Psychometric analyses of computer adaptive practice data

A new window on cognitive development

Hofman, A.D.

Publication date
2018

Document Version
Final published version

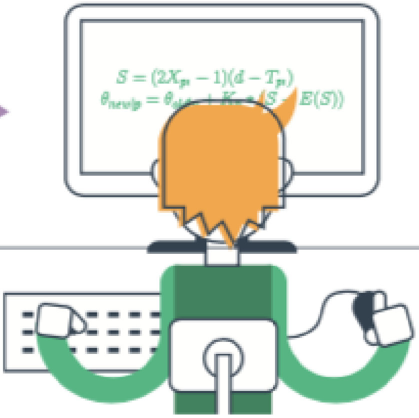
License
Other

Link to publication

Citation for published version (APA):
Hofman, A. D. (2018). *Psychometric analyses of computer adaptive practice data: A new window on cognitive development*. [Thesis, fully internal, Universiteit van Amsterdam].

Psychometric Analyses of Computer Adaptive Practice Data: **A New Window on Cognitive Development**

Abe Hofman


$S = (2X_{pi} - 1)(d - T_{pi})$
$\theta_{new|p} = \theta_{old|p} + K_p(S - E(S))$

Psychometric Analyses of Computer Adaptive Practice Data: A New Window on Cognitive Development

ABE HOFMAN

Hofman, Abe D.
Psychometric Analyses of Computer Adaptive Practice Data: A New Window on Cognitive Development

Ph.D. Thesis University of Amsterdam
Printed by Ridderprint, Ridderkerk
ISBN: 978-94-6299-934-3
Cover by Melina Ruijter (www.melina.amsterdam)
This publication is typeset in LaTeXusing the Memoir class

This reseach was supported by the Netherlands Organisation for Scientific Reseach
(NWO Research Talent Grant: 406-11-163)

Psychometric Analyses of Computer Adaptive Practice Data: A New Window on Cognitive Development

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K. I. J. Maex

ten overstaan van een door het College voor Promoties ingestelde

commissie, in het openbaar te verdedigen in de Agnietenkapel

op vrijdag 20 april 2018, te 14:00 uur

door Abe Dirk Hofman

geboren te Utrecht

Promotiecommissie

Promotor:	Prof. dr. H. L. J. van der Maas	Universiteit van Amsterdam
Copromotoren:	Dr. I. Visser	Universiteit van Amsterdam
	Dr. B. R. J. Jansen	Universiteit van Amsterdam
Overige leden:	Prof. dr. D. Borsboom	Universiteit van Amsterdam
	Prof. dr. H. M. Huizenga	Universiteit van Amsterdam
	Prof. dr. G. K. J. Maris	Universiteit van Amsterdam
	Dr. R. A. Kievit	Cambridge Neuroscience, MRC
	Dr. M. Straatemeier	Oefenweb
Faculteit:	Faculteit der Maatschappij- en Gedragswetenschappen	

Contents

1 General Introduction

This chapter is partly based on: van der Maas H. L. J., Kan K.-J., Hofman, A. D. and Raijmakers M. E. J, (2012). Dynamics of Development: A Complex Systems Approach. *Handbook of Developmental Systems Theory & Methodology.*
And: Brinkhuis M. J. S, Savi O. A., Coomans, F., Hofman A. D., van der Maas H. L. J. and Maris G. (Submitted). Learning As It Happens: Advances in Computerized Adaptive Practice.

1.1 Introduction

Large longitudinal data sets are required to answer fundamental questions on cognitive development and learning (Adolph, Robinson, Young, & Gill-Alvarez, 2008; P. C. Molenaar, 2004). To capture the developmental patterns, data should be collected while people develop or learn. Math Garden, a web-based adaptive training and monitoring system for primary education, is developed to do so, to capture the development in mathematical skills with intensive time-series for large numbers of children (Klinkenberg, Straatemeier, & van der Maas, 2011; Straatemeier, 2014).[1] Math Garden includes a wide set of games that students can use to practise well-defined mathematical skills (e.g., counting and multiplication) and math-related skills (e.g., proportional reasoning and working-memory). Each game consists of a large item bank with items of various degrees of difficulty so that children can be presented with items that match their proficiency.

Math Garden started out as a research project in 2007 at the University of Amsterdam and was commercialized in 2009 in response to the increasing popularity of the system. This was the start of Oefenweb (`www.trackandteach.com` and `www.oefenweb.nl`), a spin-off company that develops and hosts Math Garden and other adaptive practice systems for Dutch and English language learning, statistics (Groeneveld, 2014; Klinkenberg, 2014) and typing (van den Bergh, Hofman, Schmittmann, & van der Maas, 2015). Currently (August, 2017), ten years later, Math Garden covers 22 domains including in total 32.720 items, totaling 831,280,316 responses of 713,985 users, collected roughly at a rate of one million responses a day.[2] This dissertation explores this unique but complicated data set with the aim to elucidate processes of cognitive development. We first explain what Math Garden is and describe the psychometric properties of Math Garden. Second, we highlight some of the research that has been performed with Math Garden and relate chapters of this dissertation to these previous studies. Third and finally, we introduce possibilities for innovative analyses of longitudinal data sets from Math Garden, which are elaborated in chapters of this dissertation. Each chapter of this dissertation is introduced shortly.

1.2 Key Concepts of Math Garden

Math Garden is designed for children to practice math at school or at home instead. Since children play voluntarily, the system must be rewarding and engaging. Math Garden was designed to be playful, adaptive and provide direct feedback after every response; this is accomplished in a number of ways. First, tasks are setup as games and reward for practice is immediate and visible in Math Garden. Figure 1.1 shows a screenshot of Math Garden's landing page. Each plant in the garden represents a game that can be selected to start a practice session. Children have to visit the games regularly to prevent the plants from withering. Second, items match proficiency, which is very diverse among children, see for

[1] in Dutch the system is called Rekentuin; see `www.rekentuin.nl`

[2] the most frequent player of the school year 2016-2017 solved 48,231 items.

