Old wine in a new bottle? Re-examining speeded performance in mental arithmetic using linear mixed models

by

Chunyun Ma

A thesis submitted to

the Faculty of Graduate and Postdoctoral Affairs

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Psychology

Carleton University

Ottawa, Canada

Abstract

Theory construction in cognitive psychology relies on choosing a suitable statistical model for data analyses. I hypothesized that the general linear model (GLM), which is currently used for analyzing data in the field of mental arithmetic by most researchers, is not suitable for analyzing data that is typical in this field. In particular, mental arithmetic data have three features that make them unsuitable for GLM analyses: repeated measurements from individual participants, a mixture of categorical and continuous predictors, and unevenly distributed observations across levels of predictors. I proposed an alternative approach using linear mixed models (LMM), as a better candidate. I tested the hypothesis by applying GLM and LMM to three archival datasets typical in the field of mental arithmetic. Across the re-analyses, LMM consistently showed advantages over GLM. LMM was used successfully on unbalanced designs (Chapter 5) and better preserved the continuous nature of independent variables (Chapters 3, 4, and 5). I discussed the findings and provided advice for researchers who want to adopt LMM in their studies.

Acknowledgement

It was quite a journey. I am now at the end of it.

Thank you, Jo-Anne. Without your guidance, I would be lost so many times during this journey.

Thank you, Andrea. Your approach to statistics has forever changed mine.

Thank you, Warren. Because of you, I learned how to write.

Thank you, Robin and Cecilia. Our writing group has been a constant source of moral support for me in the past three years.

Thank you, Chang. Our friendship kept me sane during the most difficult times.

Thank you, Patricia, Andrew, and Andrea. Time spent with you, my adopted family, always fills me with warmth and joy.

Thank you, Hooman. Being part of your boot camps is the best gift I could imagine during the dark days towards the end of this journey.

Thank you, mum and dad. No amount of words can express my gratitude to you.

So,

Let the new journey begin.

Table of contents

List of Figures

List of Tables

CHAPTER 1: Introduction

"Reaction time data alone cannot settle this question but can, at least, provide conditions that must be met by any model of the multiplication process."

— Parkman (1972)

The question Parkman referred to in 1972 is this: Do people have all the solutions to single-digit arithmetic problems stored in their long-term memory and simply retrieve the answer when solving a problem? Or do they use some procedure, such as counting, to arrive at the answer? More importantly, if people use retrieval, how are arithmetic facts represented in long-term memory? Despite decades of research, the question of how arithmetic facts are mentally represented has not been settled. Thus, in this thesis I use modern statistical techniques to shed further light on this issue.

As Parkman suggested in 1972, response time data collected in the subsequent decades have indeed provided conditions against which models of representation could be evaluated. A prominent example of such conditions is the problem-size effect, a pattern that has been repeatedly observed in experiments on mental arithmetic. This pattern can be generally described as such: the bigger the size of numbers in a problem, the more difficult it is to solve it (where difficulty can be defined by differences in speed or accuracy). For example, people solve 9x8 more slowly than 3x4. Response time data have been used to both develop and evaluate models that offer explanations for the problem-size effect. These explanations are equally plausible but make different assumptions about underlying representations and processes, as I will elaborate below. For this field to move forward, each specific model needs to be tested so that they can be revised and synthesized into a more comprehensive theory. An often overlooked aspect

of this model-testing process, however, is whether the statistical model used in data analyses is faithful to the data. Theory building relies heavily on making inferences from empirical findings, and hence, a less-than-satisfactory statistical model may misrepresent raw data and lead to erroneous conclusions.

This dissertation tackled the problem of mental representation from the angle of statistical models applied to data analysis. A faithful analytical tool alone cannot be the only test of a theory. Nevertheless, it is a necessary component to the endeavour. To begin with, in this chapter I review the status quo of theory on mental arithmetic. I first present a group of representative findings related to the problem-size effect. Reviewing them provides a glossary and introduces key concepts that I refer to throughout this document. Next, I review selected theories that are currently active in the field of mental arithmetic. Rather than setting up the expectation that I will reveal the best model by the end of the current dissertation, this part serves to highlight that models co-exist but are not always compatible with each other.

Chapter 1 serves as a backdrop against which the remaining chapters will unfold. In Chapter 2, I review the current practice of analyzing data on mental arithmetic — what statistical models have been used so far in this field, point out shortcomings of these models, and propose an alternative approach. In Chapters 3 through 5, I apply the alternative statistical model to three archival datasets. These datasets contain a diversity of manipulations both of the stimuli — arithmetic problems — and of the arithmetic knowledge owned by individuals, and thus provide me with opportunities to systematically examine problem characteristics such as problem size, problem format, individual differences such as strategy choices, and educational experience as predictors

of latencies. Each dataset features a different combination of these predictors, resulting in data structures of various complexities, and thus allowing me to explore the pros and cons of each statistical model reviewed in Chapter 2. And finally, in Chapter 6, I summarize the findings from previous chapters and return to the evaluation of the theoretical models initiated in Chapter 1.

Part 1: Key Findings in The Field of Mental Arithmetic

Index of problem size. Problem-size effects are ubiquitous in the field of mental arithmetic. They can be operationalized in many ways, but are broadly captured by the description that performance decreases as problem size increases. The different variables that have been used to index problem-size complement each other and each potentially offers a different view of the same phenomenon. One approach is to relate indices of problem size to the hypothesized structure of the mental representation. Within the context of single-digit arithmetic, several structural definitions of problem size exist, all defined in relation to the operands (e.g., 3 and 4 for 3 + 4 or 3 × 4): product (Ashcraft, 1987; Campbell, 1995), squared sum (Stazyk, Ashcraft, & Hamann, 1982), maximum operand (Groen & Parkman, 1972), and sum (Campbell, 1995). These structural indices of problem size are differentially predictive according to samples and tasks; however, only in some models are they tied directly to the mental representations or hypothesized calculation processes (Groen & Parkman, 1972). Table 1-1 shows examples of problem-size indices derived from different structural predictors. In the current research, I focus on two of them: sum and product. I used product to index problem size in both Chapter 3 and 5, so that results in my re-analyses are comparable to the original studies. In Chapter 4, problem size was not used as a predictor in the original analysis. I used sum to index

problem size for addition and product for multiplication. Note that the results remained the same when I used product for both operations. I presented the former in the present thesis.

Note that in this thesis I have focused on two of the four arithmetic operations, that is, addition and multiplication. There were two main reasons for this focus. First, the data sets that I had available and were most suitable for evaluating my alternative statistical model only included these operations. Second, addition and multiplication are sufficiently different, yet similar, to provide an interesting contrast in terms of potential mental representations (e.g., Miller, Perlmutter, & Keating, 1984). A third but less central reason is that the existing models of mental representation were all developed in relation to addition and multiplication and thus my focus is consistent with the way in which the literature has developed (e.g., Ashcraft & Stazyk, 1981; Campbell, 1995).

Table 1-1 *Five Structural Predictors of Problem Size for Selected Example Problems*

Problem	Min operand	Product	Squared sum	Max Operand	Sum
2 + 3	2	6	36	3	5
5 + 4	4	20	81	5	9
7 + 9	7	63	256	9	16
2 x 3	2	6	36	3	5
5 x 4	4	20	81	5	9
7 x 9	7	63	256	9	16

Effects found in erroneous responses. In chronometric studies on mental arithmetic, response times on trials in which participants made calculation errors are typically discarded as uninformative because latencies on these trials are uninterpretable (e.g., Groen & Parkman, 1972; Ashcraft & Battaglia, 1978; Miller et al., 1984). However, the frequency of errors is usually analyzed as evidence of the mental processes in relation

to problem characteristics (Ashcraft, 1992). For example, Campbell and Graham (1985) analyzed the patterns of errors made by people while solving multiplication problems and gained valuable insights, which formed the core of Campbell's network-interference model. Similarly, the neuropsychological models proposed by McCloskey, Caramazza, and Basili (1985) and by Dehaene and Cohen (1997) were largely based on patterns of errors made by individuals with specific types of brain damage. However, latency data are most relevant for my proposed alternative analyses and thus they are the focus of this thesis.

Error rates. On average, participants are more likely to err on large problems than on small problems. *Figure 1-1* exemplifies such a trend in which mean error rates are calculated for multiplication problems across operand families, that is, involving a 1, a 2, and so on through the 9 times-table. Note that the departure from this trend — a low error rate for the 5 times-table — is not due to sampling error but rather is a consistent finding reported in many studies of North American participants (Campbell, 1994; LeFevre, Bisanz, et al., 1996; LeFevre & Liu, 1997). Variability in error rate as a function of problem characteristics other than problem size indicates problem size is not the only relevant independent variable.

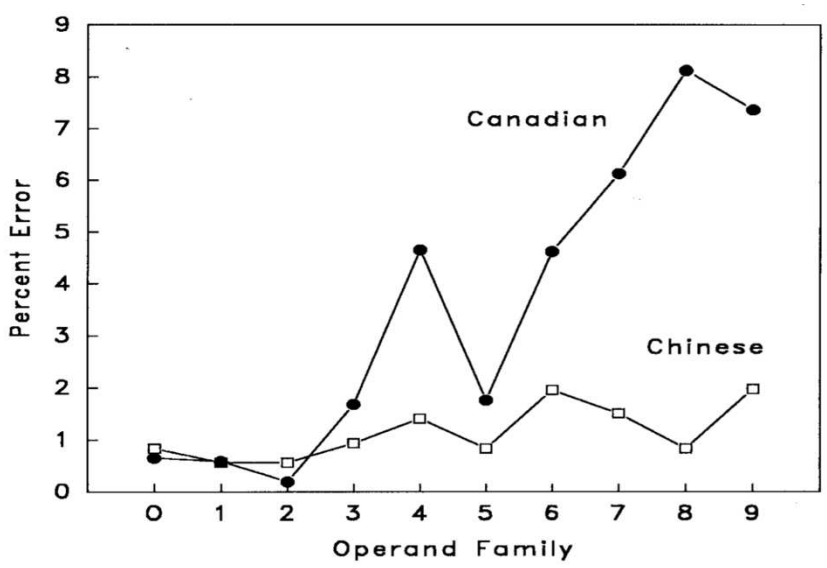

Figure 1-1 Error rate as a function of operand family, from LeFevre and Liu (1997), Figure 1.

Error type. Most addition errors are within +/- 2 of the correct answer (Campbell, 1995). For example, 90% of errors on 7+8=15 come from the set {13, 14, 16, 17}. Similarly, in multiplication, the majority of errors are within the same times table as the correct answer, off by 1 or 2 steps, known as table-related errors. For example, 45 is a common error of 6x9, in which 5x9 = 45 is in the same 9-times table as 6x9 = 54. Other errors are answers to unrelated single-digit problems that are numerically close. For example, a common mistake for 6x9 is 56, which is the correct response for 7x8 (see Table 1-2). Another category is termed "operand intrusion errors" because the answer to a related problem 'intrudes' into the error response, such as 8x*4*=2*4*, which is both table-related and an operand intrusion error. Solvers rarely respond with answers that are not part of the multiplication table. Researchers assume these errors reflect the mental representation of arithmetic facts. For example, Campbell (1995) interpreted operand

errors as evidence that an arithmetic problem is not only associated with its correct answer, but also with numbers that are related to either one of its operands.

Table 1-2 *Types of Calculation Errors on Multiplication Problem for the Example problem, 6x9*

Possible responses	Combination of operands corresponding to the response	Correctness/Type of error
	Table-related errors	
42	**6**x7	from the 6 times table (-2)
45	5x**9**	from the 9 times table (-1)
48	**6**x8	from the 6 times table (-1)
56	7x8	Close in magnitude, shared decade
	Operand Intrusion errors	
63	7x**9**	from the 9 times table (+1); operand intrusion error (6)
	Non-table errors	
59		Intrusion error, operand "9" intruded in the response; 5 is the correct decade digit
69		Intrusion error, both operands "6" and "9" intruded in the response; incorrect decade and unit digit

Effects found in latencies. With a few exceptions, all latency-based findings for the problem-size effect share the basic pattern that response latency increases as problem size increases (see Figure 1-2; Experiment 1 from LeFevre, Bisanz, et al., 1996). However, this pattern is modified by several factors. Next, I introduce some of these factors either because they make up an important part of theories mentioned in the present thesis or they are used as predictors in Chapter 3-5.

Tie. Problems that have identical operands, such as 3+3, are solved faster than non-tie problems (Groen & Parkman, 1972; LeFevre, Shanahan, & DeStefano, 2004;

Miller et al., 1984). In LeFevre, Bisanz et al. (1996), the average latency on tie problems was 1068 ms, compared to 1453 ms on non-tie problems. Furthermore, the slope of the problem size effect for tie problems is much shallower compared to non-ties (Figure 1-2). When latencies are plotted against size as the product of two operands, reaction time increases 3.4 ms per unit increase in the size of the problem for tie problems versus 19.3 ms per unit increase for non-tie problems. This difference in slopes between tie and non-tie problems in relation to problem size raises important questions about the source of the tie advantage (e.g., LeFevre et al., 2004) and thus important in developing models.

Five advantage. The advantage for five-operand problems is specific to multiplication. Multiplying a number by five is faster than solving other non-tie multiplication problems with similar problem sizes, for example, 5x9 versus 6x8 (see Figure 1-2). LeFevre, Bisanz et al. (1996) found that five-operand problems were solved 320 ms faster than non-tie problems on average. As shown in *Figure 1-2*, the slope of the problem-size effect for fives was very similar to that for other non-ties problems in that study, whereas the slope for ties was shallower. Thus, the tie advantage and the five advantage may have different sources, and understanding these differences may be important for models of mental arithmetic.

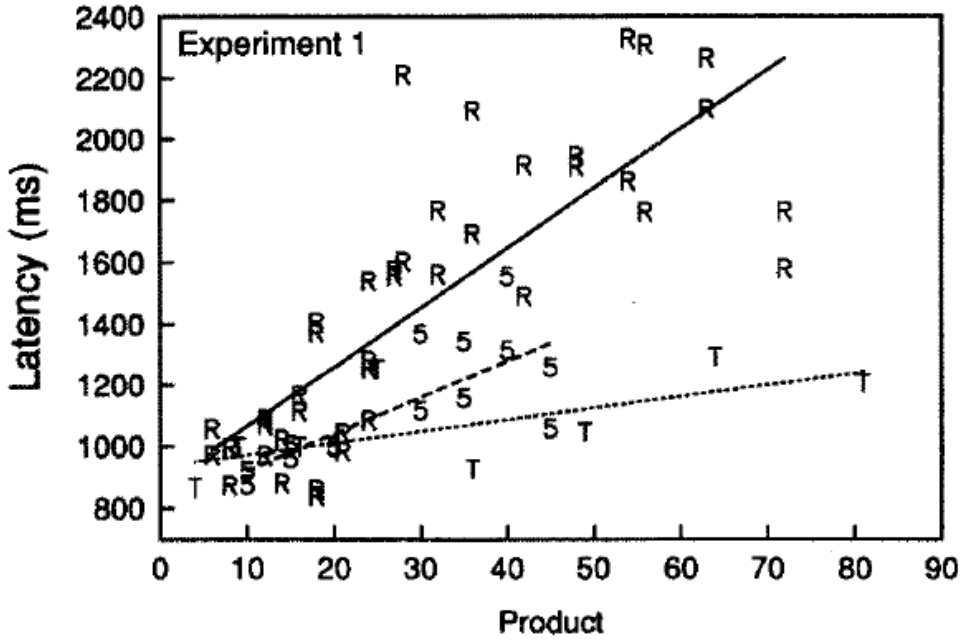

Figure 1-2 Response latencies for ties (T), five-operand (5), and all other problems (R = regular) as a function of the product of the operands (Figure 1 from LeFevre, Bisanz et al., 1996, p. 293).

0's and 1's problems. Any multiplication or addition problem with an operand of 0 or 1 can be solved by rules, hence the name "rule-based problems" (Dagenbach & McCloskey, 1992; Jost, Khader, Burke, Bien, & Rösler, 2009). For example, N x 0 = 0, N x 1 = N, and N + 0 = N can be solved by applying rules or principles such as "anything times 0 is zero". These problems are solved very quickly (650 - 870 ms per item; Fayol & Thevenot, 2012), compared to regular problems such as 5+7, which usually take one second or longer. Moroever, 0's and 1's problems do not show problem-size effects compared to the rest of arithmetic problems (Campbell & Metcalfe, 2007). Speed and accuracy data on these problems provide important conditions against which models can be checked (Campbell & Beech, 2014; Chen & Campbell, 2016). 0's and 1's problems

are not the focus in the present thesis and therefore were excluded from analyses in Chapter 3-5.

Format. Participants solve arithmetic problems in digit form horizontally most often (e.g., 4+5). Problems also are presented in vertical, word, auditory, and pseudohomophone formats (see Table 1-2). Some of these formats are common to daily life, for example, digits in horizontal and vertical form; others only exist in the laboratory. For example, pseudohomophones are non-words that when pronounced sound like words, such as wun, tue, and siks. Psychologists use problems in unusual formats mainly to test hypotheses concerning cognitive representations of arithmetic facts. Using both digit and word formats, for example, Campbell and colleagues (Campbell, 1994, 1999; Campbell & Fugelsang, 2001; Campbell & Alberts, 2009) tried to answer the question "are mental processes of solving arithmetic independent of input format/modality?" Campbell and Alberts (2009) found that word format slows people down, and the deficit associated with word format is worse as problem size increases. They interpreted the result as evidence supporting "modality-specific representations" in arithmetic processing. I listed problem formats commonly seen in the literature in Table 1-3, and flagged those that appeared in the current research.

Table 1-3 *Five Different Formats of the Same Addition Problem 3+5*

Horizontal (Chapter 3 -5)	3 + 5	(LeFevre & Liu, 1997)
Vertical	3 + 5 ?	(Trbovich & LeFevre, 2003)
Aural	/θriː/ /plʌs/ /fʌɪv/	(LeFevre, Lei, Smith-Chant, & Mullins, 2001)
Word (Chapter 5)	three + five	(Campbell, 1994)
Pseudohomophone (Chapter 5)	thrie + fyve	(Pyke & LeFevre, 2009)

Effects found in self-reports. An arithmetic problem can be solved by retrieval, if one has memorized the fact in the past, or by a reconstructive procedure, which invariably involves one or more intermediate steps. For example, the problem 8 + 7 can be solved by retrieving "15" from memory, or by decomposing "7" into 2 and 5, and then adding 8 + 2 + 5. Decomposition is a common strategy for Chinese-educated individuals (Geary, Bow-Thomas, Liu, & Siegler, 1996). Other common non-retrieval procedures include counting (i.e., count up from 8 to solve 8+3) and transformation (i.e., transform 8+8 into 8x2 and retrieve the answer). As problem size increases, so does the percentage of trials solved by procedures, either within a person or across persons (Campbell & Xue, 2001; LeFevre, Bisanz, et al., 1996). Given that non-retrieval procedures generally take longer than retrieval, more frequent use of procedures on larger problems may be one source of the problem-size effect. In LeFevre, Sadesky, and Bisanz (1996) and LeFevre, Bisanz et al. (1996), the strength of the problem-size effect on retrieval-only trials was noticeably reduced for multiplication and became trivial for addition. Their findings supported the hypothesis that the problem-size effect may be, at least in part, an artifact of averaging across latencies with different underlying solution processes.

Summary. In general, problem-size effects may be reflected in one the following three trends as problem size increases: 1) prolonged reaction time, 2) increased error rates, or 3) more frequent usage of time-consuming procedures. Furthermore, there are departures from this monotonic size-efficiency trend, such as tie and five effects. Such complexity motivates theorists to construct formal conceptual models to integrate these diverse phenomena. I will review four models in the next section, all of which were designed to explain how arithmetic facts are represented in human brain.

Part II: Existing Theories of Mental Arithmetic

Each model reviewed in this section includes some account of the problem-size effect. Underlying these models is the assumption that a solution process involves sequential (albeit potentially overlapping) stage of 1) encoding, 2) calculation of an answer, and 3) production of a verbal response. Specifically, encoding refers to the process of parsing information, such as operands in an arithmetic problem, for subsequent conversion to a memory code which can then be stored, retrieved, and operated upon as per task demands. Moreover, the calculation phase could be memory retrieval or another nonretrieval procedure such as counting or derived facts. Another assumption shared by all models reviewed here is that arithmetic facts have been committed to memory by most adults, even though some facts may be susceptible to temporary retrieval failure. Furthermore, these models also assume that when retrieval fails, nonretrieval procedures will be invoked to solve a problem. Models reviewed here differ in their assumptions about the source of variability in the problem-size effect. Based on these assumptions, models can be grouped into two categories: familiarity and interference.

Familiarity-based models. Ashcraft's (Ashcraft, 1987) semantic network model rests on the assumption that frequency of practice is a central factor in forming associations of various strengths between arithmetic problems and their answers. Problems that are practiced more often, that is, those that are more familiar to a person, form stronger associations with their answers and are solved more quickly on average. Ashcraft and Christy (1995) surveyed grade 1-6 textbooks and found that smaller problems were presented more often than large problems. To the extent that textbook frequency is a good approximation of relative frequencies with which children and adults encounter arithmetic problems, this result could explain why solution times slow down as the problem size increases.

Ashcraft's familiarity-based model offers an intuitive explanation for the problem-size effect. However, the model is limited by the difficulty of precisely quantifying frequency with which people are exposed to each arithmetic problem. The most one could say is that people encounter small problems more often than large ones, beyond which more precise predictions are hard to make. Without more precise predictions, it is hard to compare theories. Furthermore, mixed findings on frequency of tie problems in textbooks cast doubt on a familiarity account (Ashcraft & Christy, 1995; Verguts & Fias, 2005). It may be premature to rule out familiarity as a source of the problem-size effect, but concluding that it is the only source of problem-size effect also overly simplifies the matter. Familiarity makes no prediction about error pattern and cannot easily explain why it is more common for solvers to make certain errors but not others. In short, familiarity is insufficient on its own as an account of the problem-size effect.

Interference-based models. The remaining three models to be introduced belong to the interference category. Models in this category share the assumption that retrieval failure or slowdown is due to interference from erroneous associations (e.g., 6x9 may be incorrectly associated with 45, 63, or 56) co-existing with the correct association. These erroneous associations are often called "competing associations" by modellers in this field (Campbell, 1995; Verguts & Fias, 2005). Based on how these erroneous associations are formed, theories can be further grouped by whether interference is acquired or intrinsic. Siegler's ASCM model (1988) belongs to the former, whereas Campbell's (1994, 1995) and Verguts and Fias' models (2005) belong to the latter.

Siegler's ASCM model. The essence of Siegler's model is rooted in his observation that children make occasional mistakes while learning simple addition and multiplication. For example, children may solve 6x9 via repeated addition (i.e., 9+9+…). Sometimes, their computing procedure may generate the correct answer (i.e., 54); occasionally, computation may go awry and generate incorrect answers. For example, repeated addition errors might occur if a child loses count and adds nine only five times (9 + 9 + 9 + 9 + 9 = 45). According to Siegler, these incorrect answers are also committed to memory in the acquisition process, on the assumption that associations form links in memory automatically.

