*. ISBN 3-900051-07-0, URL

<http://www.R-project.org/> (version 3.5.3).

Perneger, T. V. (1999). Adjusting for multiple testing in studies is less important than other concerns. *Bmj*, 318(7193), 1288.

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, 27(7), 1036-1042. <https://doi.org/10.1177/0956797616645672>

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* 1(1), 43-46. doi: <https://doi.org/10.1136/bmj.318.7193.1288a>

Savitz, D. A., & Olshan, A. F. (1998). Describing data requires no adjustment for multiple comparisons: A reply from Savitz and Olshan. *American Journal of Epidemiology*, 147(9), 813-814. doi: 10.1093/oxfordjournals.aje.a009532

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological bulletin*, 111(2), 352. doi: <http://dx.doi.org/10.1037/0033-2909.111.2.352>

Swanson, H. L., Kong, J., & Petcu, S. (2018). Math Difficulties and Working Memory Growth in English Language Learner Children: Does Bilingual Proficiency Play a Significant Role?. *Language, speech, and hearing services in schools*, 49(3), 379-394. doi:

[https://doi.org/10.1044/2018\\_LSHSS-17-0098](https://doi.org/10.1044/2018_LSHSS-17-0098)

Stryker, S. B., & Leaver, B. L. (Eds.). (1997). *Content-based instruction in foreign language education: Models and methods*. Georgetown, Washington D.C.: Georgetown University Press.

UiL-OTS, (n.d). The Zep programming language. Retrieved from <https://uilots-labs.wp.hum.uu.nl/software/zep/>

Van Rinsveld, A., Schiltz, C., Brunner, M., Landerl, K., & Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter?. *Learning and instruction*, 42, 72-82. doi: <https://doi.org/10.1016/j.learninstruc.2016.01.003>

Veenker, T.J.G. (2018). The Zep Experiment Control Application (Version 1.16) [Windows 10]. Utrecht Institute of Linguistics OTS, Utrecht University. Available from <http://www.hum.uu.nl/uilots/lab/zep/> .

Verhoeven, P. S., & van Baal, A. (2011). *Doing research: The hows and whys of applied research*. The Hague, the Netherlands: Eleven International Publishing.

Vogt, W. P., & Johnson, B. (2011). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. New York, USA: Sage publication.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New-York, USA: Springer-Verlag.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499*.