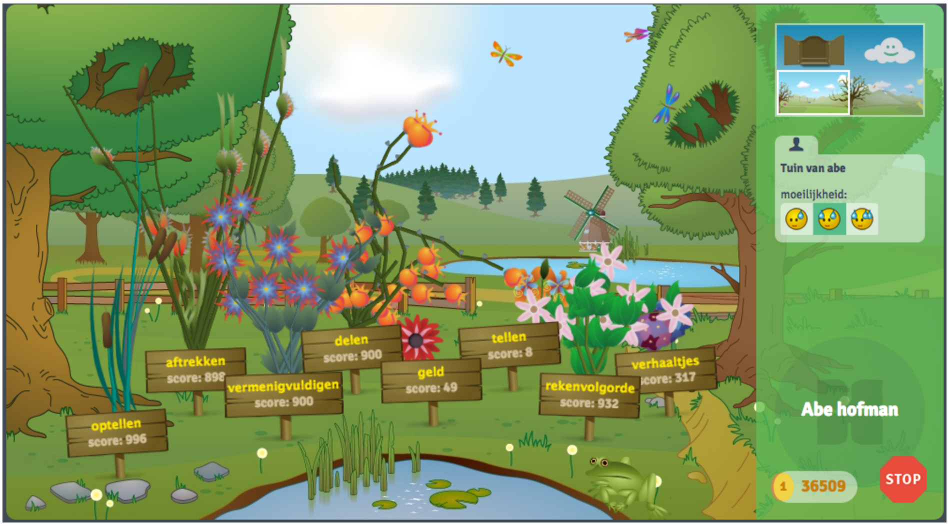


Figure 1.1: The landing page of Math Garden. Each plant represents a game and the score on the plant's tag represents the ability of the player (on a scale of 1 to 1000). Players can select game difficulty level by clicking the figure on the right with one, two or three 'sweat' drops. These are associated with the chance of solving the items correctly of either 90, 75 or 60 percent respectively.

example Straatemeier (2014, p. 13) and Dowker (2005). Third, feedback is presented after each response to facilitate personalized learning and to reward effort.

1.3 Psychometrics of Math Garden

The basis of Math Garden is an extension of classic computerized adaptive testing (CAT) methods. CAT is a testing method based on item response theory, which consists of a large family of item response models. The method used by Math Garden relies on the simplest item response model, the 1-PL or Rasch model:

$$P(x = 1|\theta_p, \beta_i) = \frac{exp(\theta_p - \beta_i)}{1 + exp(\theta_p - \beta_i)}$$

where the probability of a correct response depends on the difference between the person ability (θ_p) and the item difficulty (β_i). The Rasch model non-trivially assumes one-dimensionality (a single underlying trait is measured by all items within the item bank) and conditional independence (the response probabilities are independent conditionally on the underlying trait).

In CAT, the order of presentation of items depends on how a person responded to previous items (Wainer, 2000): if the immediately preceding response was correct, a more difficult item is presented next and vice versa. The advantage of using CAT is that abilities, such as arithmetic ability, can be estimated using fewer items than in standard tests.

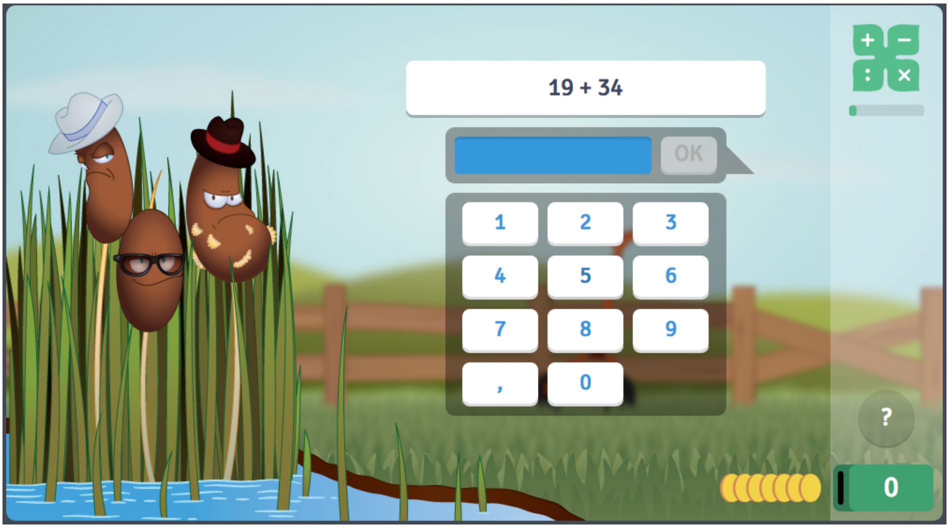


Figure 1.2: An example item of the addition game. Players can use either the keyboard or click a numeric keypad on the screen to submit their answer, and if they are unsure of the answer they can use the question-mark button on the right. The coins at the bottom visualize the implemented scoring rule (see section Psychometrics of Math Garden) and are collected after a correct response and subtracted after an incorrect response.

Currently, CAT is primarily used for testing, but in Math Garden it is used for testing and training at the same time. Therefore, Klinkenberg et al. (2011) introduced Math Garden as a computer adaptive practice (CAP) system.

Math Garden uses an extended CAT technique based on two crucial innovations: (1) incorporating both speed and accuracy in game scores and (2) updating children's scores and item difficulties real-time using the Elo-algorithm.

The first innovation, the use of both accuracy and response time when updating ability and difficulty estimates, takes place by means of a new scoring rule (Maris & van der Maas, 2012; Klinkenberg et al., 2011). Response times are included because they provide important additional information about a child's ability, and makes computerized practice more game-like. In Math Garden, items usually have a time limit of twenty seconds. According to the implemented signed residual time (SRT) scoring rule, the score equals the remaining time (twenty seconds minus response time) for correct responses, but -1 times the remaining time for incorrect responses (see Figure 1.3). As a consequence, guessing is risky, and if a child has no clue about the answer, he or she can best refrain from responding, which provides a score of zero. The SRT scoring rule is presented visually through coins at the bottom of the screen that refer to the remaining time (see Figure 1.2). Due to this visualisation, with a coin disappearing every second, even very young children can understand the scoring rule. This new scoring rule has two important advantages. First, it solves the notorious speed-accuracy trade-off problem (Wickelgren, 1977) since children now know how speed and accuracy are weighted in the scoring. Second, Maris and van der Maas (2012) have shown that under certain mild statistical assumptions, this scoring rule implies a standard two-parameter IRT model, in which discrimination is a linear function of the time limit of an item. Therefore, much is known about the properties of the measurement model, for example about the marginal and conditional distributions of the model estimates.

The score can formally be expressed as:

$$S = (2X_{pi} - 1)(d - T_{pi}),$$

where X_{pi} is 1 for correct and 0 for an incorrect response for player p at item i, T_{pi} is the response time and d the imposed deadline. The expected score ($E(S)$) that follows from the measurement model based on the SRT scoring rule is:

$$E(S|\theta_p, \beta_i) = \frac{exp(2d(\theta_p - \beta_i)) + 1}{exp(2d(\theta_p - \beta_i)) - 1} - \frac{1}{(\theta_p - \beta_i)}.$$

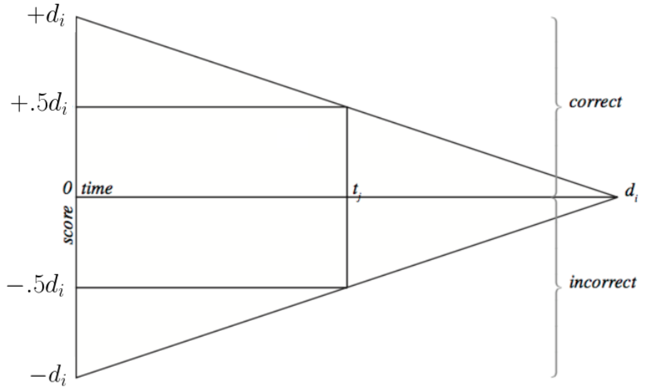


Figure 1.3: A visualization of the signed residual time scoring rule. If a player provides a correct response at time t_j his score is the remaining time. If the response is incorrect the score is -1 times the remaining time.