Given that the learning process is error prone, over time a given arithmetic problem may be associated with multiple answers at various strengths, only one of which is the correct one (Siegler, 1988). Association strengths of generated answers on a given problem acquired in the learning process would form a distribution, such as the one shown in *Figure 1-3*. Some answers are more peaked, that is, have stronger associations

with a given problem than others. For example, the largest peak in *Figure 1-3* is to the correct answer, but the answer '56' is also linked to 6 x 9. Practice strengthens the association between an arithmetic problem and its correct answer, and hence the relative strength of this association in proportion to the associative strengths of all answers to the problem (Siegler & Shipley, 1995). The stronger the association between an arithmetic problem and its correct answer *relative* to all possible associations, the more peaked the distribution is. For a retrieved answer to be stated, its relative strength has to exceed the *confidence criterion*, an internal threshold set by individual solvers (Siegler & Shipley, 1995).

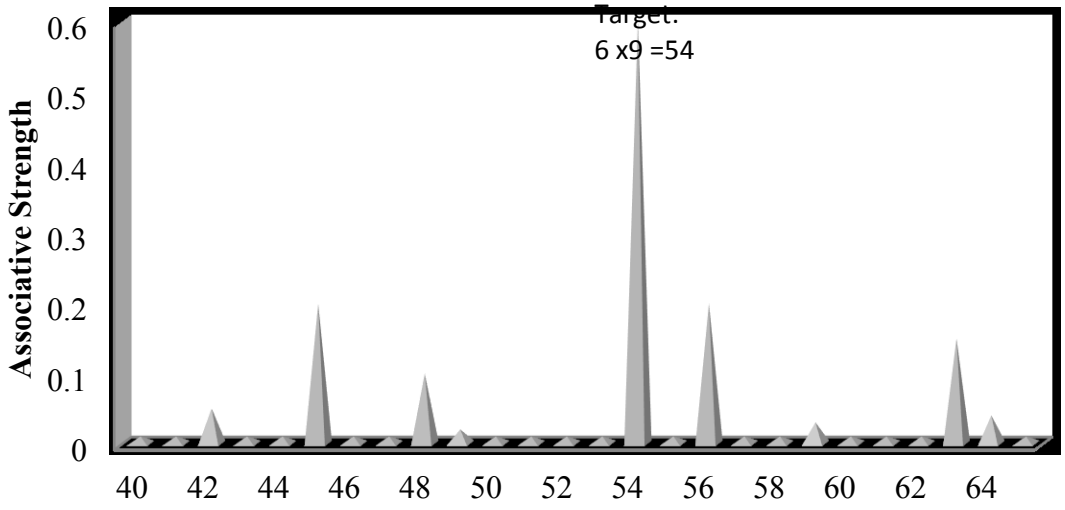

Figure 1-3 Distribution of associations between a multiplication problem and its candidate answers.

Siegler and Shipley (1995) further reasoned that, as problem size increases, over- or under-counting is more likely to occur, which would lead to higher error rates. These erroneous responses, once committed to memory, reduce the relative associative strength between a problem and its correct answer, because the total associative strengths for all

possible responses sum to 1. On this view, large problems tend to have less peaked distributions, that is, flat distributions compared to small problems. Siegler made two more assumptions to explain how peakedness relates to the length of solution times.

1) Retrieval unfolds with an iteration of searches until an answer retrieved on a single attempt surpasses the predetermined confidence criterion or the searching has reached the upper limit of the search length, after which a backup strategy will be used.

2) The probability of retrieving an answer at a retrieval attempt is proportional to its associative strength.

When a problem has a flat distribution of associations, incorrect answers are more likely to be retrieved on a given attempt compared to those in a peaked distribution. These erroneous responses, once they are rejected for not surpassing the confidence criterion, lead to another retrieval attempt. As a result, the correct answer to this problem is less likely to be retrieved on an early attempt, lengthening the solution time. Taken together, Siegler's model accounts for the problem-size effect through the effect of interfering erroneous responses acquired during learning processes.

Campbell's Network Interference Model. Whereas Siegler's model attempted to explain children's performance, Campbell's (1994, 1995) model heavily draws on his analysis of error patterns from adults' mental arithmetic performance. In one study (Campbell & Graham, 1985), Campbell found that ninety percent of the errors adults made on multiplication problems were answers to another basic number combination in multiplication, rather than a random value within the possible range (0-81). Of these errors, multiples of the problems' operands (e.g., 42 for 6x8) far outnumbered the rest of

table errors (79% vs. 14%). Such findings led Campbell to propose the network interference model (Campbell, 1995).

In Campbell's (1995) model, a presented problem (e.g., 3x7) activates other problems similar in magnitude (e.g., 3x6, 4x7, 3x8), due to spreading activation. Campbell introduced Welford's function (1960) as an index of magnitude effects, specifically, log(L/(L-S)), where S stands for the smaller and L for the larger of the magnitudes. This formula captures the finding in the number comparison literature that larger magnitudes are harder to differentiate than smaller ones. Larger problems tend to activate larger candidate answers (e.g., 7x8 -> 72, 64, 63, **56**, 54, 49, 48, etc.) whereas smaller problems activate smaller candidates (e.g., 3x2 -> 2, 3, 4, **6**, 8, 9, etc.). Therefore, it is harder, and takes longer, for solvers to discriminate answers to larger problems from simultaneously activated answers than for small problems, hence the problem-size effect.

To account for the departure of ties and 5-operand problems from the monotonic problem-size effect, Campbell's model assumes they each constitute a special category. He further assumes interference is greater within than between categories. Thus, ties and fives suffer from less interference than nonties as a result of their smaller category membership.

Interacting-Neighbours Model. The notion of interference plays a major role in Campbell's theory, although he regards interference as intrinsic to mental arithmetic, rather than as a result of learning history. Similarly, Verguts and Fias (2005) also used interference as a central explanation feature of their Interacting-Neighbours Model. However, in Verguts and Fias' model, interference is only one way in which neighbouring candidate solutions may interact with the target solution; they can also

facilitate retrieval of the target solution, depending on their corresponding decade and unit parts. If the neighbouring solution has the same unit or decade as the target solution, the former facilitates retrieval of the latter. Otherwise, the neighbour competes with the target and this competition results in increased latency and error rates. For example, both 4x6 and 4x8 are neighbours for 4x7; 4x6 cooperates with 4x7 because their answers have the same decade digit (e.g., 24 and 27) whereas the answer of 4x8 is 32 and therefore competes with the answer to 4x7 (28). The combination of facilitation and interference predicts patterns of solution time.

Along with the assumption that problems with answers that share features compete or cooperate during answer activation, Verguts and Fias' (2005) model added constraints to the set of candidate answers a problem may activate, an assumption that was based on ERP evidence that activation spreads only to table-related problems within a small unit distance (up to two) from the original problem (Niedeggen & Rösler, 1999). For example, 4x7 activates problems only as distant as (4±2)x7 or 4x(7±2).

Verguts and Fias (2005) also assumed that commutative problems are only represented once in a max × min organization. For example, only 7x5=35 is stored in memory according to their model; a participant would transform 5x7 to 7x5 in a pre-processing stage (which was not explicitly described) before solving it. It naturally follows from such a structure that large problems have more inconsistent (i.e., competing) neighbours on average than small problems, and hence are solved more slowly. This escalating competition exacerbated by problem size is less severe for ties, partly because they reside on the diagonal and have fewer neighbours (*Figure 1-4*).

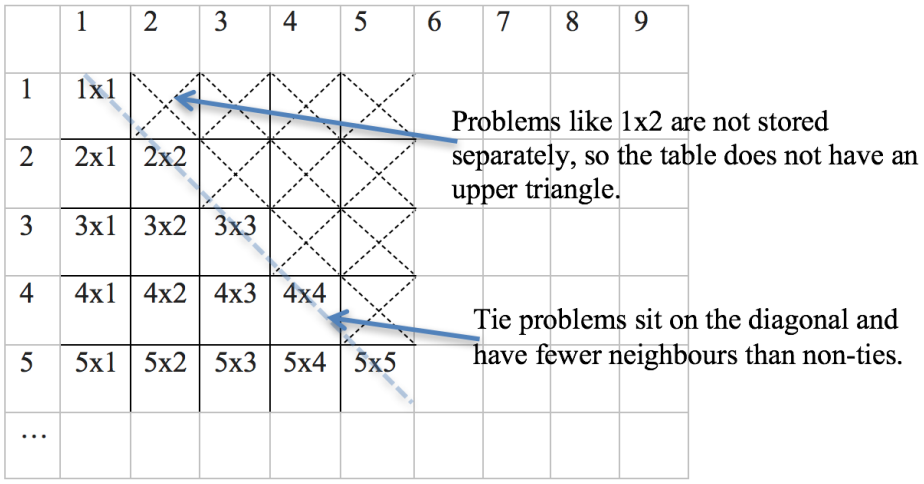

Figure 1-4. The pattern of connections assumed in Verguts and Fias' model.

Summary. Verguts and Fias' model is the most successful among existing models in that it provides a very close simulation of existing patterns of data. This model is the only one that has predicted problem size, tie, and five effects and a size by tie interaction in both latency and error rate without needing to assume separate categories for ties and 5x problems. However, Verguts and Fias acknowledged that their model does not apply to simple addition. Their model is also restricted by its assumption about the structure of mental representations of arithmetic facts, that is, only one of a commuted pair is stored in memory. Even though this assumption may be true for multiplication among Chinese- and possibly European-educated students (Butterworth et al., 2001), there is little evidence to suggest the same assumption would hold true for North American students or for simple addition (but see Butterworth et al., 2001, which assumed one operand order for addition solutions).

This review provides a sketch of the field and highlight the fact that multiple theories of mental arithmetic exist. All models offer plausible accounts of how arithmetic problems and behaviour are connected (i.e., what arithmetic knowledge is present and

how is it organized inside individuals' brains). However, these theories do not always agree with each other and different theories capture only subsets of the existing data (e.g., problem size but not tie effects). In the current thesis, I hypothesized that the use of particular statistical models may have limited researchers' ability to test certain predictions, or to find important patterns within different data sets. Accordingly, in Chapters 3 to 5, I compare statistical models with regards to their fitness to capture the patterns of typical data sets in mental arithmetic. First, in Chapter 2, I discuss the statistical models that have usually been adopted, and describe the advantages related to using an alternative approach, linear mixed modeling.

CHAPTER 2: Methodology

All studies that were cited in Chapter 1 relied on the general linear model (GLM) to explore problem-size, tie, five-operand, and strategy effects in adults' arithmetic (specifically ANOVA or regression). However, researchers have needed to use simplified predictors to accommodate the restrictions of the GLM. In this chapter, I illustrate how the GLM has been historically applied to mental arithmetic data and describe challenges in these analyses. Then I introduce linear mixed modeling (LMM) and describe how it could meet these challenges.

For demonstration purposes, I have created an artificial dataset with two participants and features typical in studies of mental arithmetic (Table 2-1). Each individual solves ten arithmetic problems of incrementing sizes, numbered Size1 to Size10. Response time is the dependent variable (DV) and problem size (Size1 to Size10) the independent variable (IV). Unless otherwise stated, I use problem size and product of two operands interchangeably in the current chapter, because product is the primary index for problem size in my research. Note that, in real datasets, problem size typically has 20-30 different values (e.g., sum or product of two single digits such as 3 and 5). Individuals also reported how they solved each problem: that is, whether they retrieved the answer from memory or via a multi-step procedure such as 8+4 = 8 + 2 + 2 = 10 + 2 = 12.

Table 2-1 *An Artificial Dataset Including Latencies on Ten Arithmetic Problems that varied in Problem Size and Self-Reported Strategies*

Partici-pant	Size1	Size2	Size3	Size4	Size5	Size6	Size7	Size8	Size9	Size10
01	1010_O	1001_M	923_M	1007_M	897_M	1007_M	950_M	1088_O	1016_M	899_M
02	998_M	1010_M	1092_O	992_M	846_M	896_M	817_M	1044_O	1091_O	1072_O

Note. Each reaction time is tagged with self-reported strategies to indicate the mental process underlying each solution. M = memory retrieval, O = other (i.e., non-retrieval). Non-retrieval trials are highlighted in grey.

The first part of this chapter reviews the current practice of data analysis in this field. The second part provides an overview of the statistical techniques used in following chapters.

Repeated Measures ANOVA (RM-ANOVA)

Researchers who use RM-ANOVA in their data analysis typically transform problem size, a continuous within-subject factor, into a variable with a limited number of levels, typically two, but sometimes three or four (e.g., Ashcraft & Stazyk, 1981; Thevenot, Fanget, & Fayol, 2007). As will be described later, this simplification reflects a trade-off between ease of interpretation and statistical power (Maxwell & Delaney, 1993). Regardless of how many categories the continuous variable of problem size is turned into, the ramifications remain the same. In the current thesis, I will focus on scenarios in which problem size has been dichotomized, but I acknowledge that it is not the only way problem size has been treated in the literature.

To illustrate the effect of dichotomizing a continuous variable, I split the problem size in the artificial dataset shown in Table 2-1 into small (size 1 - 5) and large (size 6 -

10), as has been done in most studies of this field which applied ANOVA to this type of dataset. *Figure 2-1* shows the data aggregation process.

Participant	Size1	Size2	Size3	Size4	Size5	Size6	Size7	Size8	Size9	Size10
01	1010O	1001M	9[illegible]3M	1007M	897 M	1007 M	950 M	10[illegible]8O	1016M	899M
02	998M	1010M	10[illegible]92O	992M	846M	896M	817M	10[illegible]4O	1091O	1072O

Note. Each reaction time is tagged with a self-reported strategy to indicate the mental process underlying each solution. M = memory retrieval, O = other (i.e., non-retrieval). Non-retrieval trials are highlighted in grey.

Participant ID	Small	Large
01	1007	1115
02	1003	874

Figure 2-1 Aggregating raw data (latencies) over problem size before applying RM-ANOVA.

Applying a RM-ANOVA model to the aggregated data, response latency is decomposed into the grand mean (β_0), deviation from the grand mean for participant *i* (S_{0i}), treatment effect of problem size, and residual e_t:

$$RT_{ki} = \beta_0 + S_{0i} + \beta_1 Size_k + e_{ki} \quad (2.1)$$

RT_{ki}: Response latency (DV) for subject *i* when problem size is k, aggregated from raw data.

β_0: conditional mean response latency for small problems.

S_{0i}: deviation from conditional mean for participant i.

β_1: average difference in response latency between small and large problems.

$Size_k$: Dichotomous IV transformed from the originally continuous problem size, comprising two possible values, 0 = small, 1 = large.

e_{ki}: residual associated with the observation for participant j when problem size = k; this term includes observation-level error and potential interaction between subject and problem size.

Next, I explain two major shortcomings of RM-ANOVA when applied to mental arithmetic data, namely number of random effects permitted in a model and dichotomizing continuous predictors.

Fixed- vs. random-effects. Equation 2.1 captures some individual differences by allowing participants to have different overall response latencies, namely S_{0i}. These values are assumed to be *random effects* in that another replication of the same experiment with a different sample of participants would presumably generate a different set of numbers, each representing the deviation of an individual participant's overall latency from the grand mean.

However, beyond a randomly varying variable that represents individuals' overall latencies, a RM-ANOVA does not allow for additional random effects. For example, the slope of problem size β_1 in Equation 2.1 is assumed to be constant from one replication to another of the same experiment, namely, it is assumed to be a *fixed effect*. Individual deviations from this value are considered random error and are included in the residual variance.

By definition, Equation 2.1 describes a mixed-effects model because it has both fixed and random effects. However, the term "mixed model" is often used interchangeably with LMM, as in "linear mixed-effects model" (LMM, Barr, Levy,

Scheepers, & Tily, 2013), whereas a RM-ANOVA model such as (2.1) is far less frequently labeled as a mixed model in the literature. Because the 'mixed' nature of the RM-ANOVA is not emphasized in the naming convention, users of these two techniques may be given the impression that LMM differs from RM-ANOVA because one is a mixed-effects model and the other is not. Instead, a central difference is that LMM allows for more random effects (slopes as well as intercepts), whereas in RM-ANOVA, only one random effect is allowed.

Continuous vs. dichotomous problem size. In RM-ANOVA, problem size is frequently transformed into a dichotomous variable by dividing the set of problems into two groups with an equal number of stimuli in each group. This approach equates to assuming that all problem sizes in one category (e.g., small) are interchangeable and that two adjacent problem sizes that flank the variable boundary are categorically different. Occasionally, categorical divisions of problem size were selected to reflect functional differences, as when small addition problems are defined as those with sums of 10 or less, because, on average, problems with sums of 10 or less are solved differently than those with sum greater than 10 (LeFevre, Sadesky, & Bisanz, 1996). In other situations, however, as when problem size is defined in multiplication such that small problems have products of 25 or less, the distinction is based on dividing the set of problems into two groups with the same number of problems per group (e.g., Campbell, 1994, 1995). As a consequence of dichotomizing problem size, variability in latencies due to systematic differences among problems within the same category (e.g., small) is ignored prior to analysis. Such aggregation could reduce power and inflate Type I error (Maxwell & Delaney, 1993).

Despite the limitations of dichotomizing the problem size variable, this approach has the benefit of retaining participants who would otherwise have missing cells if the unit of aggregation was smaller and so fewer data points were available per category. For example, in studies where self-reports were collected on how participants solved each arithmetic problem, researchers examined the relations between problem size and latency/accuracy data on retrieval trials only (LeFevre, Bisanz et al., 1996; LeFevre, Sadesky & Bisanz, 1996). However, as is clear from *Figure 2-2*, once non-retrieval trials are removed from the data, some cells are empty and thus, data points become unevenly distributed across individuals. Applying RM-ANOVA to this type of dataset would become problematic because any individual with missing cells would be deleted from the analysis, undermining its power. One way to circumvent this problem, as practiced by many researchers in this field, is to dichotomize single-digit problems into small and large groups (Campbell, 1994; Campbell & Alberts, 2009; Noël, Fias, & Brysbaert, 1997). This approach reduces the risk of losing participants who use non-retrieval occasionally (bottom of *Figure 2-2*). However, this way of analyzing such data only "masks" the problem of an unequal number of data points going into each cell mean for each individual (Hoffman & Rovine, 2007, p. 106)

Participant ID	Size1	Size2	Size3	Size4	Size5	Size6	Size7	Size8	Size9	Size10
01	*NA*	1001_M	$92\_{}_M$	1007_M	897_M	1007_M	950_M	*NA*	1016_M	899_M
02	998_M	1010_M	*NA*	992_M	846_M	896_M	817_M	*NA*	*NA*	*NA*

M = memory retrieval. NA = not available. Reaction time on these trials are excluded because they are solved by non-retrieval

Participant ID	Small	Large
01	1007	1115
02	1003	874

Figure 2-2 Data aggregation to overcome the problem of missing cells in ANOVA.

Dichotomizing problem size has another benefit: It eases the interpretation when a second variable interacts with problem size. As explained by Ashcraft and Stazyk (1981), "the separation of problems into small vs. large is admittedly somewhat crude …[the advantage of doing it is that] information concerning interactions between problem size and other factors is more conveniently represented and understood in the analysis of variance framework …" (p. 187). This rationale is illustrated in Campbell and Alberts (2009), in which they examined the effects of different formats (digit vs. word) on people's mental arithmetic performance. They found that format interacted with problem size such that the problem-size effect was more prominent on word problems than on digit problems. Being able to extract interpretable patterns from data may explain why Campbell and colleagues have consistently used RM-ANOVA in over fifty published articles, even though his theory attributed the problem-size effect to an essentially continuous variable – the magnitude of the sum/or product (Campbell, 1987; Campbell & Oliphant, 1992; Campbell, 1995).

Ashcraft and Stazyk (1981) were accurate in categorizing a small versus large divide as crude but convenient. Due perhaps to the lack of progress in statistical methods for dealing with continuous and categorical variables simultaneously, this practice persisted among later researchers and rarely did people question the practice (cf. LeFevre Bisanz et al., 1996; LeFevre & Liu, 1997). Despite its success in easing interpretation and circumventing loss of participants when analyses were conducted in which retrieval and non-retrieval trials were separated, dichotomizing an inherently continuous variable has statistical and theoretical consequences.

As I have illustrated, when researchers model behavioural data separately for each type of underlying mental process with RM-ANOVA, it is almost inevitable that problem-size will be reduced to a categorical, usually dichotomous, variable. Clearly, issues of dichotomizing and use of RM-ANOVA methods are heavily intertwined.

Multiple Regression

The same artificial dataset can also be used to illustrate an analysis using multiple regression. The relationship between problem size and latency can be modeled as:

$$RT_{ki} = \beta_0 + \beta_1 Size_k + e_{ki} \quad (2.2)$$

This model implicitly assumes all observations are independent, $e_{ki} \sim N(0, \sigma^2)$, iid, whereas repeated observations collected from the same person are likely correlated to a greater degree than observations collected from different people. Hence, applying the model to raw within-subject data erroneously assumes that more information is available than is warranted, because correlated observations reduce the effective sample size to a smaller number than the actual number of observations available. In practice, therefore,

data are aggregated across participants before further analysis (e.g., Ashcraft & Stazyk, 1981; Groen & Parkman, 1972; LeFevre, Sadesky, et al., 1996). The RT in Equation 2.2 represents mean (or median) latencies aggregated across participants for each problem-size value. Although such aggregation prevents inflated Type I error, it also lowers the efficiency and power of the analysis.

Multiple regression collapsed across participants was mostly used in the early stage of research in this field, particularly in the 1980s (e.g., Ashcraft & Stazyk, 1981; Miller et al., 1984). One reason researchers might prefer regression is because one can enter multiple predictors and thus rule out less predictive ones, using changes in R^2 to index the quality of predictors. At that time, theories were being developed about why and how problem size relates to latency. Different possible structural predictors, structural in the sense of capturing the relations among the stimuli, were entertained as indices of mental representation, including operand size, answer size (e.g., sum, product), answer size squared, minimum operand, maximum operand, and so on (LeFevre, Sadesky, & Bisanz, 1996; Miller et al., 1984). The shift from regression to a focus on RM-ANOVA occurred gradually during the 1980s. One possible reason is that an early study (Ashcraft & Stazyk, 1981) generated a lot of interest and was replicated by other researchers who used similar ways of analyzing the data. However, relatively few researchers continued to use that approach, after demonstrations of alternative techniques (see below; LeFevre & Liu, 1997; LeFevre, Sadesky, & Bisanz, 1996) that were based on an influential paper which made it clear why regression averaging across people was not suitable in cognitive studies that used repeated measures designs (Lorch & Myers, 1990).

By-Person Regression

Because regular regression aggregates data across people and completely neglects individual differences, some researchers have run separate regressions on each individual's latencies against problem size and used individual regression estimates as data in a subsequent between-subjects analysis (LeFevre, Sadesky, et al., 1996; De Visscher & Noël, 2014; described for language data by Lorch & Myers, 1990). Although by-person regression takes into account both item-level and person-level variation and thus is a significant improvement over regular regression, it is unwieldy and may be biased. Specifically, regression estimates for each individual have different reliabilities, which might lead to biases in unknown directions (Hoffman & Rovine, 2007). In a balanced design, however, its result may be very similar to LMM. In Chapter 3, I compare results from applying LMM and by-person regression to the same dataset.

Linear Mixed Modelling (LMM)

For the artificial dataset, consider the relationship between the trial-level predictor, problem size (i.e., 1 through 10), and the trial-level outcome variable, response time, for a single participant. These 10 data points can be summarized in a regression equation such as Equation 2.3:

$$RT_t = \beta_0 + \beta_1 Size_t + r_t \quad (2.3)$$

where RT_t is the observed response time on trial t, β_0 is the intercept, or expected response latency when problem size is zero, β_1 is the slope, or expected rate of change in response latency associated per one unit increase in problem size for trial t, and r_t is the error term, which represents unexplained variability in response latency on trial i. Errors are assumed to be normally distributed, $r_t \sim N(0, \sigma^2)$, iid.

Now scale Equation 2.3 up to reflect a random sample of *i* participants from the entire population. A total of *i* regression equations can be written, one for each individual. To represent them succinctly, I use subscript *i* to differentiate these equations:

$$RT_{ti} = \beta_{0i} + \beta_{1i}Size_{ti} + r_{ti} \quad (2.4)$$

where r_{ti} is still assumed to be normally distributed with homogeneous variance across individuals, $r_{ti} \sim N(0, \sigma^2)$. β_{0i} and β_{1i} represent variables which have values that vary over individuals.

Unlike Equation 2.3 in which coefficients β_0 and $\beta_1$were fixed effects, coefficients β_{0i} and β_{1i} in Equation 2.4 are now allowed to differ for each individual *i* in the population. Estimates of these coefficients for each individual, that is, β_{0i} and β_{1i}, are then aggregated to find the population estimates using the following equations:

$$\beta_{0i} = \gamma_{00} + u_{0i} \quad (2.5)$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \quad (2.6)$$

Substituting the expressions for β_{0i} and β_{1i} in Equations 2.5 and 2.6 into Equation 2.7 yields a combined model:

$$RT_{ti} = \gamma_{00} + \gamma_{10}Size_{ti} + u_{0i} + u_{1i}Size_{ti} + r_{ti} \quad (2.7)$$

γ_{00}, estimated conditional mean response latency when problem size is zero

γ_{10}, estimated mean rate of change in response latency for every unit increase in problem size

u_{0i}, deviation of person i from the conditional mean response latency

u_{1i}, deviation of person i's slope from the mean association between problem size and response latency (gamma10).

r_{ti}, deviation of person i on trial t from person i's own mean response latency, or otherwise unexplained variability on a given trial.

Both the hierarchical form (Equations 2.5 to 2.6) and combined form, Equation 2.7, are widely used. They highlight different aspects of a LMM. I will use these forms of the model interchangeably as convenience dictates.

Note that in Equation 2.4, trial-level outcomes were regressed on trial-level predictors. In contrast, in Equations 2.5 and 2.6, individual intercepts/slopes were treated as outcomes, and can be regressed on some individual-level predictors (not shown in the example above). Together, they form a two-level LMM. This feature is the main reason why the technique is also referred to as *hierarchical* linear model or *multilevel* modelling: variables are introduced into the model at different "levels" to explain variability. In the current example, Level-1, the lowest level, refers to the trial-level (Equation 2.4), whereas Level-2 refers to the individual level (Equations 2.5 and 2.6). In experiments on mental arithmetic, typical Level-1 variables include problem size, tie status, and 5-operand status, and represent trial-to-trial differences within individuals , whereas level-2 variables could be gender, math skill, or training condition, representing differences between individuals.

In Equation 2.5, individual estimates of coefficients are conceived as being *randomly* sampled from a population. The statistics β_{0i} and β_{1i} are commonly described as *randomly varying*; a defining feature of LMM. Intuitively, allowing the model to

include a random *intercept* (β_{0i}) acknowledges that some people are generally faster in solving arithmetic problems than others. Likewise, the random *slope* in Equation 2.6 (β_{1i}) captures individual differences such that some participants may exhibit steeper increases in RT from small to large problem size than others. Moreover, LMM assumes these individual differences (i.e., variance) existing in the data are a sample of the population, a reasonable assumption, and therefore the model estimates the population variance by adjusting the magnitude of sample variance to account for the fact that sample estimates tend to be smaller than their population counterparts. Such discrepancy between sample and population estimates is overlooked in GLM. As a consequence, standard errors tend to be smaller (underestimated) in GLM compared to LMM.

The need to incorporate random effects beyond random intercepts, so that findings from one study can be generalized, has long been recognized. As early as the 1970s, psycholinguists recognized that stimuli used in a study were almost always a sample of a larger pool of stimuli, and yet conclusions were generalized to all the stimuli (H. H. Clark, 1973). At the time, it was computationally challenging to include both random intercepts for participants and items in a single analysis. Researchers can either aggregate data across items and conduct a RM-ANOVA on conditional means for each subject (F_1) with a random intercept for subjects, or aggregate data across subjects and conduct a RM-ANOVA on conditional means for each item (F_2) including a random intercept for items. Neither of them fully capture variances in the outcome, however. In a seminal paper, Clark (1973) proposed F_{min}, an index derived from F_1 and F_2, which allows testing of treatment effects while accounting for randomly sampled participants and items.

The computational challenge of simultaneously testing more than one random effect was overcome with iterative numerical procedures developed by mathematicians in late 1970s (Raudenbush & Bryk, 2002, p. 51). Early proponents of LMM adopted these algorithms and formalized approaches to parameter estimation. These algorithms usually involve maximum likelihood: Choosing estimates of parameters (mean, variance and covariance) for which the likelihood of observing the actual data is maximal. Estimation is done through an iterative process, starting with some reasonable, but inaccurate, guesses. After tens, hundreds, and or thousands of iterations, estimation converges to a stable value, which is deemed the best guess of the parameter in the population from which the sample was drawn.

A prominent advantage of LMM, therefore, is the flexibility to include random effects—as many as the experiment design dictates—although in practice model estimation becomes intractable as the number of random effects in a model increases. In a within-subjects design with three factors A, B, and C, for example, random slopes for all factors could be tested along with the random intercept. Much discussion has been generated on how many random effects should be included in a given analysis (Barr et al., 2013; T. S. Clark & Linzer, 2015; Gelman, 2005). As the number of random effects included in a model increases, it becomes less likely that the estimation algorithm will converge on stable values. Therefore, a theoretically sound model may not be testable. In this thesis, I followed the recommendation of Barr et al. (2013) to use maximal random effects structure, the upper limit number of random effects justified by the experiment design. In the next section, I describe these guidelines for setting up the structure of

random effects of a model. In Chapters 3-5, I implement them and describe challenges in following these guidelines, as well as tips on what to do when models fail to converge.

Maximal random effects structure. In general, a by-participant intercept should be random whenever there is more than one observation per subject. A by-item intercept should be random whenever there is more than one observation per item/stimulus. And a random slope for any effect is recommended where there is more than one observation for each unique combination of subject and treatment level (Barr, 2013, p.262).

The requirement for a random by-subject intercept should be obvious, given almost all experiments on mental arithmetic done to date use the repeated-measure design. It is less clear whether a random by-item intercept is necessary. A repeated-measure design, by definition, would have each participant solve all arithmetic problems. Even after data trimming, the majority of items (arithmetic problems) will still have multiple observations each, creating non-independence among data. One way to account for such non-independence is to include a random by-item intercept. Similar to a random by-*subject* intercept which accounts for non-independence among observations from a given individual and between-person variability, a random by-item intercept acknowledges that observations made on the same item/stimulus are dependent because characteristics of the item that affect one observation would presumably also affect another. Furthermore, a random by-item intercept accounts for between-item variability, above and beyond what has already been accounted for by item-related predictors such as problem size and problem type.

Equation 2.8 demonstrates a LMM with a maximal random-effects structure. Compared to Equation 2.7, it contains all the previous random and fixed effects, with an

additional random by-item intercept, designated by "I_{0i}". This model is also known as cross-classified or crossed random effects model (Hoffman & Rovine, 2007).

$$RT_{ti} = \gamma_{00} + \gamma_{10}Size_{ti} + u_{0i} + u_{1i}Size_{ti} + I_{0i} + r_{ti} \quad (2.8)$$

Although including a random by-item intercept accounts for dependency among data, doing so also assumes that the set of stimuli used in a study is a subset of a much larger, sometimes infinite, number of stimuli. For example, researchers in linguistics who use words as stimuli would need to include a random by-item intercept in the analysis whenever the goal is to generalize findings to the entire population of words even though only a small number of representative words are used in the study for practicality. A random by-item intercept accomplishes both goals simultaneously: accounting for dependency and allows generalization to items not tested in a study. Contrary to such studies on language, however, research on single-digit mental arithmetic often use the entire set of arithmetic problems that exist. Including a by-item intercept may overestimate between-item variability: Estimates of population parameters are usually slightly larger than the sample statistics to compensate for the fact that sample variance is usually smaller than that of the population. A second option exists which would model between-item variability without the risk of inflating it: create k-1 dummy variables to represent all items in the experiment, where k is the number of items. However, this option is too onerous in practice. I chose to include the random by-item intercept in this dissertation and err on the cautious side.

Lastly, almost all item-related predictors, including problem size and problem type, should have random slopes. For example, a person who solves all single-digit

multiplication problems twice, 2x2 through 9x9, would have 2 or more trials per level of problem size, that is, multiple trials per combination of subject and treatment level. Similarly, this person would contribute 20 or more trials per level of problem type, a categorical variable. Essentially, inclusion of random slopes recognizes idiosyncrasy, that is, different people respond to different levels of a factor differently. One person may experience a large increase in RT as problem size increases whereas another person may respond equally quickly to a variety of problems, or the RT difference between tie and non-tie problems may be trivial for one person but substantial for another.

I have just introduced LMM as an alternative to analyze repeated measure data on mental arithmetic. In summary, LMM can properly model the inter-related nature of data typical for research in mental arithmetic. It can also avoid over-simplifying variables, suitable for datasets with a mix of continuous and categorical predictors. Next, I introduce the piece-wise model which would allow researchers to test the possibility that problem size has a non-linear relationship to RT.

Piecewise regression

Piecewise regression, more commonly known as “spline regression", is a technique for modelling a non-linear relationship between a continuous predictor and an outcome, with two or more segments of connected lines (*Figure 2-3*). Although this modelling technique is applicable beyond least square regressions, and I will apply it to a LMM in Chapter 3, it is best understood when introduced in a least squares regression.

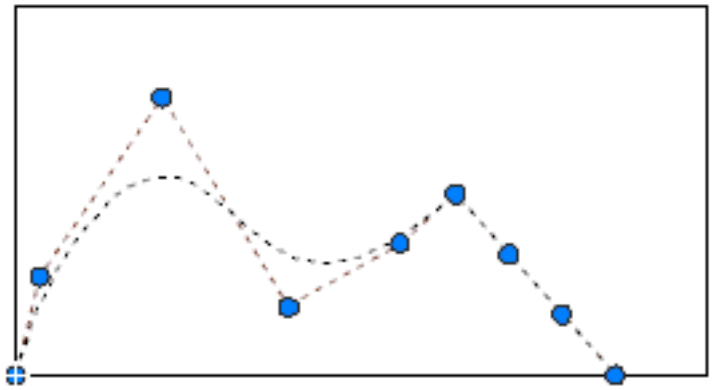

Figure 2-3. A spline regression approximating a curvilinear relationship with 8 connected linear segments.

For the same artificial dataset I have been using throughout this chapter, consider a non-linear relationship between problem size and response time. Suppose past research suggests that slope might be different for problems with size smaller than 5 and problems larger than 5, such that the rate of change in response latency slows down past problems with size larger than 5. Instead of a straight line, the hypothetical relationship between problem size and response latency would have a kink at size = 5 (*Figure 2-4*). To test this hypothesis, I construct a model that maps onto the hypothetical shape and examine its coefficients and fit.

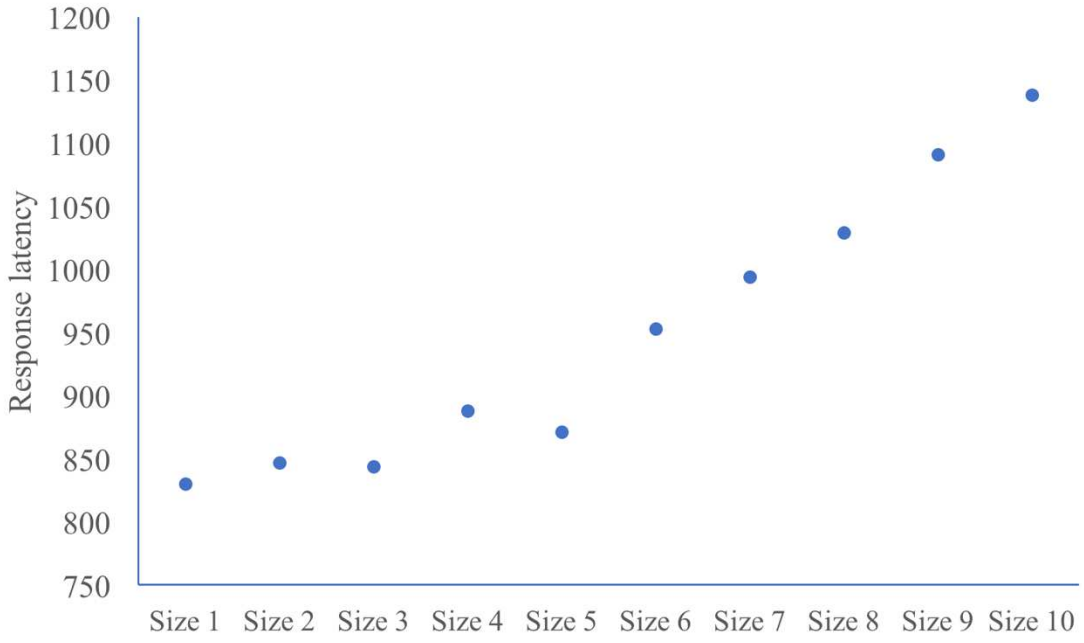

Figure 2-4 A hypothetical relationship between problem size and response latency with a kink at size = 5.

I introduce a dummy variable D to a simple regression in which response latency is a function of problem size. The equation for the dummy regression is:

$$\overline{RT_i} = \beta_0 + \beta_1 Size + \beta_2 (Size - 5) \times D \quad (2.9)$$

D = 0 when size < 5, D=1 when size > 5.

Equation 2.9 amounts to modelling two linear segments, one for problems smaller than 5 and one for problems larger than 5. This model allows different trajectories of latency against problem size.

When D = 0 (size < 5), $\overline{RT_i} = \beta_0 + \beta_1 Size$

When D = 1 (size > 5), $\overline{RT_i} = \beta_0 + \beta_1 Size + \beta_2 (Size - 5) = (\beta_0 - 5\beta_2) + (\beta_1 + \beta_2) Size$

When size = 5, the first equation predicts $\overline{RT} = \beta_0 + 5\beta_1$, and the second equation also predicts $\overline{RT} = (\beta_0 - 5\beta_2) + 5(\beta_1 + \beta_2) = \beta_0 + 5\beta_1$. The two segments would join, that is, have the same predicted values at size = 5. Therefore, a piecewise model is also known as a dummy regression with continuity restrictions. Because of this smoothing/continuity constraint, piecewise regression is a member of the spline model family, named after the tool "used by shipbuilders and drafters to construct smooth shapes having desired properties." (Racine, 2017) The term "knots" refers to transition points in a piecewise model where line segments meet each other, such as at size = 5 in the previous example. Piecewise regression is an alternative to assuming linear slopes and may allow researchers to test specific hypotheses about combinations of slopes for continuous variables.

Summary

In this chapter, I introduced three modelling techniques that have historically been applied to data on mental arithmetic. I also described problems in using these techniques and how an alternative, LMM, could potentially overcome these problems. In the following three chapters, I will use modelling techniques introduced in this chapter and compare their pros and cons when applied to each dataset. To end this chapter, I describe the general approach I adopted in Chapter 3-5.

Approach to Analyses in Chapters 3 to 5

In Chapters 3 to 5, I describe the results of re-analyses of three archival data sets, one per chapter. Each analysis started with data cleaning and organizing that minimially affected the actual data, including importing the data file into R, renaming variables, and creating dummy variables. Next, I ran diagnostic analyses on the data. Diagnostic

analyses for LMM are similar to those for ANOVA/regression in that they also revolve around residuals. Therefore, prior to diagnostic evaluation, I decided what models were appropriate based on the research question. This step included making a decision about the structure of random effects. It required some trial and error during preliminary analyses, especially if the model included four or more predictors.

Depending on the results from the diagnostic analysis, I proceeded with the full or a reduced sample and conducted the analysis proper. Afterwards, I compared results from using LMM either to the original analysis with GLM (Chapter 3 and 4) or to a comparable sample analyzed with GLM (Chapter 5). I also discussed pros and cons of each method based on these results. I conducted all analyses in R (version 3.3.3), an open-source language and environment for statistical computing (R development core team, 2017), freely available in the CRAN archive (http:// cran.r-project.org).

CHAPTER 3: Re-analysis of LeFevre and Liu (1997)

Introduction

In this chapter, I compare analyses using LMM to those of a by-participant regression for a dataset first published by LeFevre and Liu (1997). To understand how educational experience is related to numerical skills, LeFevre and Liu asked Chinese and Canadian adults to solve single-digit multiplication problems and measured their speed and accuracy. Results revealed distinct response time and accuracy patterns between the two cultural groups in both speed and accuracy as a function of problem size and problem categories. Specifically, they found that overall, participants solved tie and five-operand problems faster than regular problems and that Chinese-educated participants were faster than Canadians. However, comparisons across groups showed that only Canadian participants displayed an advantage on five-operand problems and the problem-size effect was smaller for Chinese than for Canadian participants. Furthermore, the slope of the interaction between problem size and tie status was greater for Canadian than for Chinese participants. Thus, this data set illustrates all of the central findings for single-digit multiplication, but they are qualified by educational experience (i.e., Chinese versus Canadian).

I hypothesized that the standard LMM analysis and the by-participant regression would show very similar results because the current study was a balanced design with equal number of participants in each culture group and each participant completing the same trials (Baayen, Davidson, & Bates, 2008). Both LMM and by-participant regression permit analysis of within- and between-person variance without dichotomizing problem size into small and large. Although by-participant regression risks inflating Type I error

when there is significant item variability (i.e., participants receive different sets of stimuli; Baayen, Davidson, & Bates, 2008), in the current dataset participants responded to the same set of problems and so this risk was not an issue. Subsequent to the comparison of the results of the LMM and by-participant regression, inspection of the data suggested deviations from linearity in the pattern of problem-size effects. Accordingly, I used a piece-wise approach with LMM to provide a more precise fit to the data.

Method

Description of data

Twenty Chinese and 20 Canadian participants solved 300 multiplication problems each. Problem types included 0-operand (e.g., 3x0), 1-operand (e.g., 6x1), 5-operand (e.g., 5x7), ties (e.g., equal operands such as 4x4), and regular (any problem that is not five or tie, nor has 0 or 1 in the operands). Trials on which participants made errors, 0-operand, and 1-operand problems were excluded from the current analysis. As a consequence, each person effectively contributed 192 data points at most, ranging from 128 to 192 each and totalling 7343 trials.

Analysis Plan

I applied three statistical approaches to the dataset. In all three analyses, problem type was treated as a categorical variable and coded with two dummy variables, TIEvREG (hereafter referred to as ties) and FIVEvREG (referred to as fives), in which regular problems served as the reference group. Table 3-1 illustrates my coding scheme. Another predictor, problem size (referred to as size), was defined as the product of the

operands for each problem, and thus is a continuous variable ranging from 4 (2x2) to 81 (9x9). Next, I explain analytical strategies specific to each approach.

Table 3-1 *Coding scheme of problem type, a categorical variable with three levels*

	TIEvREG (ties)	FIVEvREG (fives)
Tie	1	0
Five	0	1
Regular	0	0

Multilevel Modelling. I applied a two-level linear model to the dataset, in which level-1 units consisted of the repeated measures for each individual, and the level-2 unit was the individual. At level-1, the response time of each individual problem is a function of its size, type (ties, fives) and interactions among them. At level-2, the person's educational background (i.e., 1 = Chinese; 0 = Canadian, labeled culture) was a predictor for the intercept and all slopes. This level-2 predictor was coded as a dummy variable in which Canadian served as the reference group.

By-participant regression. Latency was regressed on problem size and type for each individual; regression coefficients were then compared across culture groups (i.e., Canadian vs. Chinese) using t-tests. This "slopes-as-outcomes" approach was adopted in the original study (LeFevre & Liu, 1997).

Piece-wise LMM. In this variation of the basic LMM, the problem-size effect was allowed to have different slopes for small (size ≤ 25) versus large problems (size > 25).

Results

Data preparation

This section is organized by two questions: 1) how many random effects should be included in a LMM model and, 2) whether the data met assumptions required by LMMs.

Intercept, problem size, and both dummy coded variables for problem type (ties, fives) are candidates to be included as random effects in a LMM model, to address the possibility that mean response time and the effects of a predictor on response time may vary across individuals. I examined spaghetti plots of both raw and predicted latencies against 1) problem size (size), 2) ties (i.e., the first dummy variable for problem type contrasting tie and regular problems), and 3) fives (i.e., the second dummy variable for problem type contrasting five and regular problems). These plots show general associations between variables separately for each person, overlaid on a single figure, and are useful in assessing whether relations between predictor and outcome variables are similar or different across participants. Plots of latencies regressed on each of the three predictors suggested that the intercept and all three predictor slopes should be modeled as random effects, because of a high degree of variability between persons.

A histogram of the dependent variable, response time, revealed violations of normality. Accordingly, I applied a log-transformation to the variable which visually improved its normality. I also examined other assumptions including heteroscedasity and linearity and conducted sensitivity analysis with potential outliers. I ran parallel analyses with and without these potential outliers. The results remained the same before and after

each potential outlier was excluded. Therefore, I proceeded to conduct the analysis with the full sample.

Multilevel Modelling

Models to be tested. I tested two parallel models, one with raw RT as the dependent variable (hereafter referred to as the rawRT model, see Equation 3.1) and one with the log-transformed RT (hereafter referred to as the logRT model, see Equation 3.2). Based on common sense, if two models are not meaningfully different (i.e., coefficients significant in one model are also significant in another), then I present results of the rawRT model because it affords intuitive interpretation. Otherwise, the model with the log-transformed variable will be interpreted.