The second innovation is the use of an 'on the fly' Elo estimation algorithm (Klinkenberg et al., 2011), which originates from chess competitions (Elo, 1978). Elo-estimation provides a self-organizing testing system in which both the ability estimates of children and the difficulty estimates of items are continually updated in real-time immediately after a child's response. The reliability of the Elo-estimation system has been well studied analytically and in simulations (Batchelder & Bershad, 1979; Glickman, 1999; Pelánek, 2016; Pelánek, Papoušek, Řihák, Stanislav, & Nižnan, 2016). The most prominent advantage of this system is that it does not require the time-intensive and expensive procedure of pre-testing items as in normal CAT, which involves estimating item difficulty beforehand with hundreds of responses to every single item in the item bank that also contains about a thousand items.

In the Elo algorithm the updated estimates are based on a weighed sum of previous estimates and the difference between the observed and the expected score:

$$\theta_{new|p} = \theta_{old|p} + K_p * (S - E(S))$$

$$\beta_{new|i} = \beta_{old|i} - K_i * (S - E(S)).$$

The K-factor serves as a smoothing function that determines the weight of the current response in the updating of the parameters and implements a bias-variance trade-off. The K-factor in Math Garden increases when children repeatedly score below or above the expected score or when they are new to the system (see Klinkenberg et al. (2011), Straatemeier (2014) or the General Discussion Section for a more detailed description of the algorithm).

1.4 Research with Math Garden

The growing popularity of Math Garden provides researchers with an invaluable data set. The research in the current thesis can be broadly categorized by three different approaches. The first approach is based on direct analyses of person ability and item difficulty parameters that follow from the system. Earlier research with this approach was conducted by Klinkenberg et al. (2011), van der Ven, Straatemeier, Jansen, Klinkenberg, and van der Maas (2015); van der Ven, Klaiber, and van der Maas (2017) and Gierasimczuk, van der Maas, and Raijmakers (2013). Klinkenberg et al. (2011) show that the person parameters of various arithmetic games correlate highly with more traditional tests. Additionally, the work of van der Ven et al. (2015, 2017) shows that item parameters match the effects predicted by different theoretical models about mathematics. Furthermore, Gierasimczuk et al. (2013) and van der Maas and Nyamsuren (2017) show that the person and items parameters in both the Deductive Mastermind game and a Number Series game can be explained by substantive models developed for these cognitive tasks. These results provide support for the validity of some of the games in Math Garden. In Chapter 2, we use this approach as well and analyze the item parameters of the counting game to investigate different enumeration strategies. We further support our findings with a computerized experiment at two primary schools.

A second research approach is aimed at understanding the cognitive strategies used by players in Math Garden. To this end, a cross-sectional sample is constructed based on responses (accuracy and response times) to items of a subset of children who played a certain game on a regular basis. In this approach, 'raw' data are analyzed with an extended latent variable model that can capture more detailed processes compared to the current measurement model of Math Garden. In Chapter 3, we study the rules that children use to solve items from the well-known balance-scale task. We provide a comparison of a rule-based model and an information-integration model on two different data sets: a more traditional paper-and-pencil data set and a data set collected with Math Garden. In Chapter 4 we investigate the strategies involved in multiplication using extended IRT models. More specifically, we test whether the abilities involved in fast responses are the same as those involved in slow responses. These studies elucidate strategies that children use in solving items and thereby also provide pointers for providing children with feedback as well as improvements of the Math Garden system.

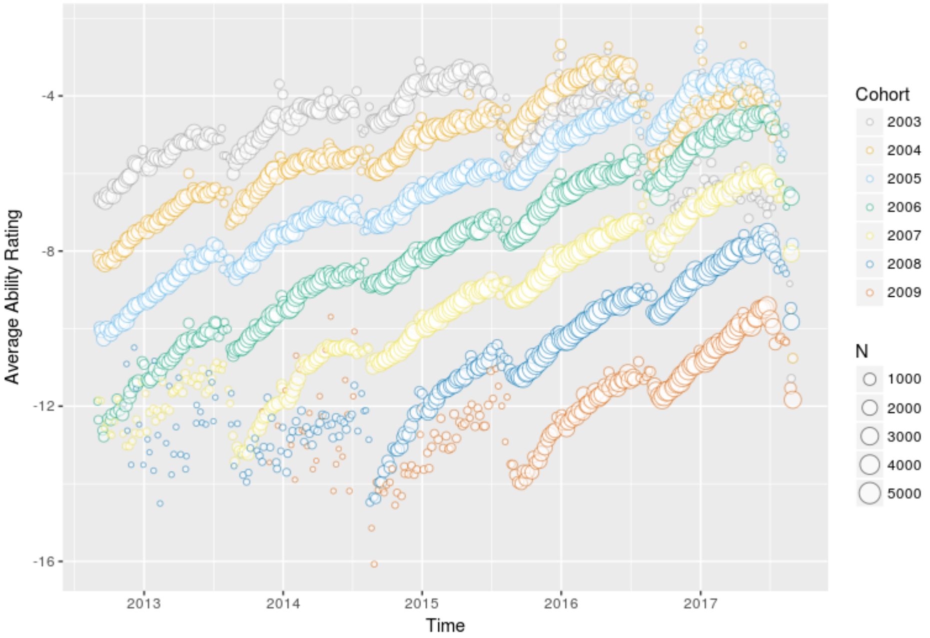


Figure 1.4: The development of the average ability on the addition game of children from different birth cohorts. Each coloured line represents a cohort of children born in a certain year, who gradually become older during data collection. The size of the dots represents the number of children who played within that week.

Both research approaches, and most of the published work so far, are based on cross-sectional data. However, the tracking system of Math Garden also allows us to investigate developmental processes using a longitudinal subset of the data. The third research approach concerns such longitudinal studies. To illustrate the longitudinal data, Figure 1.4[3] shows the average ability estimates of a single domain (addition) during four consecutive schools years of seven cohorts of children (each cohort consists of children born in the same year). This figure depicts data from 274,383 children who played the addition game, and indicates that on average the addition ability of children increases over time. However, some interesting additional patterns are observed: (1) during holidays large dips are found

[3] The figures are created with the *R* package *ggplot2* (Wickham, 2016)

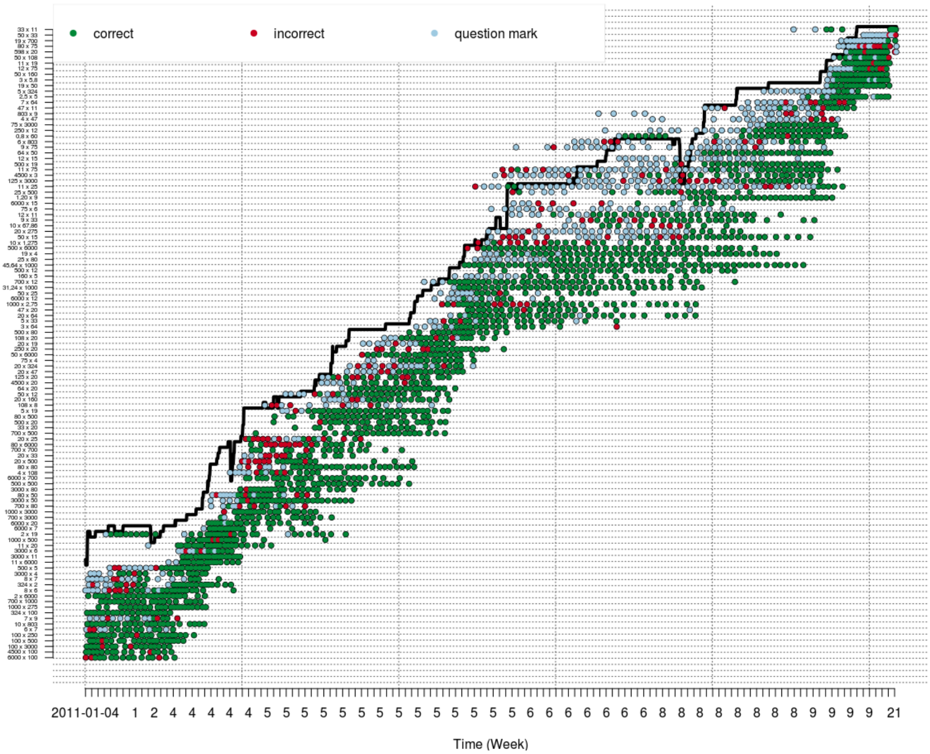


Figure 1.5: Responses to a set of multiplication items from a single child over time. The items on the y-axis are ordered on the estimated item difficulty (low is easy, high is hard). The black line depicts the child's estimated ability.

in the average rating and (2) the biggest progress is observed at a young age, and progress slowly diminishes while children grow older.

In the second part of this thesis we present several analyses of longitudinal subsets of the data. In Chapter 5, we investigate the links between the development of counting and the development of addition and between the development of multiplication and the development of division. To this end, we analyze the ability estimates with latent change scores models to compare both a mutualism (van der Maas et al., 2006a) and a g-factor (Jensen, 1998) account of development. In Chapter 6, we investigate the developmental processes of learning to touch type using the response times of keystrokes of a group of children who followed a typing course using the Type Garden (Typetuin; `www.typetuin.nl`).

In Chapter 7, we present different learning analytics aimed at providing descriptives of times-series of responses of children to single items. These data stem from a subgroup of children who visit Math Garden almost daily and have played for extended periods. The data provided by these children are rich in quantity and in dynamics, as Figure 1.5

illustrates. The responses of this child show that for some items he is not able to provide correct responses, but after a set of question-mark or incorrect responses he seems to learn the correct response and as a consequence his rating increases. In Chapter 7 we further describe these and other developmental patterns and collect different analytics to investigate the stability of the responses over time.

In sum, the present dissertation builds on earlier research with Math Garden, and extends the analysis of item and ability parameters to other domains and to links between domains. Most importantly, the dissertation shows analyses of longitudinal data, allowing the study of learning and cognitive development as it happens. The examples in this dissertation go beyond snapshots of what develops and also show the dynamics of developmental processes. This dissertation shows that Math Garden data, although not easy to analyze, provide a new window on cognitive development.

2 The Role of Pattern Recognition in Children’s Exact Enumeration of Small Numbers

This chapter is published as: Jansen, B. R., Hofman, A. D., Straatemeier, M., Bers, B. M., Raijmakers, M. E., & Maas, H. L. J. (2014). The role of pattern recognition in children’s exact enumeration of small numbers. *British Journal of Developmental Psychology, 32* (2), 178-194

Abstract

Enumeration can be accomplished by subitizing, counting, estimation, and combinations of these processes. We investigated whether the dissociation between subitizing and counting can be observed in 4- to 6-year-olds and studied whether the maximum number of elements that can be subitized changes with age. To detect a dissociation between subitizing and counting, it is tested whether task manipulations have different effects in the subitizing than in the counting range. Task manipulations concerned duration of presentation of elements (limited, unlimited) and configuration of elements (random, line, dice). In Study 1, forty-nine 4- and 5-year-olds were tested with a computerized enumeration task. Study 2 concerned data from 4-, 5-, and 6-year-olds, collected with Math Garden, a computer-adaptive application to practice math. Both task manipulations affected performance in the counting, but not the subitizing range, supporting the conclusion that children use two distinct enumeration processes in the two ranges. In all age groups, the maximum number of elements that could be subitized was three. The strong effect of configuration of elements suggests that subitizing might be based on a general ability of pattern recognition.

2.1 Introduction

Subitizing, the ability to rapidly and accurately enumerate a small set of elements (Kaufman, Lord, Reese, & Volkmann, 1949), is a component of number sense, which is essential for proficient math performance (Jordan, Kaplan, Locuniak, & Ramineni, 2007; Kroesbergen, van Luit, van Lieshout, van Loosbroek, & van de Rijt, 2009). Deficient subitizing is suggested to underlie lagging math skills of children with dyscalculia (Schleifer & Landerl, 2011). Despite extensive work on subitizing (starting with Kaufman et al., 1949), the question whether subitizing is a separate process, dissociable from estimation and counting is still actively investigated in various domains, such as neuropsychology (e.g., Dehaene & Cohen, 1994; Demeyere, Lestou, & Humphreys, 2010; Harvey, Klein, Petridou, & Dumoulin, 2013; Nan, Knösche, & Luo, 2006), psychonomics (Watson, Maylor, & Bruce, 2007), and developmental psychology (e.g., Schleifer & Landerl, 2011). In this article, we use a new method to test whether subitizing is a separate process. This new method also allows us to investigate the relation between subitizing and pattern recognition.

The standard method to differentiate between subitizing and counting involves a bilinear model, which statistically describes the shape of the function relating response times (RTs) to numerosity. It combines a regression function with a small positive, near-zero slope for the subitizing range; and a regression function with a larger, positive slope for the counting range. Reeve, Reynolds, Humberstone, and Butterworth (2012) additionally represented the dissociation between subitizing and counting by a change from a linear to an exponential function. The transition point is often set between 3 and 4 (e.g., Akin & Chase, 1978; Klahr & Siegler, 1978; Mandler & Shebo, 1982; van Oeffelen & Vos, 1982; Trick

& Pylyshyn, 1993) but is closer to 3 for children (Maylor, Watson, & Hartley, 2011; Svenson & Sjöberg, 1983)

The present study uses an alternative method, put forward by Trick (2008), to investigate the existence of a distinction in preschoolers' enumeration skills and a possible development of the subitizing range. The procedure circumvents two problems of the standard method. First, individual differences in the transition point can only be detected by estimating statistical models per participant (see Balakrishnanl & Ashby, 1992; Plaisier, Tiest, & Kappers, 2009, 2010). This requires many administrations of each numerosity to each participant, which is unfeasible with children. Also, simulations demonstrate that modern techniques (e.g, Muggeo, 2003) often fail in detecting transition points and differences in slope, when ranges of numerosities are small and slopes of the regression models differ in size, but not in sign (Julious, 2001; Muggeo, 2003). Second, good fit measures that allow model comparison are lacking. Trick argues that the hypothesis of a distinction is supported if manipulation of task conditions changes performance in the ranges differently. For example, using differently coloured elements speeds enumeration in the counting range, but not the subitizing range.