Level 1:

$$RT_{ti} = \beta_{0i} + \beta_{1i}Size_{ti} + \beta_{2i}Tie_{ti} + \beta_{3i}Fives_{ti} + \beta_{4i}Size * Tie_{ti} + \beta_{5i}Size * Fives_{ti} + r_{ti}$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\text{Culture}_i + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}\text{Culture}_i + u_{1i}$$
$$\beta_{2i} = \gamma_{20} + \gamma_{21}\text{Culture}_i + u_{2i}$$
$$\beta_{3i} = \gamma_{30} + \gamma_{31}\text{Culture}_i + u_{3i}$$
$$\beta_{4i} = \gamma_{40} + \gamma_{41}\text{Culture}_i$$
$$\beta_{5i} = \gamma_{50} + \gamma_{51}\text{Culture}_i \quad (3.1)$$

Level 1:

$$RT_{ti}^{\log} = \beta_{0i} + \beta_{1i}Size_{ti} + \beta_{2i}Tie_{ti} + \beta_{3i}Fives_{ti} + \ldots \quad (3.2)$$

Note. Equation 3.2 is partially written because it is identical to Equation 3.1 except for the dependent variable at Level 1. Both equations are written in the hierarchical form of LMM. All the βs are randomly varying coefficients, representing unique intercept and slopes for individual i. All the γs are fixed effects and all the us are random effects. r_{ti} is the unexplained residual of person i on trial t.

The logRT model with all four random effect converged successfully, whereas the rawRT model with the same number of random effects did not. A rawRT model converged when I eliminated either the random effect for the second dummy variable for

problem type (fives), u_{3j}, or the random effect for problem size, u_{1j}. Results from both converged rawRT models are comparable and thus I will not discriminate between them in the following text. Note that the rawRT model with one fewer random effect and the logRT model still have the same number of fixed effects. I compared the coefficients for the fixed effects in the rawRT model to that from the logRT model. They differ in terms of the magnitude and direction of the coefficient for the size x fives interaction. In the logRT model, a two-way interaction between size and five status occurs because 5-operand problems have a bigger problem-size effect than regular problems. However, in the rawRT model, 5-operand problems have a similar problem-size effect as regular problems. Because of this discrepancy between coefficients in the logRT model and the rawRT model, I chose to interpret results from the logRT model (Table 3-2).

Table 3-2 *Estimates from Two-level Linear Models Predicting Response Time (RT, log-transformed) for 7343 Trials and 40 Participants (20 Canadian; 20 Chinese)*

		Estimate	*t*
Fixed effects	γ_{00} Intercept	**7.08***	**236.87**
	γ_{10} Size	**0.0089***	**17.14**
	γ_{20} Tie	0.025	1.16
	γ_{30} Fives	**-0.15***	**-6.90**
	γ_{40} Size*Tie	**-0.0053***	**-12.53**
	γ_{50} Size*Fives	**0.0017***	**2.82**
	γ_{01} Culture	-0.027	-0.64
	γ_{11} Size*Culture	**-0.0052***	**-7.11**
	γ_{21} Tie*Culture	-0.065	-2.13
	γ_{31} Fives*Culture	**0.12***	**3.91**
	γ_{41} Size*Tie*Culture	**0.0043***	**7.34**
	γ_{51} Size*Five*Culture	-0.0011	-1.26
Random effects	level-1 σ^2	0.2	
	level-2 Intercept τ_{00}	0.13	
	level-2 Size τ_{11}	0.0021	
	level-2 Size*Tie τ_{22}	0.058	
	level-2 Size*Fives τ_{33}	0.046	

* $p < .05$ by estimation. The lme4 package in R, which I used to run LMM in the current thesis, does not provide *p* values for coefficients in mixed modelling, partly because statisticians are still debating whether an approximated degree of freedom of denominator in an *F* test is meaningful. As a rule of thumb, I examined whether or not the absolute value of the t-statistic exceeded 2 (Baayen et al., 2008).

The logRT model contains all three random slopes, imposing the least restriction on how much the effect of a predictor is allowed to vary across individuals. Also, log-transformation ameliorates non-normality of the dependent variable. However, a

drawback of this model is the difficulty of interpreting its coefficients. In raw RT models, for example, a coefficient can be interpreted as "for every one-unit increase in problem size, the RT increases by x ms". However, in a log RT model, the coefficient for problem size means "for every one-unit increase in problem size, the log RT increases by a factor of e^x". To circumvent such non-intuitive interpretations, I chose to plot point estimates on the raw scale (transforming the estimates back to ms units from the logRT values), and interpret coefficients in the context of the plot (*Figure 3-1*).

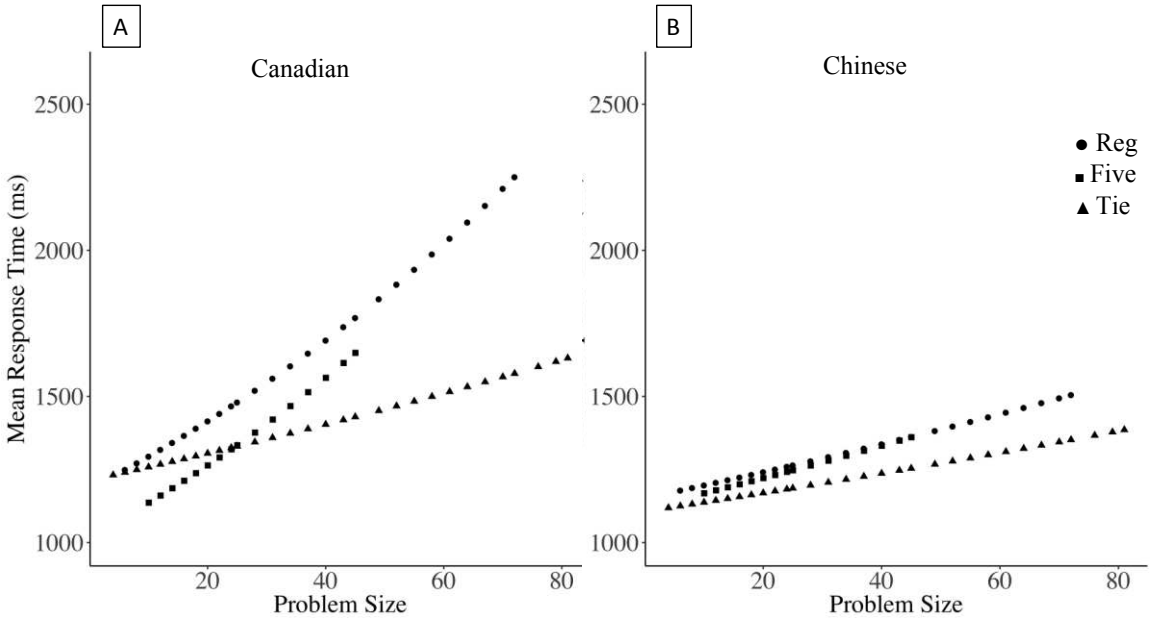

Figure 3-1 Point estimates for A) Canadian and B) Chinese participants estimated from the logRT model.

Size*Culture. Both groups of participants slowed down as problem size increased. Nevertheless, the problem size effect was shallower for Chinese than for Canadian participants on average. For a typical Canadian participant, for example, the predicted RT difference between the problem 2x4 and the problem 8x9 is 1000 ms (i.e.,

2300 – 1300 ms). In contrast, a typical Chinese participant is predicted to perform only 300 ms more slowly on 8x9 versus 2x4.

Fives*Culture. On average, Chinese participants appear to be equally fast on fives and regular problems. However, Canadians were slower on regular problems compared to five problems. This two-way interaction was NOT modified by problem size (i.e., the three-way interaction was not significant). This pattern can also be discerned by comparing the two plots. The dots and squares almost overlap with each other in the right-hand plot, whereas they are far apart in the left-hand plot.

Fives*Size. The five-operand advantage is greater on small than on large problems. For example, 2x5 is particularly easy compared to regular problems of comparable size (2x4, 2x6, etc.). As problem size increases, the speed difference between fives and regular problems reduces. As a result, the slope for 5-operand problems is slightly steeper than for regular problems. This pattern is similar across Chinese and Canadian participants (i.e., there is no three-way interaction with culture). Note that this differential problem-size effect on fives and regular problems, driven by a ">" shaped interaction, was detected in the log RT model but not in the raw RT model. In Chapter 5, I will provide another example in which the shape of an interaction is sensitive to the scale of the dependent variable. I will hold off the discussion about this phenomenon till then.

Size*Tie*Culture. For Canadian participants (panel A of Figure 3-1), the slope for tie problems is distinctively different from that for regular problems. Such a pattern is much less salient for Chinese participants (panel B of Figure 3-1). Canadians have a disadvantage on regular problems compared to tie problems, whereas Chinese have a

much smaller disadvantage for regular than tie problems, although they are still faster on tie problems. The lack of a two-way interaction between tie status and culture suggests both Canadian and Chinese participants show a tie advantage.

Summary. Both Chinese and Canadian participants became slower as problem size increase. However, this trend was more prominent among Canadians and bears a distinct problem-type signature when it is compared to the results for the Chinese, who had a modest tie advantage and no five advantage.

LMM versus By-participant Regression

How do results using LMM compare to that in the original study using by-participant regression? I replicated the procedure reported in the original paper on the current dataset (which was almost identical, save for a few trials due to different data cleaning procedures). *Figure 3-2* provides compelling evidence that the two statistical approaches provide highly similar results, at least when applied to the current dataset.

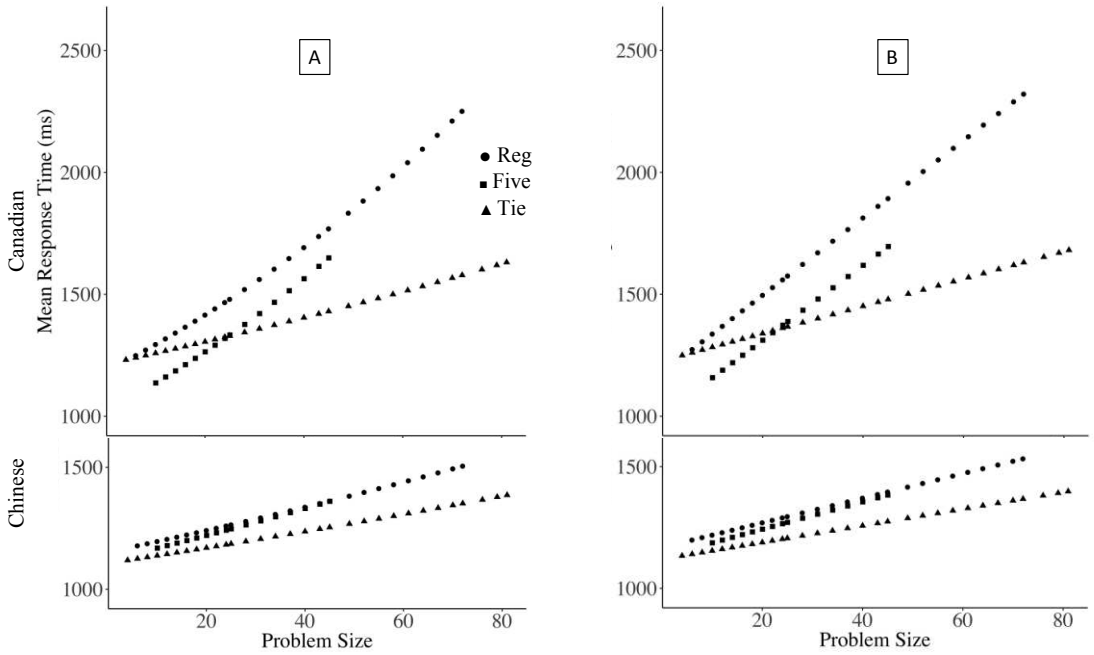

Figure 3-2 Predicted values of response time across problem size and type for Canadian and Chinese from A) LMM versus B) by-participant regression.

Comparing coefficients in both LMM and by-participant regression reveals a subtle difference. Note that the coefficients for LMM and by-participant regression in Table 3-3 are not directly comparable because RTs have been log-transformed in LMM but not in by-participant regression. In LMM, Size x Fives (γ_{50}) is significant whereas the three-way interaction with culture (γ_{51}) is not, which means that both Chinese and Canadian groups have steeper problem-size slopes on five-operand problems than on regular problems. This inference was not supported by results from the by-participant regression because the interaction of size x five status (i.e., Size x Fives) was not significant in either group. According to the by-participant regression, therefore, regular and five-operand problems have the same slope of problem-size effects.

Table 3-3 *Comparison of the fixed slopes of the multilevel model vs. coefficients of the by-participant regression*

	LMM		by-participant regression		
			Canadian	Chinese	
	Estimate	*t*	Estimate	Estimate	*t*
γ_{10} Size	**0.0089***	**17.14**	**15.87***	**5.06***	
γ_{20} Tie	0.025	1.16	**49.11***	**-47.88***	
γ_{30} Fives	**-0.15***	**-6.90**	**-173.76***	**-36.99***	
γ_{40} Size*Tie	**-0.0053***	**-12.53**	**-10.26***	**-1.62***	
γ_{50} Size*Fives	**0.0017***	**2.82**	-0.50	0.55	
γ_{01} Culture	-0.027	-0.64			
γ_{11} Size*Culture	**-0.0052***	**-7.11**			**4.95***
γ_{21} Tie*Culture	-0.065	-2.13			**-2.55***
γ_{31} Fives*Culture	**0.12***	**3.91**			**2.98***
γ_{41} Size*Tie*Culture	**0.0043***	**7.34**			**-4.91***
γ_{51} Size*Five*Culture	-0.0011	-1.26			-.41

Note. Statistics in the last column are based on a two sample *t*-test (*df*=38). For example, $t = 4.95^*$ means the coefficients of *Size* differ across groups.

Another apparent discrepancy between the results from LMM and the by-participant regression is worth noting. The coefficient for the tie effect (γ_{20}) is non-significant in LMM. However, in the by-participant regression, this coefficient is significant for the Canadian participants, but not the Chinese participants, suggesting that the results from LMM contradict those from the by-participant regression. This discrepancy reflects the differences in how main effects and conditional effects appear in the analyses. Specifically, the non-significant tie coefficient in the LMM indicates that response times for a tie versus a regular problem of size=0, when extrapolated, are not

different. This coefficient is conditional on all other predictors (being zero) in the model. The interaction of Size x Tie x Culture (γ_{41}), in contrast, reflects that the tie effect differs for the Chinese and Canadian groups. Thus, when interpreting regression coefficients in LMM it is important to always start with the highest interaction term and to only interpret a lower term when the higher term has failed to reach significance. Applying this approach resolves the discrepancy between the two analyses.

Summary. The analyses using LMM and the by-participant regression resulted in almost identical findings, save for the size x five interaction which was only detected in LMM. This discrepancy is most likely due to the fact that the interaction depends on the scale of dependent variable, rather than difference in sensitivity between LMM and by-participant regression. Both approaches revealed a tie advantage that differed in size for the Chinese and Canadians and a five advantage that was unique to Canadians. Using LMM, one can also conclude that the problem-size effect is slightly larger on 5-operand problems than on regular problems. Thus, these two approaches, at least for the present dataset, result in very similar conclusions. LMM has been validated as an appropriate tool in the context of problem-size effect for handling repeated-measure data which have a mix of continuous and categorical predictors. These analyses also suggest that the results of LMM and by-person regression are very similar in balanced data sets.

Piece-wise LMM

This additional analysis was intended as an exploration. When I plotted response times over problem size for each individual (see *Figure 3-3*), I noticed that most individual plots have a non-linear shape. Such nonlinearity suggests that a linear relationship in LMM may be a source of model mis-specification. To capture the non-

linearity of the problem-size effects, I performed a secondary analysis by introducing a piecewise component into LMM.

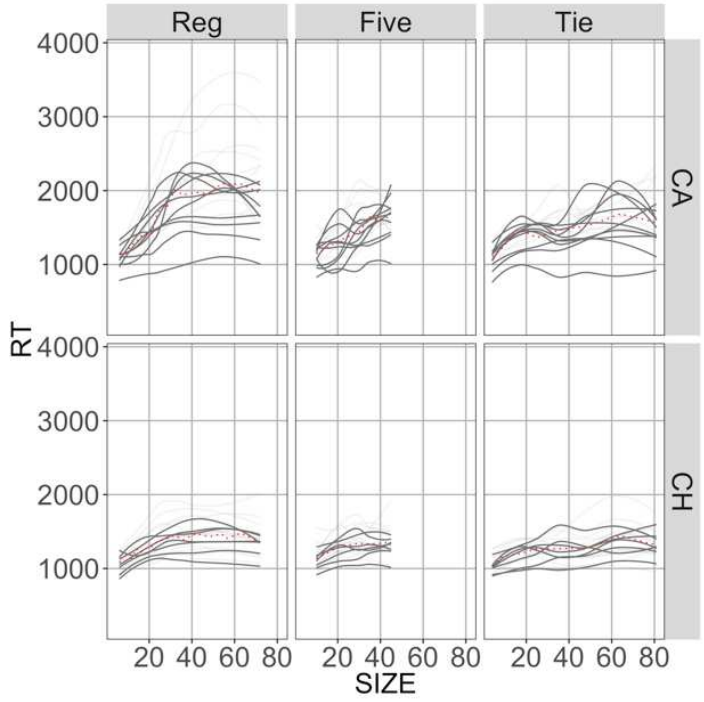

Figure 3-3 Spaghetti plot of size by culture (Canadian versus Chinese) by problem type (Regular, Fives, Tie). Each solid line represents one individual. Dotted lines represent mean latencies across individuals.

A few technical details need to be clarified before I present the findings. Instead of choosing the turning point via a data-driven approach, I placed the knot at size=25, echoing existing theories and computational models on problem-size effect (Campbell, 1994; Verguts & Fias, 2005). All tests remained the same as in the previous LMM model, except a single slope describing the problem size effect was replaced with two joined slopes, corresponding to small and large problems respectively. I plotted the results as I did previously. *Figure 3-4* compares plots from both linear and piece-wise models.

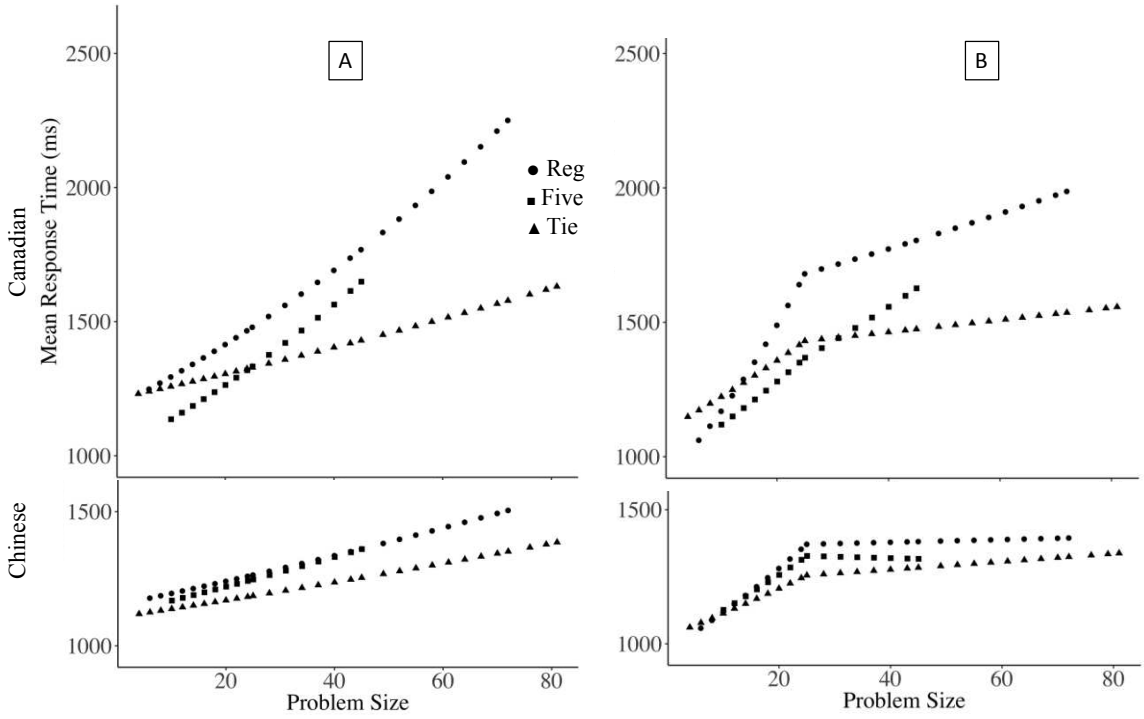

Figure 3-4 Predicted values of response time across problem size and type for Canadian and Chinese from A) LMM versus B) piece-wise model.

For both Canadians and Chinese participants, the knot at size=25 does make a noticeable difference in the trend, with the exception of 5-operand problems for Canadian participants. For Canadians, the problem-size effect became shallower for problems over 25, so much so that the predicted difference between 8x9 (regular) and 9x9 (tie) is roughly 400 ms, compared to a prediction of 800 ms in the linear model. For Chinese, the problem-size effect almost disappeared for problems over 25, whereas such an effect, albeit small, was still visible in the linear model. According to the piecewise model, large problems (size > 25) are indistinguishable from one another with respect to solution times for Chinese participants.

Summary. The piece-wise model captured the non-linearity between response time and problem size. According to this model, response time slows down unevenly rather than steadily as problem size becomes larger.

Discussion

I applied three statistical approaches to a dataset and compared patterns of the problem-size effect resulting from each approach. By-participant regression and multi-level modelling gave almost identical results in the current dataset. Piece-wise modeling is by far the most accurate, judging by its similarity to individual plots of the raw data and the improved model fit based on a likelihood ratio test ($\chi^2(11) = 746.43$, $p < .001$).

Comparison of LMM and By-Participant Regressions

It is interesting to see that LMM and by-participant regression gave similar results, even though they employed different estimation techniques -- maximum likelihood for LMM and ordinary least squares regression for the by-participant regression. This is not surprising given the nature of the current dataset, in which each participant solved the same set of problems and therefore item variability was minimized. It remains to be seen whether results would still be comparable between the two approaches when applied to a dataset in which each person receives a randomly selected set of stimuli.

The advantage of LMM over by-participant regression in the current dataset was subtle. LMM is more parsimonious; it saves one from having to run all the individual regressions and then combine the coefficients. All typical software packages are already configured to run LMM but by-participant regression requires brute-force effort with no added benefits. Also, model comparison is more straightforward with LMM than with by-

participant regression. For example, a researcher who wants to compare two forms of relationship between the outcome and predictors (e.g., linear vs. quadratic), s/he only needs to run two models using LMM, whereas by-participant regression would require him/her to rerun every regression at the individual level.

One drawback of LMM, however, is that it is more complex to use than traditional modelling techniques such as ANOVA and regression. As the number of predictors in a model increases, it becomes challenging to decide which predictors should have random effects while still having a successfully converged model.

In summary, although by-participant regression is better than ignoring all the individual differences, or turning a continuous variable into a discrete one, as is often done in ANOVA, it is still a workaround. I therefore argue that whenever by-participant regression is appropriate to a dataset, one should choose LMM instead.

Comparison of Linear and Piecewise Analyses

The discrepancy between results from linear and piece-wise model is contrary to intuition. According to the piece-wise model, slopes prior to the knot are steeper than those past the knot. Problem-size effects seemed stronger on small problems than large problems, even though large problems seem harder, and thus we expect their response times to be more sensitive to size increment than those of small problems. How can these discrepancies be explained?