This study is the first to apply Trick's procedure with preschoolers. It is remarkable that developmental studies of enumeration are scarce (but see Benoit, Lehalle, & Jouen, 2004; Starkey & Cooper, 1995) because enumeration processes develop in preschoolers and not in adults. We present developmental data from both an experimental (Study 1) and a field study (Study 2). Study 1 includes a controlled manipulation of task conditions in a selected (small) sample, whereas Study 2 is less controlled, but has a very large sample, from a wide background, with many repetitions per participant.

Time limit was manipulated in Study 1. Performance in a condition with limited presentation duration was compared to performance in an unlimited time condition. Starkey and Cooper (1995) conclude that performance of children from 2 to 5 years of age is accurate in the subitizing but not the counting range in a limited time condition. However, only 2-year-olds' performance was compared in two time conditions. If subitizing and counting are the same process, manipulation of time limit is expected to affect performance comparably in the subitizing and the counting range. If subitizing is a separate process, variation in time limit is expected not to affect performance in the subitizing range because subitizing would ensure high performance. Limited time would deteriorate performance in the counting range as counting would be impossible and children probably resort to estimation.

Configuration of the elements to be enumerated was manipulated in both Study 1 and Study 2, using random, dice, and line configurations. The use of these various configurations is theoretically important because it allows studying the nature of subitizing, which is thought to be related to pattern recognition (Mandler & Shebo, 1982). Dice configurations form familiar patterns that may be processed holistically. Enumeration performance improves when presenting familiar configurations (e.g., four dots presented as vertices of a square) as compared to random configurations (e.g., Benoit et al., 2004).

As random, dice, and line configurations differ minimally for numbers in the subitizing range, it is expected that configuration would not affect performance in the subitizing range. Presenting elements in dice patterns is expected to benefit performance in the counting range. Line configurations also form a pattern (*line*), but it does not relate to a specific number of elements. Presenting elements in a line may either facilitate performance in the counting range, because elements are easily detected, or complicate performance because one may easily skip or recount elements (Towse & Hitch, 1996). The age variety in both studies, combined with the manipulation of configurations, allows for studying development of the subitizing range when presented with various configurations.

Manipulating the configuration of the elements is particularly important for investigating the relation between subitizing and pattern recognition. If familiarity of configurations has a larger effect on performance than the number of elements in the configuration, this is an important indication that pattern recognition is central in subitizing.

2.2 Study 1: An Experiment

Method

Participants

Participants came from two Dutch primary schools. Socio-economical status was high in one school (high SES-school) and low in the other (low SES-school). Parents either signed informed consent to allow their child's participation or were informed and could refuse participation. The Local Ethics committee approved of the procedures. A total of twenty-six 4-year-olds (58% girls) and thirty-seven 5-year-olds (51% girls) participated.

The final sample consisted of nineteen 4-year-olds (M = 4.59 years, SD = 0.24, 47% girls) and thirty 5-year-olds (M = 5.43 years, SD = 0.33, 53% girls), who completed at least 75% of the problems in both versions. Children missed problems due to inattentiveness. Inclusion was independent of age group, $\chi^2(1) = 0.57, p = .452$, gender, $\chi^2(1) = 0.77, p = .380$, and condition, $\chi^2(2) = 3.93, p = .140$.

Material

Task presentation was on 15-inch laptops, with a screen resolution of 1,024 x 768 pixels. The screen was viewed from a distance of about 40 cm. Red dots (RGB-values: 255, 0, 0; diameter: 1 cm; 1.4°) were presented in a screen-centred black-bordered white square (10 x 10 cm; 14° x 14°). The minimum interdot distance was 1.79°. The square covered an 8 x 8 cm (11° x 11°) matrix. Response buttons representing numbers 1-7 and a question mark were displayed at the bottom of the screen.

All subjects performed a task version with presentation duration limited to 250 ms and a task version with unlimited presentation duration. Order of versions varied randomly. Four consecutive screens appeared: (1) fixation cross (500 ms), (2) presentation of dots

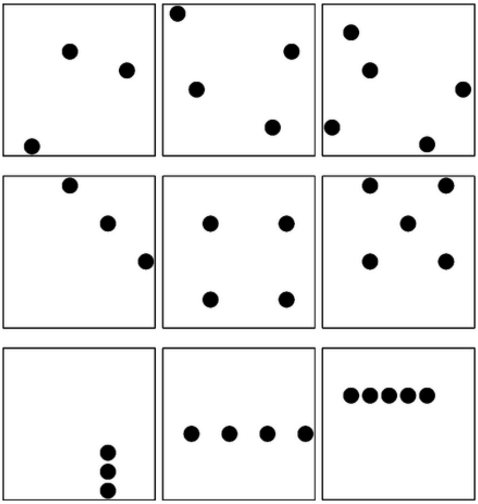


Figure 2.1: Study 1: Enumeration problems with three, four, and five elements (from left to right) in three types of configurations. From top to bottom: random, dice, and line.

(250 ms in limited time version; user-terminated in unlimited time version), (3) mask (solid blue square of 10 by 10 cm; RGB-values: 0, 100, 255; 1,000 ms), and (4) response buttons (experimenter-terminated).

Two general example problems and two version-specific examples were part of the task. Test problems were presented in four blocks of six problems each, yielding 24 problems. In each block, displays of 1-6 elements were presented randomly. Hence, each numerosity (1-6) was presented four times. Finally, 10 dots were presented in a line on screen.

Participants were randomly placed in the random, line, or dice condition. In the random condition, elements were spread in an unsystematic way. Four different displays were used for each number. In the line condition, evenly spaced elements were aligned on a horizontal or vertical line. Two different horizontal and two different vertical configurations were used for each number, varying distance between elements and length of complete display. In the dice condition, elements were presented in a dice pattern. The same dice configuration was used for each number, but absolute distance between elements varied, thereby varying the size of the entire display between displays of the same numerosity. Position of the display in the white square varied. Figure 2.1 shows examples of each configuration.

Procedure

Individual task administration took place in a quiet room at school. Instructions were printed on screen and read out loud by the experimenter. She told that 'grand dad collected berries for a bird near his house' and announced that berries would be shown after presentation of a small cross. At presentation of the elements, she asked: 'Can you tell me how many berries grand dad collected? Tell me how many and I will click on the answer'. Dots were presented until the child answered or indicated he/she did not know the answer. Correct responses varied from 1 to 6, whereas response buttons indicated numbers 1-7. The experimenter clicked on the button corresponding to the stated number or on the question mark in the absence of a response. The second example followed.

The experimenter introduced the two versions by encouraging the child to pay extra attention in the limited time condition and explaining that the berries would only disappear if the child pressed the space bar in the unlimited time condition.

In the final task, the experimenter asked 'Can you count these dots for me?' and pointed to the 10 dots on screen. Time was unlimited. The experimenter noted whether the child was able to count up to six or made any errors. We assured that counting skills were sufficient for counting to six so that performance variations could be attributed to experimental manipulations.

Results

Multivariate ANOVA with error rates (number of errors expressed as a proportion) in the unlimited and limited time condition as dependent variables and order and configuration as between-factors indicated no main effect of order, $F(11, 33) = 0.99, p = .476$, and no interaction effect between order and configuration, $F(22, 68) = 1.34, p = .181$. Hence, data from different orders were combined.

A mixed ANOVA with between-subjects factor configuration (three levels: random, dice, line) and age (two levels: 4 and 5 years) and within-subjects factors duration (two levels: Limited and unlimited) and numerosity (six levels: 1, 2, 3, 4, 5 or 6 elements) was performed on the error rates. In fact, displays with one element were similar in the three configurations.

The main effect of configuration was significant, $F(2, 43) = 5.47, p = .008$, $\eta_p^2 = .40$. Post-hoc analyses using Tukey's HSD indicated that error rates of random and line presentations were significantly higher than those of dice presentations. The main effect of age was not significant, $F(1, 43) = 2.78, p = .103$. The main effect of duration was significant, $F(1, 43) = 75.79, p < .001$, $\eta_p^2 = .64$, with lower error rates in the unlimited than the limited time condition. Finally, the main effect of numerosity was significant, $F(5, 215) = 56.19, p < .001$, $\eta_p^2 = .57$. Post-hoc analyses demonstrated that error rates increased with every additional element, except for the increase from 1 to 2 elements and the increase from 5 to 6 elements.

However, main effects were qualified by interactions between configuration and duration, $F(2, 43) = 7.84, p = .001$, $\eta_p^2 = .27$, configuration and numerosity, $F(10, 215) = 5.27, p < .001$, $\eta_p^2 = .20$, and duration and numerosity, $F(5, 215) = 13.56, p < .001$, $\eta_p^2 = .24$, and three-way interactions between configuration, duration, and numerosity, $F(10, 215) = 2.54, p = .006$, $\eta_p^2 = .11$, and between configuration, age, and numerosity, $F(10, 215) = 2.07, p = .028$, $\eta_p^2 = .09$. Remaining two-, three-, and four-way interactions were not significant.

ANOVAs with factor configuration were performed by numerosity and by duration to further investigate the interaction between configuration, duration, and numerosity. Alpha level was divided by the number of ANOVAs performed for each duration condition (6) and adjusted to .008. In the limited time condition, error rates did not differ between configurations for one, two, and three elements, $F(2, 46) = 0.01, p = .990$; $F(2, 46) = 1.28, p = .289$; $F(2, 46) = 0.79, p = .460$, for 1-3 elements, respectively. The error rates for four elements just did not differ between configurations, due to the corrected alpha level, $F(2, 46) = 4.13, p = .022$. Error rates did differ between configurations for five elements, $F(2, 46) = 12.57, p < .001$, $\eta_p^2 = .55$, and six elements, $F(2, 46) = 7.98, p = .001$, $\eta_p^2 = .35$. Post-hoc analyses showed that error rates were higher for random and line patterns than for dice patterns and for both five and six elements. Error rates did not differ between configurations for any numerosity in the unlimited time condition, $F(2, 46) = 0$; $F(2, 46) = 1.13, p = .332$; $F(2, 46) = 1.75, p = .185$; $F(2, 46) = 0.47, p = .630$; $F(2, 46) = 4.26, p = .020$; $F(2, 46) = 0.25, p = .782$, for 1-6 elements, respectively.[1] Note that the configuration effect just did not reach significance for enumerating five elements in the unlimited time condition, due to the corrected alpha level.

The three-way interaction was also investigated by testing the duration effect, by numerosity, for random configurations only, performing paired-samples t-tests. Alpha level was adjusted to .008. Error rates of enumerating one and two elements did not significantly differ between the two duration conditions, $t(13) = 1.47, p = .165$ for one element; $t(13) = 1.00, p = .336$ for two elements. The duration effect for enumerating three elements just did not reach significance, due to the corrected alpha level, $t(13) = 2.69, p = .019$. In contrast, error rates for enumerating four, five, and six elements were higher in the limited compared with the unlimited time condition, $t(13) = 5.70, p < .001, r = .85$; $t(13) = 4.94, p < .001, r = .81$; $t(13) = 5.95, p < .001, r = .86$ for four, five, and six elements.[2]

Repeated-measures ANOVAs with age as between factor and numerosity as within

[1] Chi-square tests were performed as well because distributions of error rates were possibly not normal. A chi-square test, testing dependence of configuration and error rates was performed by number, by duration. Alpha level was divided by the number of chi-square tests per duration condition (6) and adjusted to .008. In the limited time condition, configuration and error rates were independent for numerosities one, two, three, and four, $\chi^2(2) = 0.020, p = .990$; $\chi^2(6) = 5.22, p = .516$; $\chi^2(6) = 11.15, p = .004$; $\chi^2(8) = 17.86, p = .022$, for 1-4 elements, respectively. Configuration and error rates were clearly dependent for numerosities five, $\chi^2(8) = 23.17, p = .003$, Cramer's V = .49, and six, $\chi^2(8) = 20.80, p = .008$, Cramer's V = .46: Error rates were lower than expected for dice configurations and higher than expected for random and line configurations. In the unlimited time condition, configuration and error rates were independent, $\chi^2(4) = 2.56, p = .634$; $\chi^2(4) = 3.92, p = .417$; $\chi^2(6) = 8.37, p = .212$; $\chi^2(8) = 12.05, p = .149$; $\chi^2(8) = 5.66, p = .685$ for 2-6 elements, respectively. No errors were made on problems with one element, in the unlimited time condition. Hence, results of the chi-square tests replicated results of the ANOVAs.

[2] Chi-square tests, testing the dependence of error rates and duration, were performed by number, for random configurations. Alpha level was adjusted to .008. Duration and error rates were independent for one, two, and three elements, $\chi^2(4) = 2.15, p = .71$; $\chi^2(4) = 1.04, p = .90$; $\chi^2(4) = 6.09, p = .19$. Duration and error rate were dependent for four, $\chi^2(4) = 15.43, p < .01$, Cramer's V = .74, five, $\chi^2(4) = 13.16, p = .01$, Cramer's V = .69, and six elements, $\chi^2(4) = 18.24, p < .01$, Cramer's V = .81. Lower error rates on the unlimited, compared with the limited time task, caused the dependence. Results of the chi-square tests replicated results of the ANOVAs.