One possible explanation of these effects of problem size on either side of the knot at the product 25 is that the data are affected by a selection bias in that only correct trials were analyzed. Incorrect ones tend to be large problems as well, and discarding them may have skewed the RT for large problems downward. In the current dataset, there

were 240 incorrect trials, and 213 of these were large problems. Here I am assuming these discarded trials, had they been correct, would have weighted the RT upward (slower RT on average, *assumption I*). Why would those trials, if they were correct, be more likely to be slow trials? Trials solved by procedures, rather than retrieval, are generally slower (LeFevre, Bisanz, et al., 1996). And I assume, once again, they would be more likely to be incorrect because procedures have high failure rates on very large problems (LeFevre, Sadesky, et al., 1996). By "very large", I mean approximately size > 40. Whenever procedures fail, a trial would be excluded from the latency analysis. For the middle-range problems (size around 25), however, non-retrieval procedures are more likely to succeed (*assumption II*). As a result, the correct trials consists of 1) small problems which are either solved by retrieval or fast procedures, 2) mid-range problems solved by retrieval and not-so-fast procedures, and 3) large problems solved by retrieval and even slower procedures. The proportion of correct trials solved by procedures, however, is bigger for mid-range problems than for large problems (a consequence of different success rate of procedures, following assumption I, II). Such a composition would result in the counter-intuitive pattern from the piece-wise analysis.

Assumption I is not directly testable, because it would be impossible to know what any individual incorrect trial would be had it been correct. However, assumption II is directly testable: procedures are slow in general, but some procedures are more error-prone than others. Outcomes from testing these assumptions would bear on the validity of assumption I. However, the above explanation will unlikely apply to Chinese participants, given their overall low error rate: Only 44 incorrect trials in total and 36 of

them were large problems. A different explanation other than data attrition is required to account for the current results from the piece-wise analysis for Chinese participants.

The current analysis demonstrated the power of piecewise models in detecting nonlinearities. The common way to test nonlinear relations is to add a quadratic term, whereas a piecewise model accomplishes this without resorting to sophisticated theories to justify the parabolic relationship. Moreover, coefficients in a piece-wise model are more straightforward to interpret than those in a quadratic form model, for example. However, a quadratic or other nonlinear function such as exponential may be preferable under certain circumstances when gradual change, rather than a discrete shift, is desired.

I have only shown in this analysis the simplest form of piecewise models. Additional knots may be added to test whether adjacent slopes are significantly different from each other. A knot may be eliminated if two surrounding slopes are the same (Panis, 1994; Cudeck & Klebe, 2002). Furthermore, Cudeck and Klebe (2002) showed how to model different knots for each individual.

Summary

Linear mixed modelling (LMM) and by-participant regression gave almost identical results for the first dataset. The current analysis showed two advantages of LMM over general linear model (GLM): 1) LMM is more parsimonious, and 2) LMM affords the opportunity to use versatile models such as a piece-wise one. In the next chapter, I apply LMM to an archival dataset originally analyzed using repeated measure ANOVA (RM-ANOVA).

CHAPTER 4: Operand Order Effect

Introduction

In the current study, I re-analyzed a dataset collected as part of an honour thesis project (Zhao, 2014). This study has its background rooted in the question of single vs. dual representation of arithmetic facts. Are both members of a commuted pair, such as 3x5 and 5x3, represented in long term memory or is only one of them represented? And, if there is a single representation, is it order specific or order free? Note that these research questions assume that all target items have already been stored in long-term memory. Chinese-educated participants, known for their well-drilled arithmetic facts, provide a good opportunity to explore this question. Thus, in the present study all participants had received their elementary schooling in China.

The current dataset was originally analyzed with RM-ANOVA using a 2 (operation) x 2 (order) within-subject design. Results from that analysis showed a significant interaction between operand order and operation. Operand order affects solution times in a systematic way for multiplication and addition, however, in opposite directions. I expect to find the same effect in the current analysis with LMM. More specifically, minimal operand followed by maximal operand for multiplication, that is, MIN x MAX, and the reversed order for addition, that is, MAX + MIN, respectively, are the preferred operand-order such that they would be solved faster than their commuted counterparts. This hypothesis is based on existing theories and evidence (Butterworth, Zorzi, Girelli, & Jonckheere, 2001; Verguts & Fias, 2005). For addition, Butterworth and colleagues' (2001) COMP model predicts MAX + MIN to be the preferred order as a legacy of the count-on strategy, a procedure children practiced when learning addition.

Although the COMP model was developed using British samples, Chinese students share similar acquisition experience of simple addition: count-on strategy is a common component of the curriculum in Chinese classrooms (Zhou & Peverly, 2005). Chinese students also are drilled on facts-to-10 and on using decomposition (to make 10s) for sums beyond 10. Thus, 8 + 6 can be solved by decomposing 6 to 2 + 4; the problem is thus reformulated as 8 + 2 = 10 + 4 (Zhou & Peverly, 2005). This strategy is highly consistent with the form of the number words in the Chinese language. For multiplication, Chinese children are drilled with half of the table, that is, with smaller operand first and tie entries. Accordingly, studies have shown that the order MIN x MAX is solved about 20 – 40 ms faster than the order MAX x MIN (LeFevre & Liu, 1997).

Other than the hypothesized interaction between operation and operand order, I also expected findings typically reported in this body of literature, although they are not central to the current research question. For instance, problem-size effects were expected regardless of the operation type (i.e., addition, multiplication), operand order (e.g., MAX x MIN, MIN x MAX), or problem type (e.g., fives, ties), with the possible exception of addition ties. LeFevre, Shanahan and DeStefano (2004) and Butterworth et al. (2001) have reported flat problem-size effects on addition ties, although Miller et al. (1984, p.52) reported a non-zero slope for addition ties.

I also expected to find an advantage for ties over nonties, and an interaction between tie status and problem size. Previous studies have consistently found the tie effect on multiplication problems. However, results on addition problems are somewhat mixed for Chinese-educated participants. In particular, Campbell and Gunter (2002) tested Chinese participants and did not find either a tie advantage or a tie x size

interaction; addition ties and non-ties were indistinguishable for those participants. In contrast, Canadian- and American-educated participants typically show tie advantages and tie x size interactions on addition problems (Campbell & Gunter, 2002; LeFevre, Sadesky, et al., 1996; Miller et al., 1984). Given the existing literature, I hypothesized that there would be a tie advantage and a tie x size interaction on multiplication, with an attenuated but non-zero slope on multiplication ties. Furthermore, both addition ties and nonties would exhibit shallower slopes compared to multiplication with no tie x size interaction.

Method

Description of data

Forty-three participants solved single-digit addition and multiplication problems in separate blocks. All participants were born and educated in China up to high school. However, they were attending a post-secondary Canadian university at the time of the current study. The experiment was conducted in their native language, Mandarin. Each person solved 256 problems, 128 for each operation[1], resulting in 10,880 trials in total. Features of problems varied on four dimensions, including operation, problem size, operand order (5 x 7 vs. 7 x 5), and tie status. Inaccurate and invalid trials were excluded from the current analysis. As a consequence, each person effectively contributed 256 latency data points at most, ranging from 210 ~ 250 each and totaling 10,062 trials.

[1] Participant #21 only had multiplication data.

Analysis plan

I applied a two-level cross-classified mixed-effects linear model to the dataset, in which individual participants made up the level-2 units and arithmetic problems solved by each individual comprise the level-1 units. No predictors were entered at level-2. At level-1, the response time of each individual problem was a function of its operation type (+, x), problem size (sum/product of two operands), type, operand order, and interaction terms among them. With the exception of operation type, which is straightforward, I provide a rationale of coding scheme for each predictor. Problem size was operationalized as the sum of two operands for addition and product for multiplication. Preliminary analysis suggested scaling the problem size improved the stability of parameter estimation, robustness to outlying observations, and model convergence. Therefore, problem size was standardized before entering the model.

In the literature, single-digit addition problems are often subcategorized into two types: tie and non-tie problems (Campbell, 1994, 1995; Groen & Parkman, 1972; LeFevre et al., 2004). An extra category, fives, has been identified for single-digit multiplication problems due to its distinct behavioural signature compared to tie and other non-tie multiplication (Campbell & Graham, 1984). In the current analysis, however, I choose not to distinguish between fives and other non-tie multiplication problems. Furthermore, findings in Chapter 3 suggested that Chinese-educated participants did not exhibit different behavioural patterns on fives and other non-tie multiplication problems, in contrast to their Canadian counterparts. Therefore, one dichotomous variable was required to code the tie status of problems.

Operand order is another two-category variable that requires a dichotomous coding. However, operand order is undefined for tie problems (either addition and multiplication). To have two predictors, tie status and operand order, simultaneously in the model, I decided to code problem type and operand order conjunctively as one predictor with three levels; SL, LS, and tie, in which SL stands for problems with the small operand preceding the large operand and vice versa. The choice between dummy coding and contrast coding is driven by my research questions: only contrast coding would allow me to directly test the main effect of operand order as well as its interaction with operation type, should there be any. As described in Table 4-1, the first contrast variable compares SL and LS problems, while the second aggregates SL and LS problems and compares them to tie problems.

Table 4-1 *Coding scheme of a categorical variable that combines operand order and tie status*

	Contrast 1	Contrast 2
MIN MAX	1	1
MAX MIN	-1	1
TIE	0	-2

Results

Data preparation

I deleted any trials with response times below 250 ms as they were deemed improbable responses (n = 283; 2.6% of data). Preliminary analysis indicated that the dependent variable, consisting of RT data, was skewed and therefore was log-transformed. I adopted the same approach as in Chapter 3 and back-transformed point

estimates before plotting them. Coefficients of tested models were interpreted in the context of plots.

Models to be tested. In the current analysis, I followed recommendations by Barr et al. (2008; cf. Chapter 2) and used the maximal random effects structure justified by the current experimental design, which includes by-subject random intercept, by-item random intercept, random slope for problem size, operation, operand order, and problem type. However, this model failed to converge. Accordingly, I eliminated the correlations between random slope and random intercept (i.e., fixed them at zero, Equation 4.2), following recommendations by Barr et al. (2008), and the model converged successfully (Equation 4.1).

$$logRT_{ti} = I_{0i} + \beta_{0i} + \beta_{1i}Operation_{ti} + \beta_{2i}Problem\_size_{ti} + \beta_{3i}Operand\_order_{ti} + \beta_{4i}Tie_{ti} + r_{ti} \quad (4.1)$$

$$\begin{pmatrix} V_{U_0} & & & & \\ 0 & V_{U_1} & & & \\ 0 & 0 & V_{U_2} & & \\ 0 & 0 & 0 & V_{U_3} & \\ 0 & 0 & 0 & r_{34} & V_{U_4} \end{pmatrix} \quad (4.2)$$

Note. In Equation 4.1, I is the by-item random intercept and βs are randomly varying by-participant coefficients. All the interaction terms among predictors were omitted from the equation but were included in the actual model. In Equation 4.2, the diagonal entries represent variance of random effects in model 4.1 whereas the off-diagonal entries represent correlations between these random effects.

Diagnostics. A box plot showed outliers on both the fastest and slowest end of the scale for response time. Subsequently, I deleted 25 observations on either side, which was less than 1% of the total observations. Estimates of fixed and random effects are comparable in the analyses of the full and reduced sample, although model convergence

was improved with the reduced sample. Therefore, I proceeded with the reduced sample. I also examined the assumptions of normality and heteroscedasity of residuals at level 1 as well as normality and linearity of random effects at level 2. No further outliers were detected.

Table 4-2 *Estimates from Two-level Linear Models Predicting Response Time*.

	Estimate	*t*
Intercept	**6.75***	**247.59**
Size	**0.030***	**3.80**
Order	0.010	1.41
Operation	**0.099***	**5.94**
Tie	**0.039***	**2.60**
Size * Order	0.00095	0.12
Size * Operation	**0.058***	**5.91**
Order * Operation	**-0.024***	**-2.28**
Size * Tie	0.00064	-0.06
Operation * Tie	0.017	0.88
Size * Operation * Order	-0.0034	-0.30
Size * Operation * Tie	**0.066***	**4.42**

Fixed effects

Size x Operation x Tie. This result is most obvious from the plot (Figure 4-1). Although tie and non-tie problems have drastically different slopes for multiplication, this pattern is absent for addition problems. Compared to tie multiplication, non-ties have a much steeper slope (i.e., stronger size effect). The largest tie multiplication (9x9) takes 100 ms longer than solving the smallest (2x2), whereas this discrepancy jumps to as large as 500 ms between 2x3 and 8x9, the smallest and largest non-tie problems. In contrast,

the slope of problem-size effects for both tie and non-tie addition problems are much shallower than their multiplication counterparts.

Order x Operation. This coefficient represents the difference between multiplication problems (when operation = 1) with different operand order. For multiplication, participants solved problems where the smaller operand was followed by a larger operand 25 ms faster than their commuted counterparts whereas in addition, they solved problems where the bigger operand followed by a smaller operand 18 ms faster than their commuted counterparts.[2] The speed difference between commuted pairs remained the same across different problem sizes, signified by the parallel slopes of problem-size effect on both operand order (*Figure 4-1*) and the lack of a three-way interaction of order x operation x problem size ($t = 0.30$).

Size x Tie. This coefficient represents the difference between the slopes on tie and non-tie addition problems (op = 0). It is a non-significant two-way interaction (t = -.06), which means the slopes of tie and non-tie problems are parallel for addition (see *Figure 4-1* (Left)).

Size x Operation. The problem-size effect, averaged across tie and non-tie problems, has a steeper slope for multiplication than for addition.

---

[2] To get this effect size, I reversed the coding of operation, with multiplication as the reference group and addition coded as 1, and ran the model again.

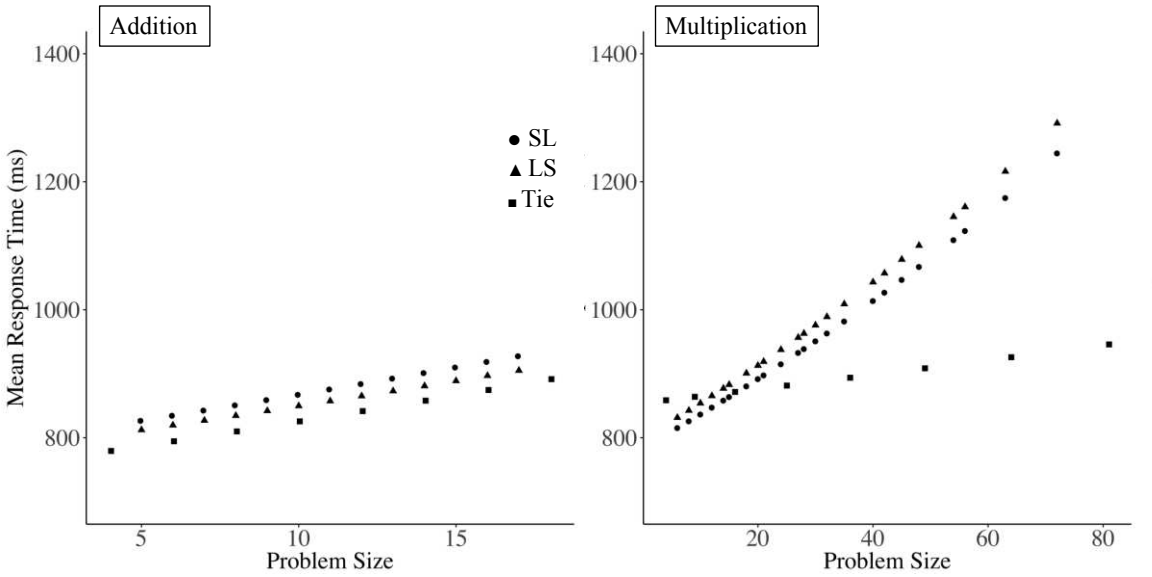

Figure 4-1 Problem-size effects for ties and non-ties for addition and multiplication. SL - small operand preceding large operand. LS - large operand preceding small operand.

Discussion

The goal of this experiment was to test the hypothesis that Chinese-educated participants process addition and multiplication in a different but consistent fashion with respect to the operand order. The present analyses supported this hypothesis. Chinese participants solved multiplication problems with the smaller operand first (i.e., MIN x MAX) faster than their commuted pairs by a small but consistent amount of time. Conversely, they solved addition problems with the larger operand first (i.e., MAX + MIN) faster than their commuted pairs.

For multiplication, operand order effects are consistent with a single-representation account. On this view, when a Chinese participant encounters a problem in the form of MAX x MIN, for example, 5 x 3, this participant phonologically encodes both operands, that is wu san (i.e., five three in Mandarin). Encoding of each operand then activates all the complete mental representations associated with that operand, such

as "wu wu er shi wu" (“five five two-ten-five”), "san si shi er" ("three four ten-two"), and so on. Of all the activated facts, only “san wu shi wu” (Mandarin for “three five ten-five”) matches with both encoded operands, because Chinese-educated participants only learned half of the multiplication table and did not practice "wu san shi wu" (Mandarin for "five three ten-five"). The participant verbalizes the answer once a match has been found. Note that such an account does not presuppose a numerical comparison stage in the mental process, even though it appears as if MIN is consistently preferred to be on the left. Participants need not to be aware of the nature of the privileged format in their memory (i.e., min preceding max). Hence, the extra time spent on solving MAX x MIN compared to solving MIN x MAX should in no way be equated to time that would be spent for making a numerical judgement of the size of two operands, instead, it represents the slightly faster matching of 3 x 5 to san wu shi wu compared to 5 x 3.

The results for addition were also as predicted. This Chinese sample showed a preferred operand order for addition problems, that is, MAX + MIN. Chinese students are explicitly taught the sum-to-10 strategy in first grade (Zhou & Peverly, 2005); counting strategies are not emphasized. To use this strategy on problems with sums over 10, one "grows" the larger operand (MAX) into 10 by breaking down MIN, then adding the remainder from MIN to 10. For example, 8+5 = 8 + 2 + 3. Chinese students are also taught to use part-whole strategies for sums less than 10 (Zhou & Peverly, 2005); they might also practice with an abacus, which emphasises part whole strategies based on five.

Consistent with the second hypothesis, a problem-size effect was observed for both operations, across the operand order and problem types, with a steeper slope on

multiplication than on addition problems. Notably, addition ties also showed a problem-size effect, albeit smaller compared to multiplication tie problems.

As predicted, multiplication problems exhibited both a tie advantage and a tie x size interaction. Overall, tie problems were solved faster than nonties. This trend is qualified by problem size, such that small ties and nonties were solved equally quickly. This finding also agrees with Campbell and Gunter's (2002) findings, corroborating their rejection of an encoding-based account for the tie advantage. As problem size increases, nonties become increasingly disadvantaged compared to ties. The same pattern has been repeatedly reported for North American samples, although with a larger discrepancy between tie and non-tie problems. Multiple theories exist to account for the tie effect. However, Verguts and Fias' theory (2005) is worth special attention here because their assumption, that only half of the multiplication table is stored in semantic memory, is more plausible in this particular sample; Chinese only learn half of the times table. According to Verguts and Fias, tie problems, being positioned on the diagonal, receive less interference from adjacent problems (described in Chapter 1). Although similar ideas underlie other theories -- that there is less interference for ties as problem size increases compared to non-ties, higher retrieval success rate, and less frequent multi-step calculations, Verguts and Fias' computational model has the advantage of being precise and comprehensive.

Similar to the results reported in Campbell and Gunter (2002) for their Chinese-educated participants, no tie x size interaction for addition was observed in the current sample. In contrast, previous studies with North American samples have consistently found the tie x size interaction for both addition and multiplication (e.g., Ashcraft, 1992;

Campbell, 1994, 1995; LeFevre, Bisanz et al., 1996). This discrepancy between the two populations suggests that, 1) for Chinese-educated participants, mental representations of addition ties and nonties may have more in common than they do in the North American population, and 2) for Chinese-educated participants, the mental processes involved in solving addition and multiplication may be fundamentally different as compared to those in the North American population.

The current analysis found a small tie advantage overall for addition, unlike Campbell and Gunter (2002) in which tie and nontie addition were indistinguishable in terms of solution times. This could be due to the fact that participants in the current study solved each problem twice, both tie and non-tie, whereas participants in Campbell and Gunter only solved tie problems once each. It is plausible that tie problems benefited more from the extra practice, and thus had a small advantage over non-ties.

The current dataset was originally analyzed with RM-ANOVA using a 2(operation) x 2(order) within-subject design (Zhao, 2014). Results from that analysis are consistent with the current one. In particular, an interaction between operand order and operation was also significant. It is tempting to think that, because RM-ANOVA is less powerful than LMM and it is merely a special case of LMM, it therefore should be replaced by LMM in all situations. Results from the current analysis caution against such practice. Despite several limitations of RM-ANOVA when applied to datasets in mental arithmetic, such as needing to dichotomize problem size and dismissing individual differences, of the two datasets I have analyzed so far, I cannot claim RM-ANOVA would produce grossly different results. However, using LMM allows problem size to be

a continuous variable and therefore retains the possibility of examining any curvilinear relation between problem size and response time.

The current analysis demonstrated two practical strategies for using LMM. Standardizing problem size so that it is not disproportionately larger than other predictors in the model seems to have improved the stability of parameter estimation. Furthermore, restricting the random effects structure by eliminating correlations between random intercepts and random slopes also improved model convergence. These tips point to some disadvantages of LMM however, specifically, that it is more complex and susceptible to misuse. For example, specifying a random effects structure that is theoretically sound while being practical has been recognized as challenging and a principled way to do so is still being debated in the literature (Baayen et al., 2008; Barr et al., 2013; Gelman, 2005).

I have applied LMM to two archival datasets, both of which have a balanced design. Results from LMM were comparable to those from GLM. In the next chapter, I analyze an archival dataset that distinguishes trials solved via retrieval versus nonretrieval based on participants' self-reports. Separating these two types of trials resulted in an unbalanced design. I apply LMM and RM-ANOVA to the same dataset and compare their results.

CHAPTER 5: The Effect of Problem Format

Introduction

The goal of the current study was to examine how the encoding stage interacts with the calculation stage of solving mental arithmetic. I used a dataset collected as part of an unpublished project (Pyke & LeFevre, 2009). Campbell and colleagues proposed that cognitive processes depend on the format of external inputs. In support of this argument, they have consistently found different behavioural patterns associated with arithmetic problems presented in a less familiar format, written words (e.g., three + six; eight x nine) versus Arabic numerals (3+6; 8x9; Campbell, 1994; Campbell & Alberts, 2009; Campbell & Epp, 2004). Campbell (1994), for example, reported steeper problem-size effects on arithmetic problems in word format compared to those in digit format. Similarly, Campbell and Alberts (2009) found exacerbated problem-size effects on addition problems in word format, which also coincided with a strategy shift. Participants reported more non-retrieval strategies on problems in word format compared to digit format, especially on large problems.

Campbell interpreted these results by referencing to the construct of "retrieval efficiency" (described in Chapter 1). Retrieval efficiency consists of two components: the efficiency of initiating a retrieval attempt and the efficiency of executing it once retrieval is attempted. Any factor that slows down solution times could do so by affecting either or both components. According to Campbell and Alberts (2009), format affects solution times mostly through affecting the first component, initiation of a retrieval attempt. Compared to digits, the word format prolongs the time to initiate retrieval, and may encourage counting and other non-retrieval procedures being attempted. As a result of

increased non-retrieval usage on large problems in word format, solution times for these problems are on average longer than that for digit problems (given that non-retrieval tends to be slower than retrieval), and hence participants show a steeper problem size effect for problems in word versus digit format.

In this view, number word is considered an unfamiliar format and is assumed to have a uniform effect on solution processes, which could interact with other factors such as problem size (Campbell, 1999). For example, Campbell and Alberts (2009) reasoned that the relatively unfamiliar word format exacerbates the disadvantage of large problems, that is, those with low memory strength, hence producing an intensified problem size effect. The current study was designed to test a competing interpretation based on the familiarity of individual operands. According to Metcalfe and Campbell (2007), familiarity with operands produces "a feeling of knowing the answer that results in a decision to attempt retrieval rather than to calculate a solution". In the present research, Pyke and LeFevre (2009) assumed that familiarity with operands can be approximated by printed frequency of number words (Brysbaert, 1995). Given that printed frequency of number words decreases with increasing operand magnitude (SUBTLEXUS data base, Brysbaert & New, 2009), word frequency may influence retrieval efficiency differently depending on operand size and therefore contribute to the problem-size effect for number-word problems.

However, the word frequency hypothesis cannot be tested with word versus digit formats because frequency of number words is confounded with problem size. To decouple these two variables and test whether or not familiarity with individual operands (i.e., printed word frequency) indeed contributes to the problem-size effect, the present

study contrasted a third format to that of digits and number words. Pseudo-homophones are letter strings which sound like words when they are pronounced. In arithmetic, they include pseudo words such as eyt and siks. These nonwords have similar frequency across magnitude (Table A1, from LeFevre, Pyke, & Penner-Wilger, 2010). If Campbell's theory is correct, the pattern of problem-size effects on pseudo words would look similar to those on real number words, because both formats are relatively unfamiliar compared to digits in the context of mental arithmetic. However, if pre-experimental experience with individual operands is responsible, in part, for the larger problem-size effect on word problems compared to digit problems, pseudo words should display similar problem-size effects as those for digit problems, both shallower than number words, because pseudo words have similar frequency across magnitude.

The following analyses primarily answers two questions. First, does the current dataset replicate Campbell and colleagues' (Campbell, 1999; Campbell & Alberts, 2009) findings for word versus digit problems? Second, does word frequency contribute to the problem-size effect? In tackling the first question, I also contrasted RM-ANOVA and LMM analyses of the data to examine whether the choice of statistical model has any impact on the conclusions.

Method

Description of data

Sixty-nine participants solved single-digit addition or multiplication problems with all combinations of operands between 1 and 9. Problem types include 1's (e.g., 6x1), tie (e.g., 4x4), and nontie (any problem that is not tie, nor has 1 in the operands). The operands were presented in three formats: digits (e.g., 2 x 5); words (e.g., two x five);

and pseudohomophones (e.g., tue x fyve). The pseudohomophones used in the present experiment include wun, tue, thrie, fowr, fyve, siks, sevin, eyt, and nyne. Participants solved problems in all three formats, but in only one operation (addition or multiplication). For each operation, participants solved the nine tie problems (e.g., 2 + 2) twice each. The 36 non-tie problems were also solved twice (e.g., 3 + 4 and 4 + 3) in each of the three formats. Thus, each participant solved a total of 90 (36 x 2 + 9 x 2) problems in each format, for a total of 270 across the experiment. Inaccurate trials and 1's problems were excluded from the current analysis. As a consequence, each person effectively contributed 140 ~ 213 latency data points to the analysis, resulting in 13,247 trials in total.

Self-reports were collected at the end of each trial. Participants chose from four options which type of strategy they had just used in solving the problem: remember, count, transform, and other. Description of these strategy types were provided to each participant at the beginning of the experiment, following the procedure described in Campbell and Alberts (2009):

- Remember: you solved the problem by just remembering the answer that just came into your head and you did not complete any intermediate steps.
- Count: you solved the problem by counting by ones.
- Transform: you solved the problem by using your knowledge of another arithmetic problem to solve the one on the screen. For example, you could use multiplication facts to help you solve addition problems or you could retrieve the answer to a different problem and then add or subtract an amount to get to the answer.

- Other: any other strategy that is not listed here.
- Unsure: you are not exactly sure how you solved the problem or you do not wish to provide a strategy report for this problem.

For the purpose of the current study, I only differentiated "remember" from the rest of the strategy types and separated trials solved with retrieval from those solved with non-retrieval.

Analysis Plan

Stage One. To replicate Campbell and Alberts' (2009) results, I excluded tie as well as pseudohomophone problems, so that digit and word format could be contrasted for both addition and multiplication nontie problems. I applied two statistical approaches at this stage: RM-ANOVA and LMM. In both approaches, problem size was operationalized as the product of two operands for both addition and multiplication (Campbell & Alberts, 2009). Replication was done in two steps: disregarding self-reported strategies and then taking them into account.

Size x Format. Analyses of problem size (small versus large) x format (digit versus word) were separately applied to addition and multiplication problems, following Campbell and Alberts' example. Small problems were classified as those with products of 25 or less (Campbell & Alberts, 2009). Four RM-ANOVA models were conducted, two for each operation, one using raw RT as the dependent variable and the other log-transformed RT (m200* in Table 5-1). Note that Campbell and Alberts' (2009) did not report results from their logRT models.

Table 5-1 *Models Tested in Each Stage*

Stage 1							
	Opera-tion	DV Trans-form	Formats Con-trasted	Size Dicho-tomized	Random Effects Structure	Retrieval	Code
ANOVA	a	r	d-w	y	int	n	m200ardw
LMM	a	r	d-w	y	int	n	m210ardw
LMM	a	r	d-w	y	max	n	m211ardw
LMM	a	r	d-w	n	int	n	m220ardw
LMM	a	r	d-w	n	max	n	m221ardw
LMM	a	r	d-w	n	max	y	m321ardw
ANOVA	m	r	d-w	y	int	n	m200mrdw
LMM	m	r	d-w	y	int	n	m210mrdw
LMM	m	r	d-w	y	max	n	m211mrdw
LMM	m	r	d-w	n	int	n	m220mrdw
LMM	m	r	d-w	n	max	n	m221mrdw
LMM	m	r	d-w	n	max	y	m321mrdw
ANOVA	a	l	d-w	y	int	n	m200aldw
LMM	a	l	d-w	y	int	n	m210aldw
LMM	a	l	d-w	y	max	n	m211aldw
LMM	a	l	d-w	n	int	n	m220aldw
LMM	a	l	d-w	n	max	n	m221aldw
LMM	a	l	d-w	n	max	y	m321aldw
ANOVA	m	l	d-w	y	int	n	m200mldw
LMM	m	l	d-w	y	int	n	m210mldw
LMM	m	l	d-w	y	max	n	m211mldw
LMM	m	l	d-w	n	int	n	m220mldw
LMM	m	l	d-w	n	max	n	m221mldw
LMM	m	l	d-w	n	max	y	m321mldw
stage 2							
	Opera-tion	DV Trans-form	Formats Con-trasted	Size Dicho-tomized	Random Effects Structure	Retrieval	code
LMM	a	r	d-p	n	max	n	m221ardp
LMM	m	r	d-p	n	max	n	m221mrdp
LMM	a	l	d-p	n	max	n	m221aldp
LMM	m	l	d-p	n	max	n	m221mldp
LMM	a	r	d-p	n	max	y	m321ardp
LMM	m	r	d-p	n	max	y	m321mrdp
LMM	a	l	d-p	n	max	y	m321aldp
LMM	m	l	d-p	n	max	y	m321mldp

Operation, a = addition, m = multiplication
Format contrasted, d-w = digit vs. word, d-p = digit vs. pseudohomophone
DV transform = dependent variable transformed. r = raw, l = log-transformed
Size Dichotomized, y = yes, n = no
Random Effects Structure, int = random intercepts only, max = random intercepts and slopes
Retrieval = self-reported strategy is one of the predictors, y = yes, n = no

To fully explore the differences between RM-ANOVA and LMM, I also applied LMM to the same subset of data: addition and multiplication nontie problems in digit versus word format. Across all LMM analyses, response time of each individual problem was regressed on its size and format (digit versus word), that is, aggregation *after* analysis, as is the common practice in LMM. This is in contrast to RM-ANOVA, in which cell means are always calculated before the main analysis, that is, aggregation *before* analysis. For each RM-ANOVA conducted above, I also conducted four LMMs as follows:

1) dichotomized problem size, same random effects structure as ANOVA;

2) continuous problem size, same random effects structure as ANOVA;

3) dichotomized problem size, maximal random effects structure;

4) continuous problem size, maximal random effects structure.

These LMMs resulted from independently manipulating whether or not problem size was dichotomized or continuous, and whether the structure of random effects was maximal or only random intercepts were included (m210* - m221* in Table 5-1).

Size x Format x Strategy. Campbell and Alberts (2009) also separated trials based on whether retrieval or nonretrieval was used in solving a problem, then contrasted the pattern of size x format interaction on retrieval versus nonretrieval trials. Those

analyses were severely underpowered, however, because individuals who used exclusively retrieval, for example, on small multiplication problems, had to be deleted from the dataset (18 out of 24 participants; p.1007). In fact, they could not test size x format for multiplication problems solved via nonretrieval because three quarters of participants would have been excluded.

Instead of replicating Campbell and Alberts (2009) with RM-ANOVA, I used LMM to examine how self-reported strategy use would moderate the size x format interaction. Unlike ANOVA , LMM has no requirement for complete data across trials for an individual. Therefore, I could retain any trials a participant contributed despite of the unbalanced design that was a consequence of separating retrieval and nonretrieval trials (cf. Chapter 2). Response time of each individual problem was regressed on its size, format (digit versus word), and self-reported strategy (retrieval versus nonretrieval). Two LMMs for each operation were conducted, one using raw RT as the dependent variable and the other log-transformed RT. Problem size was kept continuous and the random effects structure maximal (m321* - m321* in Table 5-1).

Stage Two. To understand the source of the format x size interaction, I included pseudohomophone problems in further analyses. Digit and pseudohomophone format were contrasted for both addition and multiplication nontie problems using LMM. Response time of each individual problem was regressed on its size and format (digit versus pseudohomophone). Four models thus generated (m221*dp in Table 5-1) were analogous to their counterparts comparing digit versus word format in stage one (m221dw*), all of which used maximal random effects structure and continuous problem size.

Furthermore, self-reported strategy (retrieval, nonretrieval) was added as a third predictor of RT (models m321*dp in Table 5-1), keeping problem size continuous and the random effects structure maximal. These models allowed me to compare the pattern of size x format interaction on retrieval and nonretrieval trials.

Tie problems had been excluded from all analyses so far. There are two reasons for this decision. I excluded them at stage 1 to replicate Campbell and Alberts (2009) analyses. To make results in stage 2 comparable to those in stage 1, I excluded tie problems again in stage 2. Preliminary analyses suggested that very few tie problems were solved using nonretrieval, making the estimate of coefficients that involves tie and nonretrieval unreliable. Thus, tie problems were excluded from all analyses.

Results

Data preparation

The histogram of the dependent variable, response time, revealed violation of normality. I applied log-transformation to the variable, which visually improved its normality. In the previous chapters, I ran parallel models with raw and log-transformed RT, 1) interpreted the rawRT model if results from both models were comparable, and 2) interpreted the logRT model if results were different. However, I deviate temporarily from this approach in stage one of the present study and interpret results from both models when they are different. Stage one of the current study was intended to replicate the analyses used by Campbell and Alberts (2009), in which the authors used raw RT. To ease the comparison between Campbell and Alberts and the current study, therefore, I

placed more emphasis on raw data models. In stage two of the current study, I resume the same practice as before, that is, interpret the logRT model whenever results are different.

I also examined other assumptions including heteroscadesity and linearity and conducted sensitivity analysis with potential outliers on both level-1 and level-2. Subject 47 was identified as an influential outlier. This person had extremely slow RTs on some multiplication problems. I deleted this participant and based the following analyses on the remaining 68 participants.

Stage One

At stage one, I tested whether the present findings replicated those of Campbell and Alberts (2009); that is, whether the problem-size effect was larger for word than for digit format problems. I excluded ties as well as pseudohomophone problems in order to directly compare the results to those reported by Campbell and Alberts. I first present results ignoring strategy reports. Then I proceed with analyses in which strategy was used as an additional predictor.

Size x Format. Using raw RT as the dependent variable, I conducted RM-ANOVAs for addition and multiplication separately with a focus on the format x size interaction. The results are shown in Table 5-2 and plotted in *Figure 5-1*(a). Note that Campbell and Alberts (2009) termed the format x size interaction "format cost". However, I renamed this effect as relative format cost; format cost relative to problem size. I used format cost to refer to the RT difference between two formats at a given value of problem size. Overall, word problems were solved slower than digit problems. The problem-size effect was exacerbated for word format on addition but not multiplication problems. In particular, there was an interaction of size x format on addition, as shown in

Figure 5-1 (a). The interaction of format x size was not significant for multiplication, however. Note that relative format costs are comparable in magnitude for both operations, about 100 ms (Table 5-3). However, the format x size interaction for multiplication failed to reach significance because it had a much larger variability than addition (almost ten-fold, as indicated by error bars in *Figure 5-1* (a)).

Table 5-2 *Summary of ANOVA: Raw Response Time as A Function of Problem Size and Format for Addition and Multiplication*

Effect	Operation	MS (MSE)	df1, df2	F
problem size	addition	8169698 (86219)	1, 33	**94.76*****
	multiplication	17694740 (233519)		**75.77*****
format	addition	6158615 (52053)	1, 33	**118.3*****
	multiplication	6506849 (117047)		**55.59*****
problem size x format	addition	103244 (4913)	1, 33	**21.01*****
	multiplication	74657 (47817)		1.56

Note. *** $p < 0.001$

Table 5-3 *Means and standard deviations for RT (ms) as a function of a 2(format) X 2(problem size) design*

	problem size (addition)				problem size (multiplication)			
	Small		Large		Small		Large	
format	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Digit	1014	457	1449	805	1225	727	1900	1229
Word	1384	548	1929	926	1616	895	2384	1621
Format cost	370		480		391		484	
Relative format cost	110				93			

Note. Format cost = difference between word and digit on small/large problems. Relative format cost = format cost on large problems - format cost on small problems.

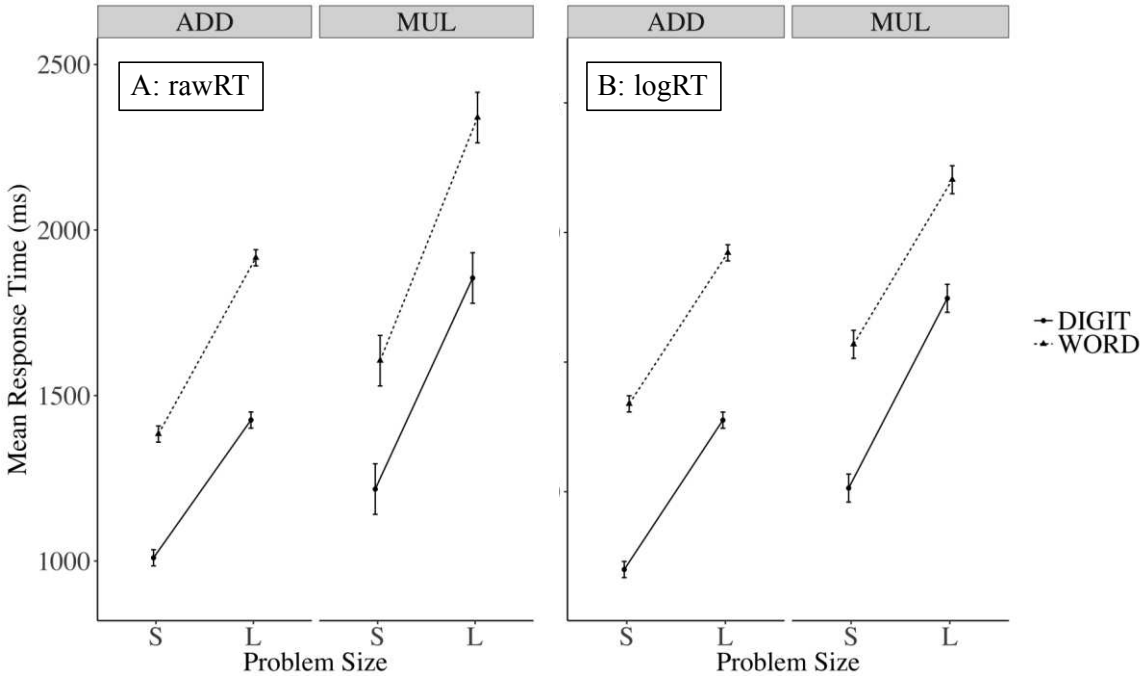

Figure 5-1 Interaction of problem size (small vs. large) and format (word vs. digit) with 95% confidence intervals: (a) raw RT, (b) log RT.

Both in Campbell and Alberts 's (2009) and the current study, participants showed a steeper problem-size effect for word compared to digit problems on addition. Furthermore, the size of the relative format cost was similar across two studies, about 100 ms (Campbell & Alberts, 2009, p.1003). Similarly, in both studies, there were nonsignificant size x format interactions for multiplication. However, the magnitude of this effect in the current sample is much larger (93ms) than in Campbell and Alberts (-9 ms). In fact, Campbell and Alberts' result surprised the authors as well (see footnote 4 on p. 1007). According to Campbell's (1995) encoding-complex theory, sizable differential

relative format costs are expected for both addition and multiplication, although the magnitude of the relative format cost should be greater for addition than multiplication.

When I used log RT in place of Raw RT as the dependent variable and conducted another two RM-ANOVAs for addition and multiplication each, however, the size x format interaction on addition disappeared. The results are shown in Table 5-4 and plotted in *Figure 5-1* (b). Main effects of problem size and format were not affected by the log transformation and remained significant. Similarly, the size x format interaction on multiplication remained nonsignificant.

Table 5-4 *Summary of RM-ANOVA: Log-Transformed Response Time as A Function of Problem Size and Format for Addition and Multiplication*

		Addition			Multiplication		
Effect	*df1, df2*	*MS*	*MSE*	*F*	*MS*	*MSE*	*F*
problem size	1, 33	3.086	.015	**209.62*****	4.33	.028	**155.37*****
format	1, 33	3.42	.027	**128.17*****	2.20	.027	**82.41*****
problem size x format	1, 33	.0003	.002	.16	.021	.006	3.79$^{\#}$

Note. *** $p < 0.001$. #: $p = .06$

For each ANOVA conducted above, I also conducted four LMMs to test whether dichotomizing problem size and the structure of random effects would affect the results. In each LMM, response time of each individual problem was regressed on its size and format at the trial level. Results from these LMMs are highly consistent with their corresponding RM-ANOVA with only one exception. LMM with the continuous problem size detected a size x format interaction on multiplication whereas ANOVA with the dichotomized problem size did not (see Table 5-2), with both analyses using raw RT and

random intercepts for subjects only. This example confirms decreased power associated with dichotomization (Maxwell & Delaney, 1993).

In summary, the current analysis replicated findings in Campbell and Alberts (2009). LMM and RM-ANOVA produced the same results in the current sample. Word format exacerbates the problem-size effect for addition, although this result is dependent on the scale of the dependent variable. Now that I have tested the size x format with both RM-ANOVA and LMM irrespective of participants' strategy, next, I test this effect as a function of strategy, examine the impact of problem format on problem size effect for retrieval and nonretrieval trials separately.

Size x Format x Strategy. The plots showing data fits are in *Figure 5-2* and *Figure 5-3*; frequency graphs showing the relative use of retrieval and non-retrieval strategies are shown in *Figure 5-4*. I refer to the latter figure to help interpret the results of the former.

For addition (raw data), LMM detected a size x format interaction when all trials were combined (stage one, also in *Figure 5-2*, #ARDW2), but this interaction becomes borderline or nonsignificant when retrieval and nonretrieval trials were separated. The slope of the problem-size effect for digit and word format is indistinguishable on both retrieval and nonretrieval trials. Comparing plot ARDW2 to ARDW3_* in *Figure 5-2* as well as in *Figure 5-4*, a parsimonious account for these results/patterns would be: a) format effect for addition problems only manifest during the encoding phase, b) the speed at which the answer to a given addition problem is accessed or calculated is not affected by problem format, and c) the size x format interaction on addition problems is primarily due to increased nonretrieval trials on large word addition problems. The last point needs

some elaboration. Nonretrieval trials are in general slower than retrieval (e.g., the two lines in ARDW3_2 are on average higher than two lines in ARDW3_1). Relatively more trials were solved via nonretrieval on large word addition than large digit addition (#1 & #2 in *Figure 5-4*). The combination of slower nonretrieval and more frequent nonretrieval usage pushed up the average RT on large word addition problems and hence the steeper slope.

Moving on to multiplication problems with raw RT, there was no size x format interaction before taking strategy into account (stage one, also in *Figure 5-3*, MRDW2). However, once strategy is entered into the equation, while the size x format interaction remained nonsignificant on retrieval trials, it became significant on nonretrieval trials (MRDW3_2). Participants showed a steeper problem-size effect on word problems than on digit problems when both are solved using nonretrieval. Campbell and Alberts (2009) found a similar non-significant interaction between problem size and format when all trials were analyzed, as well as on retrieval only trials. However, they were not able to test the effect of problem format on non-retrieval trials.

Comparing *Figure 5-3* multiplication (i.e., MRDW*) to their counterparts for addition (i.e., ARDW*), the following inferences can be made. Given the parallel slopes on retrieval trials, the encoding speed difference is sufficient to account for the main effect of format on multiplication. It is plausible that word format does not affect the access speed on a given multiplication problem. However, it seems to influence the calculating process, for example, through prolonging the execution of a nonretrieval procedure or biasing a solver towards multi-step procedures. Unlike addition, word format did not lead to increasing number of nonretrieval trials on large multiplication

problems, as shown in *Figure 5-4* (#4 & #5). This interpretation is also consistent with the nonsignificant size x format interaction when retrieval and nonretrieval trials were analyzed together (MRDW2 in *Figure 5-3*). Although nonretrieval trials take longer on average in word format than digit format, as suggested by the size x format interaction, the difference between digit and word format became diluted once retrieval and nonretrieval trials were combined, hence the nonsignificant size x format interaction in m221mrdw.