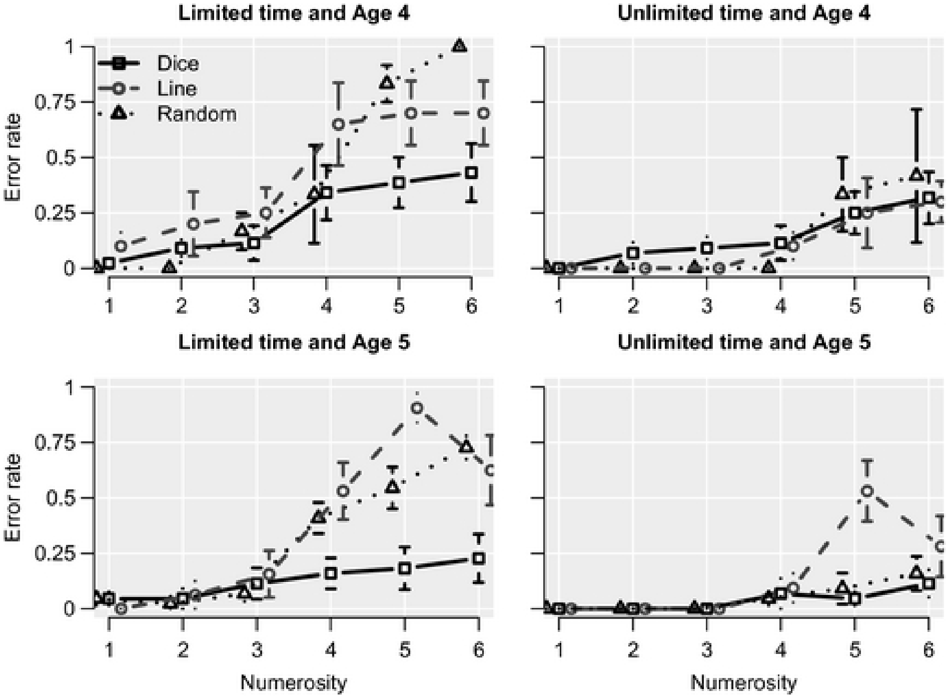


Figure 2.2: Study 1: Average error rates by numerosity, configuration, time limit, and age group. Error bars denote standard errors.

factor were conducted per configuration to further investigate the interaction between configuration, age, and numerosity. The interaction between age and numerosity was significant in random configurations, $F(5, 60) = 3.63, p = .006$, $\eta_p^2 = .23$, but not in dice, $F(5, 100) = 1.67, p = .149$, and line configurations, $F(5, 55) = 1.31, p = .131$. Figure 2.2 shows that 4-year-olds had higher error rates than 5-year-olds in random configurations, on problems with five and six elements.

Conclusion

Variations in configuration did not affect error rates when children enumerated 1-4 elements. Error rates were low for all configurations, in both the limited and the unlimited time condition. From five elements, however, dice presentations of elements were significantly easier compared with random or line presentations but only in the limited time condition. Time limits affected error rates from four elements. Hence, task manipulations affected performance in the counting range, but not the subitizing range. Children probably estimated the number of elements in the counting range, when in the limited time condition, because counting was impossible. Four-year-olds demonstrated a more steep increase in error rate than 5-year-olds from five elements, when enumerating problems

Table 2.1: Distribution of age groups, gender, and average number of attempted problems in selected set in enumeration game of Math Garden

Age	N (% of males)	Average number of made items (SD)
4	1,285 (52.8)	93.1 (106.9)
5	3,364 (49.6)	69.2 (99.2)
6	9,778 (50.7)	26.7 (54.1)
Total	14,427 (50.6)	52.3 (91.9)

Note. There were 12,302 unique players with at least one response on the selected problems. A total of 2,125 children played at least one particular problem in two different age groups.

with a random display. In conclusion, these results support the hypothesis that distinct processes are used for enumeration of elements in the two ranges (Trick, 2008), but not when elements are presented in dice patterns.

2.3 Study 2: Math Garden

Method

Participants

Data were collected between September 2010 and March 2013 with the *enumeration game*, part of project Math Garden (Klinkenberg et al., 2011). Math Garden is a Web-based computer-adaptive practice and monitoring system, used at school and at home by over 60,000 children. Children practice math skills by playing math games, linked to plants in a personal garden. Playing a game makes the plant grow. Here, we describe only those aspects that are essential to understand the data analyses below. Participating schools gave permission to use data from their students for research purposes and accepted responsibility to inform parents accordingly. Parents of private individuals (a minority of the sample) electronically signed for permission for use of their data in scientific research. Four-, 5-, and 6-year-olds were selected. Table 2.3 shows distribution of age and gender of participants who attempted at least one of the problems that were analysed. Both problem difficulties and RTs per problem were subject to analysis. Estimation of problem difficulties is explained in the next section.

Material and procedure

Fifteen enumeration problems were presented sequentially in a session, see Figure 2.3 for an example problem. For each problem, two consecutive screens appeared. The first screen showed the elements to be enumerated, a clickable keyboard with numbers 1-10, a question mark in case a participant did not know the answer, a coin bag, a row of 20 coins, and a green bar indicating game progress. Presentation duration was user-terminated, with a maximum of 20 s. A coin disappeared with each expiring second. Users clicked a response, which started the presentation of the second screen (1,000 ms in case of a correct

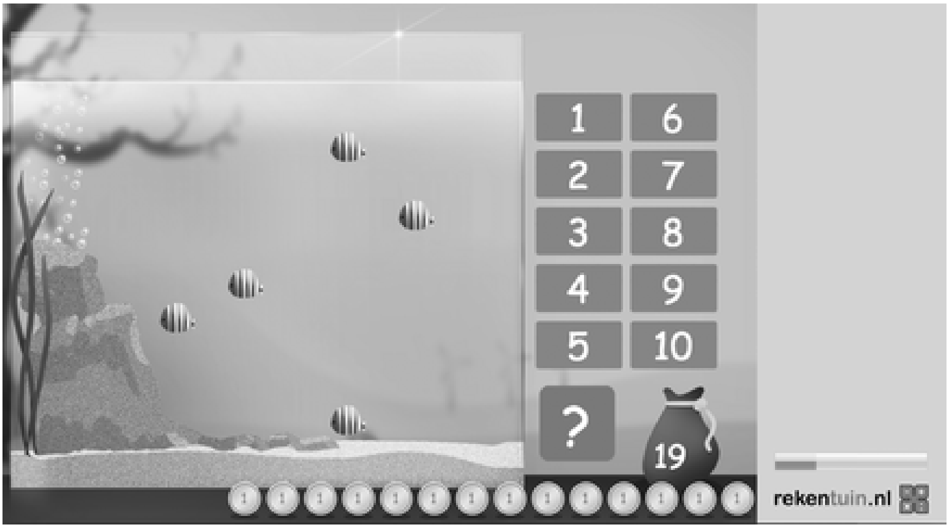


Figure 2.3: Study 2: Example of an enumeration problem in Math Garden. The example shows a random display of five elements.

response; 3,000 ms in case of an error), showing the correct response in green font and, if applicable, the false response in red font. After a correct (incorrect) response, the total in the coin bag was increased (decreased) by the number of remaining coins, with the notation that the total could not become negative. Hence, children were rewarded for fast accurate answers but penalized for fast inaccurate answers (Klinkenberg et al., 2011; Maris & van der Maas, 2012). Coins could be spent on prizes in a virtual trophy cabinet. If the question mark was clicked, no coins were won or lost and the correct response turned green.