In summary, word format leads to more nonretrieval trials without affecting the access or calculation speed on addition problems, but on multiplication, it leads to longer calculation processes instead of increasing nonretrieval usage for large problems

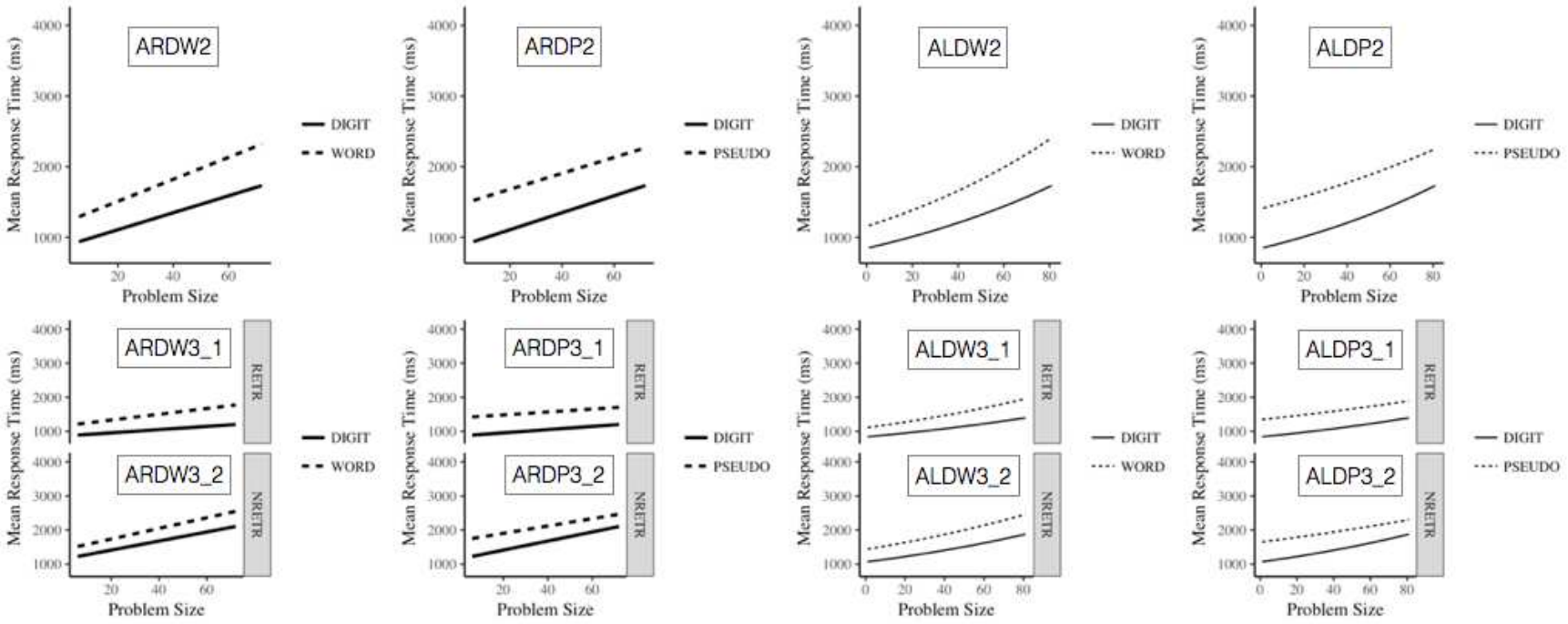

Figure 5-2 Response time as a function of problem size, format, and self-reported strategy for addition. The full set of trials are shown in the top row, and retrieval vs. nonretrieval in the top + bottom panels of the bottom row. R = raw RT, L = log RT, DW = digit versus

word, DP = digit versus pseudohomophone, 2 = two predictors (size, format), 3_1 = three predictors for retrieval trials (size, format, strategy), 3_2 = three predictors for nonretrieval trials.

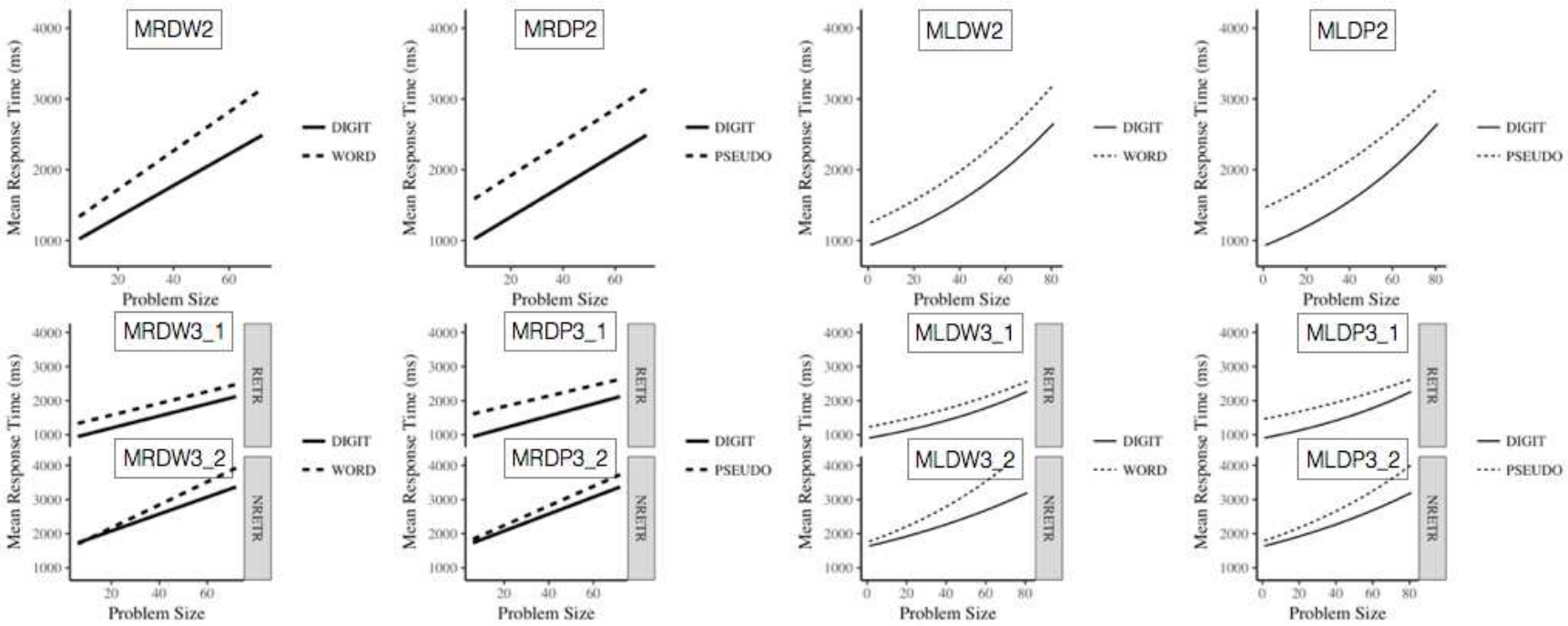

Figure 5-3 Response time as a function of problem size, format, and self-reported strategy for multiplication (M). The full set of trials are shown in the top row, and retrieval vs. nonretrieval in the top + bottom panels of the bottom row. R = raw RT, L = log RT, DW = digit versus word, DP = digit versus pseudohomophone, 2 = two predictors (size, format), 3_1 = three predictors for retrieval trials (size, format, strategy), 3_2 = three predictors for nonretrieval trials.

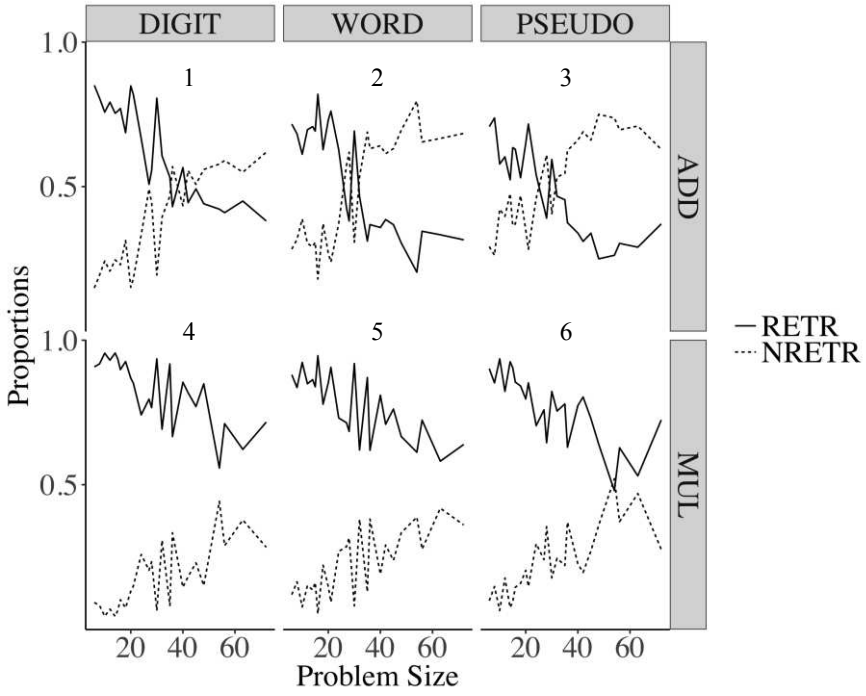

Figure 5-4 Percentages of trials solved via retrieval versus nonretrieval as a function of problem size (i.e., product) and format. Addition in panels 1 – 3; multiplication in panels 4 – 6.

Results for multiplication problems remained the same when I used log-transformed RT. However, the size x format interaction on addition problems became nonsignificant either when all trials were combined or when retrieval and nonretrieval trials were separated. The fact that results of statistical tests depend on the scale of dependent variable is typical for findings reported in this chapter. An interaction effect may be statistically significant when raw RT is the DV whereas the same effect becomes nonsignificant when log RT is the DV, or vice versa. This is because log transformation adjusted what used to be a positively skewed data distribution into a normal looking distribution. In doing so, the pile of data on the right side were squeezed whereas data on

the left side were stretched. The practical effect of this transformation is to reduce variability of data on the positive end and increasing variability on the negative end. Therefore, a positive effect flagged as significant when raw RT is the DV could become trivial once RT has been log transformed. Likewise, a trivial negative effect when raw RT is the DV could become significant after transformation.

Given the scale-dependency of results in this chapter, I emphasize the effect size instead of on statistical significance. In Table 5-5, I listed the effect size of each size x format interaction reported in previous ANOVAs (#3 & #6). I used effect size as a guide when interpreting statistical significance. The effect sizes of the size x format interaction were calculated similarly for all LMMs as they were previously for RM-ANOVAs (i.e., group mean differences), as shown in Table 5-5. These effects are easy to calculate and intuitive to understand, however, they serve as proxies for the more appropriate effect sizes, regression coefficients. The last two rows show the effects of words versus digits and for pseudohomophones versus digits for retrieval versus nonretrieval, and for the combined analyses. The superscripts indicate which effects were significant in which analyses (i.e., RM-ANOVA raw data, RM-ANOVA log-transformed data, LMM raw data, and LMM log-transformed data). For example, #3 was flagged as significant when raw RT was the dependent variable but became nonsignificant when log RT replaced raw RT.

Table 5-5 *Mean RT (ms) as a function of a 3(format) X 2(problem size) X 2(strategy) design*

	problem size (addition)						problem size (multiplication)					
	retrieval		nonretrieval		all		retrieval		nonretrieval		all	
	S	L	S	L	S	L	S	L	S	L	S	L
Digit	913	1074	1318	1785	1010	1426	1096	1584	2233	2756	1217	1855
Word	1283	1496	1605	2173	1384	1916	1486	1946	2238	3333	1605	2340
Pseudo	1453	1582	1811	2205	1593	1974	1713	2080	2292	3087	1805	2453
Word - Digit	370	422	287	388	374	490	390	362	5	577	388	485
Pseudo - Digit	540	508	493	420	583	548	617	496	59	331	588	598
Relative format cost (W-D)	[#1] 52		[#2] 101		[#3] 116[a, c]		[#4] -28		[#5] 572[c, d]		[#6] 97	
Relative format cost (P-D)	[#7] -32[d]		[#8] -73[d]		[#9] -35[d]		[#10] -121[d]		[#11] 272[c]		[#12] 10[d]	

[a] Significant in RM-ANOVA raw, [b] significant in RM-ANOVA log-transformed,
[c] significant in LMM raw, [d] significant in LMM log-transformed

Note. Relative format cost (W-D) = difference between word and digit on large problems - difference between word and digit on small problems. Relative format cost (P-D) = difference between pseudo word and digit on small problems - difference between pseudo word and digit on large problems. I indexed the statistics in the last two rows so that they can be easily referenced in the text.

Summary. My attempt to replicate Campbell and Alberts' (2009) analyses with the current data set had two components. First, without considering strategy, I found that both RM-ANOVA and LMM produced highly similar results to each other as well as to those of Campbell and Alberts. Note, however, that the size x format interaction on addition problems was sensitive to the scale of dependent variable as it disappeared once I log-transformed RT. Second, when taking into account strategy, the current analyses provides a fuller picture than those of Campbell and Alberts, although the two accounts are generally in agreement. Word format led to more frequent nonretrieval on large addition (*Figure 5-4*, #2) but to longer calculation times on large multiplication (*Figure 5-4*, MRDW*). Thus, the current analysis supported the notion that encoding stage is not insulated from the calculation stage of solving mental arithmetic (Campbell & Alberts, 2009).

The analyses in stage one were intended to replicate previous findings and does not test research hypotheses involving pseudohomophone problems. Next, I test the word frequency hypothesis outlined at the beginning of this chapter using LMM.

Stage Two

The goal of analyses at this stage is to compare digit and pseudohomophone format, and to contrast this dyad to its counterpart involving digit and word format, results of which were already reported in stage one. All models were tested using LMM with maximal random effects structure. Results from rawRT and logRT models were different. Unlike in stage one, I resumed the practice of prioritizing logRT over rawRT models when results differ between them. Therefore, unless otherwise stated, results

presented in this part were based on the logRT model. I present results on addition first, followed by those on multiplication.

Addition. On addition problems, a size x format interaction was detected on both retrieval and nonretrieval trials as well as on all trials (*Figure 5-2*, ALDP*; Table 5-5, #7 - #9). On average, addition problems were solved slower in pseudohomophone relative to digit format, but speed differences decreased as problem size increased. Note that although the main effects of format and problem size remained significant in the rawRT models, the format x size interactions were not statistically significant (Figure 5-2(a), ARDP*). Readers should interpret these results with caution.

The direction of the size x format interaction for the digit-pseudo dyad is opposite to the direction for the digit-word dyad on addition (Table 5-5, compare #3 and #9). The difference between digit and word problems intensified as problem size increased whereas the gap between digit and pseudohomohpone problems decreased. This difference in RT agrees with the different patterns of strategy reports between these two dyads. As shown in *Figure 5-4* (digit vs. pseudo, #1 vs. #3), large pseudohomophone problems were more frequently solved via nonretrieval strategies than were large digit problems, similar to the digit-word dyad, which slowed down large pseudohomophone problems on average. However, an almost comparable, if not larger increase of nonretrieval use was also observed for small pseudohomophone problems, which slowed down small pseudohomophone problems as well. The combined effect is that the problem-size effect is shallower on pseudohomophone problems than on digit problems, hence the negative interaction of size x format for the digit-pseudo dyad.

Multiplication. The size x format interaction was significant when all trials were analyzed in the log transformed model (MLDP2). On average, multiplication problems were solved slower in pseudohomophone relative to digit format, and such speed difference increased as problem size increased. This interaction was nonsignificant in the rawRT model, however. The significant size x format interaction in the logRT model was a surprising finding given the small magnitude of this interaction (10ms; see Table 5-5, #12). Its counterpart for the digit-word dyad has a larger magnitude but was nonsignificant in either logRT or rawRT model (97ms; see Table 5-5, #6).

There was also a size x format interaction on retrieval trials (MLDP3_1). The speed gap between digit and pseudohomophone problems solved via retrieval narrows as problem size increases. The same interaction was not significant in the rawRT model, however. If the significant interaction of size x format on retrieval trials in the logRT model is not a merely statistical artifact of log transformation, it can be inferred that pseudohomophone and word formats differ in terms of their relative format costs on multiplication problems solved by retrieval, because the size x format interaction on retrieval trials has a smaller magnitude for the digit-word dyad and was nonsignificant (-28ms, #4 in Table 5-5).

The size x format interaction on nonretrieval trials was nonsigificant (*Figure 5-2*, MLDP3_2), although the same interaction was detected in the rawRT model (MRDP3_2). This interaction has the same direction as its counterpart in the digit-word dyad (Table 5-5, #6 and #12). It is plausible that pseudohomophone format slowed down execution of procedures on large multiplication problems, as did word format.

Summary. Performance on pseudohomophone problems is distinguishable from that on word format problems, at least for addition. Although both formats are relatively less familiar than digits, they affect solution times differently in relation to digits. Solution time slows down on both word and pseudohomophone problems when compared to digit problems. However, the speed difference between digit and word format increases as a function of problem size whereas the speed difference between digit and pseudo word formats tends to decrease as a function of problem size. Furthermore, word format leads to more frequent nonretrieval usage compared to digits, but disproportionately on large problems. In contrast, pseudo word format leads to more nonretrieval on both small and large problems.

On multiplication problems, number word and pseudo number word formats are processed somewhat similarly in relation to digits. Both formats resulted in slowed retrieval of small problems more than large ones (Table 5-5, #4 & #10), although this trend was statistically significant only for the pseudo word format. Both word and pseudo word formats slowed down large problems more so than small ones when these problems were solved via nonretrieval (#5 & #11), although this trend was statistically significant only for word format.

Inconsistent findings between the logRT and rawRT models may undermine the credibility of results reported in the current study. Of all the results I have presented on the comparison between digit and pseudo word formats, for example, main effects remained the same in rawRT and logRT models but the significance of size x format interaction flipped between these two models. I discuss the implications of this situation in the Discussion as well as in the final chapter.

Discussion

Campbell and colleagues have repeatedly shown that encoding and processing stages are not insulated from each other (Campbell, 1999; Campbell & Alberts, 2009). Instead, simple arithmetic problems in different formats, such as 4x7 vs. four x seven, are sometimes solved via different routes and at different rates, even after encoding difficulties have been taken into account. Campbell and Alberts (2009) also proposed the mechanism of how problem format affects solution process. Specifically, atypical formats such as written number words reduce retrieval efficiency and promote nonretrieval strategies. In contrast, the current study tested an alternative explanation for the format effect. Participants solved problems in three formats: Arabic numerals, number words and pseudohomophones. Findings based on contrasting digit and number word formats replicated and extended the results of Campbell and Alberts (2009). Compared to digits, word format leads to more nonretrieval trials, but on multiplication, word format leads to slower calculation processes instead of increasing nonretrieval usage for large problems.

Findings based on contrasting number word and pseudo word formats in relation to digits did not support Campbell and Alberts' hypothesis on why word format exacerbates the problem-size effect. Both number word and pseudo word formats are relatively unfamiliar compared to digits, and yet solution times and strategy reports on pseudo word problems displayed different patterns compared to those on number word problems. The word frequency hypothesis, however, is compatible with the current findings. Assume that small digits, like small number words, are encountered more often than large digits. It is possible that the relatively higher frequency of small digits compared to large digits is an important component of what drives the problem-size

effect. However, the word frequency of pseudo number words does not decrease systematically with increasing magnitude. Therefore, the frequency-based advantage of small digit problems is lost on pseudo word problems, and hence we found a shallower problem-size effect on pseudo word format compared to digits. In short, familiarity with individual operands are at least in part responsible for problem-size effect.

I also compared ANOVA to LMM analyses in this study. With a balanced dataset, neither the structure of random effects nor dichotomizing problem size had significant impact on the results. There is no evidence to suggest keeping the structure of random effects maximal would reduce power. On the contrary, dichotomizing a continuous variable, as is the case in RM-ANOVA, may lower the statistical power of a test. The advantage of LMM over RM-ANOVA was most clear with unbalanced data, that is, when self-reported strategy was taken into account. The current study was able to examine response time for simple arithmetic solved via non-retrieval, and thus extended findings from Campbell and Alberts (2009).

One notable challenge in the current analyses was the dilemma presented by the skewed RT. Skewness of the dependent variable is typically dealt with through transformation to achieve normality when researchers are applying normality-based models, such as RM-ANOVA and LMM, in order to fulfill the normality assumption of these models. However, as was demonstrated, the models are sensitive to the distribution of the outcome. As a consequence, the interaction of primary interest (i.e. size x format) came and went depending on whether I used a raw or log-transformed dependent variable. Transforming the skewed dependent variable is a requirement of the chosen statistical model, but retaining the original scale of dependent variable is desirable for

recognizing mental processes reflected by extreme data. In the current study, I decided to place more confidence in the logRT model and to give priority to meeting assumptions of a statistical model. Researchers can back-transform predicted values into the original scale and plot them when interpreting statistics, especially if they are concerned that the log transformation would affect whether the log-transformed RTs are still interpretable as the duration of cognitive processes.

Response times are by nature positively skewed. Therefore, it is reasonable to assume existing studies that dealt with effect sizes similar to the current one (100 - 200 ms or smaller) would also have results dependent on the scale of dependent variable. As a rule of thumb, a fan-shaped interaction is more likely detected when the DV is kept raw, that is, positively skewed, whereas a reverse fan-shaped interaction is more likely detected when DV has been transformed and skewness-corrected. Given the wide use of response time in the literature as an outcome, it is important to have a principled way to reconcile conflicting results based on raw and transformed dependent variables. I made an attempt in the current analyses to do so by adopting an approach that emphasizes effect size. This example also echoes recent movement away from the autocracy of p values (Cumming, 2014). Lo and Andrews (2015) also proposed generalized linear mixed modelling (GLMM) as an alternative model which circumvents the need to transform a skewed dependent variable and still meet assumptions of the model.

CHAPTER 6: General Discussion

In this thesis, I have applied linear mixed modelling (LMM) to three datasets, all of which had been previously analyzed using general linear model (GLM). My goal was to test whether or not LMM was more suitable than GLM for modelling patterns of variation in latencies related to the problem-size effect, a pervasive finding in the field of mental arithmetic. The problem-size effect is the ubiquitous finding that arithmetic problems with large addends or multipliers, and hence large answers, take longer to solve and are more error prone than those with small operands and answers (Zbrodoff & Logan, 2005). The problem-size effect is assumed to reflect some of the principles underlying cognitive architecture of arithmetic, which explain questions such as how do people solve 7x8 (Ashcraft, 1992; Campbell, 1994; Rickard, 2005).

To uncover principles of cognitive architecture in this field, researchers have created different conditions in laboratory experiments that change the shape and chronometric characteristic of problem-size effects. Data collected in these experiments were used as the basis for making inferences about the structure and process of mental arithmetic. As described at the beginning of this dissertation, the way that these data have been analyzed historically sometimes imposes unreasonable assumptions and thus risks invalid conclusions. Given that statistical analysis serves as the basis of inferences and theorizing in this field, I proposed LMM as an alternative to traditional analysis methods because LMM eliminates some of the unnecessary assumptions and holds the promise of fitting the data better than alternatives, specifically, by-participant regression and RM-ANOVA. The current thesis was an attempt to improve the application of statistical modelling to the data sets involving simple arithmetic. The logic is this: If LMM made

any difference in a re-analysis, then the current research is a viable approach for advancing the theory. In this chapter, I briefly summarize my findings, emphasizing the differences between the traditional analyses and those I conducted using LMM.

All three datasets analyzed in this thesis feature repeated measures, with response latency as the dependent variable and a mix of continuous and categorical predictors. Three analyses tackled the same question from different angles: how do structural (e.g., problem size, tie status) and experiential factors (e.g., participants' educational background) contribute to the problem-size effect in mental arithmetic? Across three studies, a systematic change in RT on single-digit arithmetic problems strongly correlated with structural factors including the size and type of a problem. However, this pattern was modified by experiential factors to various extents, such as learning history of participants or their familiarity with operands. For example, Chinese participants showed an advantage on one operand order for addition (i.e., max + min) but the other for multiplication (min x max), partly because they drilled min x max but not max x min (Chapters 3 and 4). This experiential factor produces characteristic differences in mental processing. In contrast, Canadian-educated students show advantages on multiplication problems with 5 as an operand (Chapter 3), presumably because they practice fives using repeated addition separately from other nontie multiplication problems.

What did I find? In brief, my analyses indicate that the advantage of LMM over by-participant regression is cosmetic; both essentially gave the same results (Chapter 3). By-participant regression is less elegant and more labour-intensive than LMM but more intuitive. Users who are unfamiliar with LMM but want to avoid aggregation before

analysis and/or dichotomizing continuous predictors can use this technique on repeated measures data that have a balanced design.

Nevertheless, there were some advantages of LMM over RM-ANOVA that were exemplified in two cases, both of which featured unbalanced data. Specifically, although I only compared LMM to by-participant regression and piecewise model for the first dataset (Chapter 3) in the current thesis, a separate analysis of this dataset showed that the problem-size effect on 5-operand problems was more appropriately modelled with LMM than with RM-ANOVA (Ma, LeFevre, & Howard, 2016). ANOVA distorts the result because it requires dichotomizing problem size and thus masks the fact that 5-operand problems have a smaller range of problem size (10 - 45) compared to other non-tie problems (6 - 72). Similarly, in the third dataset (Chapter 5), the need to separate retrieval and nonretrieval trials resulted in an unbalanced dataset and missing cells. This situation is handled poorly by RM-ANOVA but with ease by LMM. In the past, researchers have tried to remedy similar problems through dichotomizing and data aggregation, but were not always successful. For example, Campbell and Alberts (2009) reported that three-quarters of participants could not be included when analyzing the data from trials solved via nonretrieval, despite data aggregation. Similarly, over two-thirds of participants in the dataset of Chapter 5 would have 3 or fewer observations in at least one cell if I were to apply RM-ANOVA to this dataset, which would render the results unreliable and unrepresentative.

Save for the above-mentioned two scenarios, LMM did not lead to different conclusions than what were already drawn using GLM. In the data analyzed in the preceding chapters, the ability to include more than one random effect in a model

afforded by LMM did not seem to affect findings. Nor did the discrete versus continuous characterization of problem size affect the conclusions. Given these results, is it worth the effort to apply LMM, given it tends to be more complicated to use than ANOVA and may require specialized software? As a partial answer to this question, I discuss the reasons why LMM did not make a difference in the current re-analyses and my stand on which technique is to be preferred.

In retrospect, the magnitude of problem-size effects examined in the current dissertation is perhaps large enough to withstand fluctuations of standard error due to changes in the structure of random effects. Specifically, RM-ANOVA only includes a random intercept whereas LMM analyses in the current thesis contain random intercepts and three or more random slopes. Yet the results in LMM largely replicated those from RM-ANOVA. However, it is almost impossible to know in advance the impact of different random effects structures on the conclusion before actually conducting the analysis. For example, mis-specifying a random slope as a fixed one for predictor A could reduce power for detecting the effect of predictor B if predictors A and B are unrelated, whereas the mis-specification could inflate the risk of Type I error if predictors A and B are confounded (Barr et al., 2013). Therefore, there is no sure way of telling whether results obtained from applying RM-ANOVA to the data would differ from that obtained from LMM. Importantly, estimates obtained in LMM, when the structure of random effects is properly set up, are less susceptible to Type-I error than those found with RM-ANOVA. From this view, LMM with a properly specified random effects structure is still preferred over RM-ANOVA.

In general, the present findings suggest that dichotomization of problem size into small and large categories seems sufficient to capture the systematic change of response time on single digit addition and multiplication, at least when the data are balanced. This conclusion has a caveat, however. Most existing theories treat problem size as a proxy for problem difficulty and make no explicit prediction beyond differences between small versus large problems. Therefore, it is hard to hypothesize anything other than a linear relationship between RT and problem size. Given that the problem size effect is robust, the power recovered from treating problem size as continuous rather than dichotomous does not affect this conclusion. Nevertheless, preserving the continuous nature of a predictor such as problem size allows one to look for a non-linear relation between independent and dependent variables. This advantage might be more obvious for situations where different cognitive processes drove the relation as the independent variable takes on different values. If, for example, I had modelled different slopes for single-digit addition with sum <= 10 and sum > 10, I could have tested the hypothesis that these two groups of problems were solved differently (e.g., retrieval vs. non-retrieval), at least to the extent that I had a reason to predict different slopes in these problem-size ranges.

In summary, the current thesis showed that the advantage of LMM over GLM was most obvious when applied to unbalanced datasets (Quené & van den Bergh, 2004). For example, researchers could use LMM to examine the problem-size effect for retrieval and nonretrieval trial separately, an analysis that used to be problematic or even impossible with GLM in the past. This type of analysis is becoming increasingly important because participants of the current generation more likely use a mix of retrieval and non-retrieval

strategies when they solve arithmetic problems, and less likely use retrieval exclusively, an assumption that dominated earlier research on mental arithmetic. It is important that behavioural data (e.g., RT) are separately analyzed based on which type of mental processes underlies these data. At the minimum, non-retrieval trials should not be mixed with retrieval trials when the purpose of the analysis is to understand mental representation of arithmetic facts in memory. In balanced datasets, analyses using LMM and GLM tend to give similar results. However, where the findings do differ, the results using LMM were more trustworthy in terms of being faithful to the raw data (e.g., the pattern for five-operand problems). Furthermore, LMM also invites researchers to ask questions that they have avoided because the traditional analysis was inadequate, such as a non-linear relation between problem size and latencies. Next, I offer some advice based on personal experience for researchers who are new to LMM and want to adopt this technique. These tips may be particularly relevant for current users of SPSS who are used to ANOVA.

Issues Related to the Use of LMM

Deciding the structure of random effects. In the current research, I have demonstrated two ways to set up the structure of random effects for a linear mixed model. Specifically, I demonstrated a visual approach in Chapter 3, where I plotted data for each individual. If individual plots look dissimilar, for instance, if some slopes appear steep and some shallow, then the slope is set as a random effect. A second approach, using the maximal random effects structure justified by the experimental design, was implemented in Chapters 4 and 5. The specific principles for deciding which effect should be random is described on page 45 as well as in Barr et al. (2013, p. 262). The same challenge

persisted when I used both approaches, namely, having to decide which random effects to keep and which ones to omit when the structure of random effects for a model was theoretically justified but the model failed to converge in practice. The current literature of LMM has recognized this challenge (T. S. Clark & Linzer, 2015). However, to the best of my knowledge, statisticians have yet to reach a consensus on the best practice to achieve a balance between accuracy and complexity. Readers can find an elaborate discussion of this topic and some tentative solutions in Barr et al. (2013, p. 275-276)

Which software should be used? LMM analysis is available in most mainstream statistical software packages, including SPSS (v.11 or higher), SAS, and R. SPSS has the convenience of a point-and-click interface. However, it is severely limited in performing model diagnostic analysis and generating graphic output for models fitted using LMM (West & Galecki, 2012). This shortcoming of SPSS makes it a less desirable choice for LMM modelling. Instead, three commonly used stats packages for running LMM analysis are STATA, SAS and R, all of which use scripting languages. Accordingly, users need to write syntax for specifying models and selecting desired options, which might entail a steep learning curve for people who are used to graphical interfaces such as SPSS. In short, researchers who are used to ANOVA and SPSS may face double challenges in applying LMM: They need to both understand the technique conceptually and learn to navigate the mechanics of implementing the model with a software package.

In the current research, I chose R instead of SAS for two reasons. R is freely available whereas SAS is proprietary. R is available across different operating systems whereas SAS is only available for Windows (mac users either run SAS on a server or a virtual machine). Another advantage of R is users can tap into the expertise of a group of

dedicated users. They are very active on the internet and always ready to answer questions from other users.

Notably, however, R also has drawbacks. For example, because R is open source and any user can contribute R modules, modules authored by different people tend to create redundancy. Typically multiple modules exist for one function such as plotting; but some modules are more suitable than others given a person's analytic goal. Spending time on selecting the most appropriate module for each task could bog down a less experienced user. One of the contributions of the current thesis is provide a full script for applying LMM to data in mental arithmetic, including model diagnosis, analysis, and visualizing results (online appendix available at https://github.com/xiaobitou/LMM). These scripts can also be adapted to analyzing data for related lines of research in math cognition that also tend to use GLM.

Limitations of the Current Thesis

There are several limitations of the current thesis. It has become clear that LMM and GLM sometimes produce comparable results. As a result, the question I raised at the beginning of Chapter 1, whether or not LMM is more suitable than GLM, has only been partially answered. To fully answer the question that I have raised, future research must define the boundary conditions: When is LMM likely to give different results than GLM? This re-framed question will likely require a larger variety of datasets than the three datasets I selected for this thesis. For example, in studies that use double- or triple-digit numbers (e.g., 35 + 79) so that only a subset of all the possible stimuli is tested, applying LMM with by-item random intercepts may lead to different results compared to GLM.

I analyzed three datasets in the current thesis. The analyses were completed over a two-year period of time, during which my understanding of LMM deepened and practical skills of implementing the analysis improved through solving a series of challenges. I applied newly acquired understanding and skill as I continued the analyses, sometimes creating inconsistency between earlier and later analyses. For example, in specifying the random effect structure, I used visual inspection in Chapter 3 and a theory-driven approach in Chapters 4 and 5. Although it is very unlikely that these inconsistencies had any impact on findings in the current thesis, if time allowed, I would re-do some of the analyses based on the best practice I have learned so far and maintain as much consistency as possible to maximize rigor.

One remaining issue that arose during these analyses was the question of whether skewed data should be log-transformed before analysis. In the analyses for this thesis, I log-transformed the dependent variable, RT, for all three datasets to meet the assumption of normality. Some findings in the first (Chapter 3) and third dataset (Chapter 5) were different for raw versus log-transformed data. For example, in Chapter 5, the format x size interaction for addition problems was detected in an RM-ANOVA when raw RT was the dependent variable but not when RT had been log-transformed. I decided to place more emphasis on findings from the logRT model in the current research. However, this decision can be criticized. Log-transforming data disproportionally affects observations in the upper tail, which tend to be those with large problem size, and hence artificially changing the pattern of problem-size effects. In other words, different cells of the design are differentially affected by the transformation. In my opinion, log-transforming the dependent variable is not an ideal solution to meet the normality assumption. Instead,

there are at least two directions for future research to take. First, in any given study, researchers could use simulation to test how robust the planned model is to violation of normality, and if its robustness is acceptable, then apply the model to raw data. Second, in view of the prevalent skewness among RT-based experiments, researchers could explore statistical models that do not require normality, such as generalized linear mixed models (i.e. GLMM; Bolker, 2015).

A couple of additional analyses could strengthen conclusions in the current thesis as well. For example, I could apply piecewise analysis to datasets in Chapter 4 and 5 as well, to replicate and extend findings in Chapter 3, that is, the problem-size effect levels off on large problems. Future research could examine problem-size effects with piecewise analyses. Ideally, these analyses should be guided by a theoretical framework so that findings could be interpreted accordingly.

Conclusions

This dissertation began with one question: would LMM be better suited than GLM for modelling data related to the problem-size effect? Additional questions were raised as I pursued the answer to this question. For example, treating problem size as continuous affords the opportunity to model a nonlinear relation between the two, however, there is currently a lack of theory to inform specifically what shape to expect. This predicament can be remedied by searching for alternative predictors of latency that are directly tied to problem difficulty, of which problem size is only a proxy. Such predictors can be derived from a computational model such as Campbell (1995) or Verguts and Fias (2005), in which the modeller assigned numbers to quantify the

associative strength/interference between each pair of operands and their sum/product. These numbers can replace problem size and be used to predict RT.

In general, LMM is preferred over GLM for analyzing repeated measures data that have a mixture of continuous and discrete predictors (Hoffman & Rovine, 2007). Granted, most researchers in the field of math cognition are familiar with RM-ANOVA and may feel reluctant to learn a different statistical technique, especially if RM-ANOVA is sometimes just as good as LMM. However, the time investment required to become proficient in LMM would be offset by the flexibility afforded by it in dealing with unbalanced data, or answering research questions that involves 5-operand problems or other special cases. LMM is capable of accommodating a wider range of data structures than RM-ANOVA and is less likely to break down with an unbalanced design. By adopting LMM in this field collectively, it would help standardize results reporting, encourage replication and comparison across studies, and ultimately contribute to theoretical advances in this field.

References

Ashcraft, M. H. (1987). Children's Knowledge of Simple Arithmetic: A Developmental Model and Simulation. In J. Bisanz, C. J. Brainerd, & R. Kail (Eds.), *Formal Methods in Developmental Psychology* (pp. 302–338). Springer New York. Retrieved from http://link.springer.com.proxy.library.carleton.ca/chapter/10.1007/978-1-4612-4694-7_9

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*(1–2), 75–106. https://doi.org/10.1016/0010-0277(92)90051-I

Ashcraft, M. H., & Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(5), 527–538. https://doi.org/http://dx.doi.org.proxy.library.carleton.ca/10.1037/0278-7393.4.5.527

Ashcraft, M. H., & Christy, K. S. (1995). The Frequency of Arithmetic Facts in Elementary Texts: Addition and Multiplication in Grades 1-6. *Journal for Research in Mathematics Education*, *26*(5), 396–421. https://doi.org/10.2307/749430

Ashcraft, M. H., & Stazyk, E. H. (1981). Menatal addition: A test of three verification models. *Memory & Cognition*, *9*(2), 185–196. https://doi.org/10.3758/BF03202334

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bolker, B. M. (2015). Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological Statistics* (pp. 309–333). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672547.003.0014

Brysbaert, M. (1995). Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, *124*(4), 434–452. https://doi.org/10.1037/0096-3445.124.4.434

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Butterworth, B., Zorzi, M., Girelli, L., & Jonckheere, A. R. (2001). Storage and retrieval of addition facts: The role of number comparison. *The Quarterly Journal of Experimental Psychology Section A*, *54*(4), 1005–1029. https://doi.org/10.1080/713756007

Campbell, J. I. D. (1987). Network interference and mental multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 109–123. https://doi.org/10.1037/0278-7393.13.1.109

Campbell, J. I. D. (1994). Architectures for numerical cognition. *Cognition*, *53*(1), 1–44. https://doi.org/10.1016/0010-0277(94)90075-2

Campbell, J. I. D. (1995). Mechanisms of simple addition and multiplication: A modified network-interference theory and simulation. *Mathematical Cognition*, *1*(2), 121–164.

Campbell, J. I. D. (1999). The surface form×problem size interaction in cognitive arithmetic: evidence against an encoding locus. *Cognition*, *70*(2), B25–B33. https://doi.org/10.1016/S0010-0277(99)00009-8

Campbell, J. I. D., & Alberts, N. M. (2009). Operation-specific effects of numerical surface form on arithmetic strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 999–1011. https://doi.org/http://dx.doi.org.proxy.library.carleton.ca/10.1037/a0015829

Campbell, J. I. D., & Beech, L. C. (2014). No Generalization of Practice for Nonzero Simple Addition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, No Pagination Specified. https://doi.org/10.1037/xlm0000003

Campbell, J. I. D., & Epp, L. J. (2004). An Encoding-Complex Approach to Numerical Cognition in Chinese-English Bilinguals. *Canadian Journal of Experimental Psychology*, *58*(4), 229–244. https://doi.org/10.1037/h0087447

Campbell, J. I. D., & Fugelsang, J. (2001). Strategy choice for arithmetic verification: effects of numerical surface form. *Cognition*, *80*(3), B21–B30. https://doi.org/10.1016/S0010-0277(01)00115-9

Campbell, J. I. D., & Graham, D. J. (1985). Mental multiplication skill: Structure, process, and acquisition. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *39*(2), 338–366. https://doi.org/10.1037/h0080065

Campbell, J. I. D., & Gunter, R. (2002). Calculation, culture, and the repeated operand effect. *Cognition*, *86*(1), 71–96. https://doi.org/10.1016/S0010-0277(02)00138-5

Campbell, J. I. D., & Metcalfe, A. W. S. (2007). Arithmetic rules and numeral format. *European Journal of Cognitive Psychology*, *19*(3), 335–355. https://doi.org/10.1080/09541440600717610

Campbell, J. I. D., & Oliphant, M. (1992). Chapter 9 Representation And Retrieval Of Arithmetic Facts: A Network-Interference Model And Simulation. In Jamie I.D. Campbell (Ed.), *Advances in Psychology* (Vol. Volume 91, pp. 331–364). North-Holland. Retrieved from http://www.sciencedirect.com/science/article/pii/S0166411508608912

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

Chen, Y., & Campbell, J. I. D. (2016). Operator priming and generalization of practice in adults' simple arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(4), 627–635. https://doi.org/10.1037/xlm0000196

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Clark, T. S., & Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods*, *3*(02), 399–408. https://doi.org/10.1017/psrm.2014.32

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods*, *7*(1), 41–63. https://doi.org/http://dx.doi.org.proxy.library.carleton.ca/10.1037/1082-989X.7.1.41

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dagenbach, D., & McCloskey, M. (1992). The organization of arithmetic facts in memory: Evidence from a brain-damaged patient. *Brain and Cognition*, *20*(2), 345–366. https://doi.org/10.1016/0278-2626(92)90026-I

De Visscher, A., & Noël, M.-P. (2014). The detrimental effect of interference in multiplication facts storing: Typical development and individual differences. *Journal of Experimental Psychology: General*, *143*(6), 2380–2400. https://doi.org/10.1037/xge0000029

Dehaene, S., & Cohen, L. (1997). Cerebral Pathways for Calculation: Double Dissociation between Rote Verbal and Quantitative Knowledge of Arithmetic. *Cortex*, *33*(2), 219–250. https://doi.org/10.1016/S0010-9452(08)70002-9

Fayol, M., & Thevenot, C. (2012). The use of procedural knowledge in simple addition and subtraction problems. *Cognition*, *123*(3), 392–403. https://doi.org/10.1016/j.cognition.2012.02.008

Geary, D. C., Bow-Thomas, C. C., Liu, F., & Siegler, R. S. (1996). Development of Arithmetical Competencies in Chinese and American Children: Influence of Age, Language, and Schooling. *Child Development*, *67*(5), 2022–2044. https://doi.org/10.1111/j.1467-8624.1996.tb01841.x

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53. https://doi.org/10.1214/009053604000001048

Groen, G. J., & Parkman, J. M. (1972). A chronometric analysis of simple addition. *Psychological Review*, *79*(4), 329.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117.

Jost, K., Khader, P., Burke, M., Bien, S., & Rösler, F. (2009). Dissociating the solution processes of small, large, and zero multiplications by means of fMRI. *NeuroImage*, *46*(1), 308–318. https://doi.org/10.1016/j.neuroimage.2009.01.044

LeFevre, J.-A., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S. L., & Sadesky, G. S. (1996). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General*, *125*(3), 284–306. https://doi.org/10.1037/0096-3445.125.3.284

LeFevre, J.-A., Lei, Q., Smith-Chant, B. L., & Mullins, D. B. (2001). Multiplication by eye and by ear for Chinese-speaking and English-speaking adults. *Canadian*

Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, *55*(4), 277–284. https://doi.org/10.1037/h0087374

LeFevre, J.-A., & Liu, J. (1997). The Role of Experience in Numerical Skill: Multiplication Performance in Adults from Canada and China. *Mathematical Cognition*, *3*(1), 31–62. https://doi.org/10.1080/135467997387470

LeFevre, J.-A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216–230. https://doi.org/10.1037/0278-7393.22.1.216

LeFevre, J.-A., Shanahan, T., & DeStefano, D. (2004). The tie effect in simple arithmetic: An access-based account. *Memory & Cognition*, *32*(6), 1019–31.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*(1), 181–190. https://doi.org/http://dx.doi.org.proxy.library.carleton.ca/10.1037/0033-2909.113.1.181

McCloskey, M., Caramazza, A., & Basili, A. (1985). Cognitive mechanisms in number processing and calculation: Evidence from dyscalculia. *Brain and Cognition*, *4*(2), 171–196. https://doi.org/10.1016/0278-2626(85)90069-7

Metcalfe, A. W. S., & Campbell, J. I. D. (2007). The role of cue familiarity in adults' strategy choices for simple addition. *European Journal of Cognitive Psychology*, *19*(3), 356–373. https://doi.org/10.1080/09541440600872001

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: Comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 46–60. https://doi.org/10.1037/0278-7393.10.1.46

Niedeggen, M., & Rösler, F. (1999). N400 Effects Reflect Activation Spread During Retrieval of Arithmetic Facts. *Psychological Science*, *10*(3), 271–276. https://doi.org/10.1111/1467-9280.00149

Noël, M.-P., Fias, W., & Brysbaert, M. (1997). About the influence of the presentation format on arithmetical-fact retrieval processes. *Cognition*, *63*(3), 335–374. https://doi.org/10.1016/S0010-0277(97)00009-7

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication*, *43*(1–2), 103–121. https://doi.org/10.1016/j.specom.2004.02.004

Racine, J. S. (2017). *A Primer on Regression Splines*. Retrieved from https://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Rickard, T. C. (2005). A Revised Identical Elements Model of Arithmetic Fact Representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 250–257. https://doi.org/10.1037/0278-7393.31.2.250

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*(3), 258–275. https://doi.org/10.1037/0096-3445.117.3.258

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31–76). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Stazyk, E. H., Ashcraft, M. H., & Hamann, M. S. (1982). A network approach to mental multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(4), 320–335. https://doi.org/10.1037/0278-7393.8.4.320

Thevenot, C., Fanget, M., & Fayol, M. (2007). Retrieval or nonretrieval strategies in mental arithmetic? An operand recognition paradigm. *Memory & Cognition*, *35*(6), 1344–1352. https://doi.org/10.3758/BF03193606

Trbovich, P., & LeFevre, J.-A. (2003). Phonological and visual working memory in mental addition. *Memory & Cognition*, *31*(5), 738–745. https://doi.org/10.3758/bf03196112

Verguts, T., & Fias, W. (2005). Interacting neighbors: A connectionist model of retrieval in single-digit multiplication. *Memory & Cognition*, *33*(1), 1–16. https://doi.org/10.3758/BF03195293

Welford, A. T. (1960). The measurement of sensory-motor performance: Survey and reappraisal of twelve years' progress. *Ergonomics*, *3*, 189–230. https://doi.org/10.1080/00140136008930484

West, B. T., & Galecki, A. T. (2012). An Overview of Current Software Procedures for Fitting Linear Mixed Models. *The American Statistician*, *65*(4), 274–282. https://doi.org/10.1198/tas.2011.11077

Zbrodoff, N. J., & Logan, G. D. (2005). What everyone finds: The problem-size effect. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 331–345). New York: Psychology Press.

Zhou, Z., & Peverly, S. T. (2005). Teaching addition and subtraction to first graders: A Chinese perspective. *Psychology in the Schools*, *42*(3), 259–272. https://doi.org/10.1002/pits.20077