Selection of each enumeration problem resulted from a match between problem difficulties and participants' enumeration skills. Both were estimated simultaneously using a computer-adaptive method (Klinkenberg et al., 2011), based on the Elo algorithm (Elo, 1978).[3] After a correct response, the estimated problem difficulty lowered (dependent on

[3] Problem difficulties and participants' skills are estimated simultaneously and updated continuously after a first choice of starting values. Starting values are based on problem size for problem difficulties and on age for participants' skills. Updating of a problem's difficulty and a participant's skills happens after the participant has solved a problem, according to:

$$\theta_{new|p} = \theta_{old|p} + K_p * (S - E(S))$$

$$\beta_{new|i} = \beta_{old|i} - K_i * (S - E(S)),$$

where θ_p is the skill estimate of participant p; β_i is the difficulty estimate of problem i; S and $E(S)$ are the score and expected score of person p on problem i. K is a function of the problem difficulty uncertainty U of the participant and the problem:

$$K_{new|p} = K_{old|p}(1 + K_{+}U_j - K_{-|i}U_i)$$

Table 2.2: Number of problems in selected set in enumeration game of Math Garden, by numerosity, by display

Numerosity	Random	Dice	Line
1	-	4^A	-
2	4	4	7
3	4	4	5
4	4	6	10
5	4	5	5
6	4	5	10

Note. A = Problems with one element could not be assigned exclusively to one of the configurations.

RT and estimated participant's skills) and the estimated skill increased (dependent on RT and estimated problem difficulty). The reverse happened after an error. Selection of the next problem depended on adjusted estimations of skill and problem difficulties. Over time, problem difficulties converged to a stable level. The next problem could be of different numerosity and/or different configuration but was chosen such that the average expected probability of a correct answer equalled .75. Hence, order of problems differed across participants.

Here, we focused on problems with 1-6 elements, in random, dice, or line configuration. Because of Math Gardens' adaptive nature, we did not collect data of children who were able to count larger numbers quickly and accurately because they were presented with more difficult items. Table 2.3 shows the number of available problems, by numerosity.

Math Garden's computer-adaptive method is based on estimation of problem difficulties and individual's skills on one and the same scale, across age groups. However, for Study 2, problem difficulties were recalculated for 4-, 5-, and 6-year-olds separately, using children's logged records of problems attempted, their answer, and RT on each problem. The estimation procedure was rerun by age group and resulted in a difficulty and an average RT for each problem, for each age group. The Euclidean distance, which is the straight-line distance between two elements, based on the grid used in Math Garden, was averaged across all elements, for each display and included as a covariate in subsequent analyses.

$$K_{new|i} = K_{old|i}(1 + K_{+}U_i - K_{-|i}U_j),$$

where $K = 0.0075$ is the default value when there is no uncertainty; $K_{+} = 4$ and $K_i = 0.5$ are the weights for the estimate uncertainty for participant p and problem i. U is uncertainty, which depends on both recency (the more recent, the lower the uncertainty) and frequency (the higher the frequency, the lower the uncertainty) of playing. Klinkenberg et al. (2011) assume that uncertainty reduces to 0 after 40 administrations, but increases to the maximum of 1 after 30 days of not playing:

$$U_{new} = U_{old} - \frac{1}{40} + \frac{1}{40}D.$$

Where D refers to the number of days without playing. A more elaborate description of the procedure of estimating problem difficulties and participants' skills is given in Maris and van der Maas (2012) and Klinkenberg et al. (2011).

Results

Problem difficulties

Although the number of problems may appear low, we contend that problem difficulties reflect participants' varying performance in enumeration, dependent on numerosity, and configuration because problem difficulties were based on responses from a large sample (see Table 2.3), had converged to stable values, and had small standard errors. Also, split-half reliability of problem difficulties, based on a split of the problems into two groups, with numerosity and configuration equally divided, was high, $r = .98, p < .001$.

Left panels in Figure 2.4 graph problem difficulties against numerosity for random, dice, and line configurations, by age group. Problems with one element are graphed in Figure 2.4 but excluded from analyses because they could not be assigned exclusively to one configuration. A univariate ANOVA with factors configuration (random, dice, and linear), age group (4-6 years), and numerosity (2-6), covariate average Euclidean distance, and dependent variable problem difficulties was performed. Average Euclidean distance did not affect problem difficulty, $F(1, 197) = 0.28, p = .60$. The main effect of configuration was significant, $F(2, 197) = 72.11, p < .001, \eta_p^2 = .50$. Post-hoc analyses indicated that problem difficulties were significantly lower for dice compared with random and line configurations ($p < .001$). The main effect of age was not significant, $F(2, 197) = 0.11, p = .89$. This is a consequence of the recalculation method in which the average problem difficulty was set to 0 for each age group. The main effect of numerosity was significant, $F(4, 197) = 248.73, p < .001, \eta_p^2 = .25$. Post-hoc analyses indicated that each additional element increased problem difficulty significantly ($p < .05$ for comparisons between displays of two subsequent numbers).

Main effects were qualified by the interaction between configuration and numerosity, $F(8, 197) = 7.66, p < .001, \eta_p^2 = .12$. Post-hoc analyses indicated that the configuration effect was not significant for two and three elements, $p > .05$. For four, five, and six elements, problems with dice displays were easier than problems with random and line displays, $p < .001$, whereas problem difficulties of the latter did not differ from each other for all numerosities, $p > .05$. The interaction between age group and numerosity was significant as well, $F(8, 197) = 7.76, p < .001, \eta_p^2 = .13$. In 4-year-olds, problems with two elements were significantly easier than problems with three elements, which were significantly easier than problems with four elements, $p < .001$. Difficulties of problems with four, five, and six elements were equally difficult in 4-year-olds, $p > .05$. In 5- and 6-year-olds, problems with two and three elements were equally difficult, $p > .05$, whereas problem difficulty increased significantly from 3 to 4 and from 4 to 5 elements. The interaction between configuration and age group was not significant, $F(4, 197) = 2.11, p = .081$. The three-way interaction was also not significant, $F(16, 197) = 1.61, p = .069$.

Summarized, dice displays, as compared to random and line displays, only lowered problem difficulties in the counting range. Moreover, 5- and 6-year-olds, but not 4-year-olds, showed the typical pattern of comparable difficulties for problems with two and three

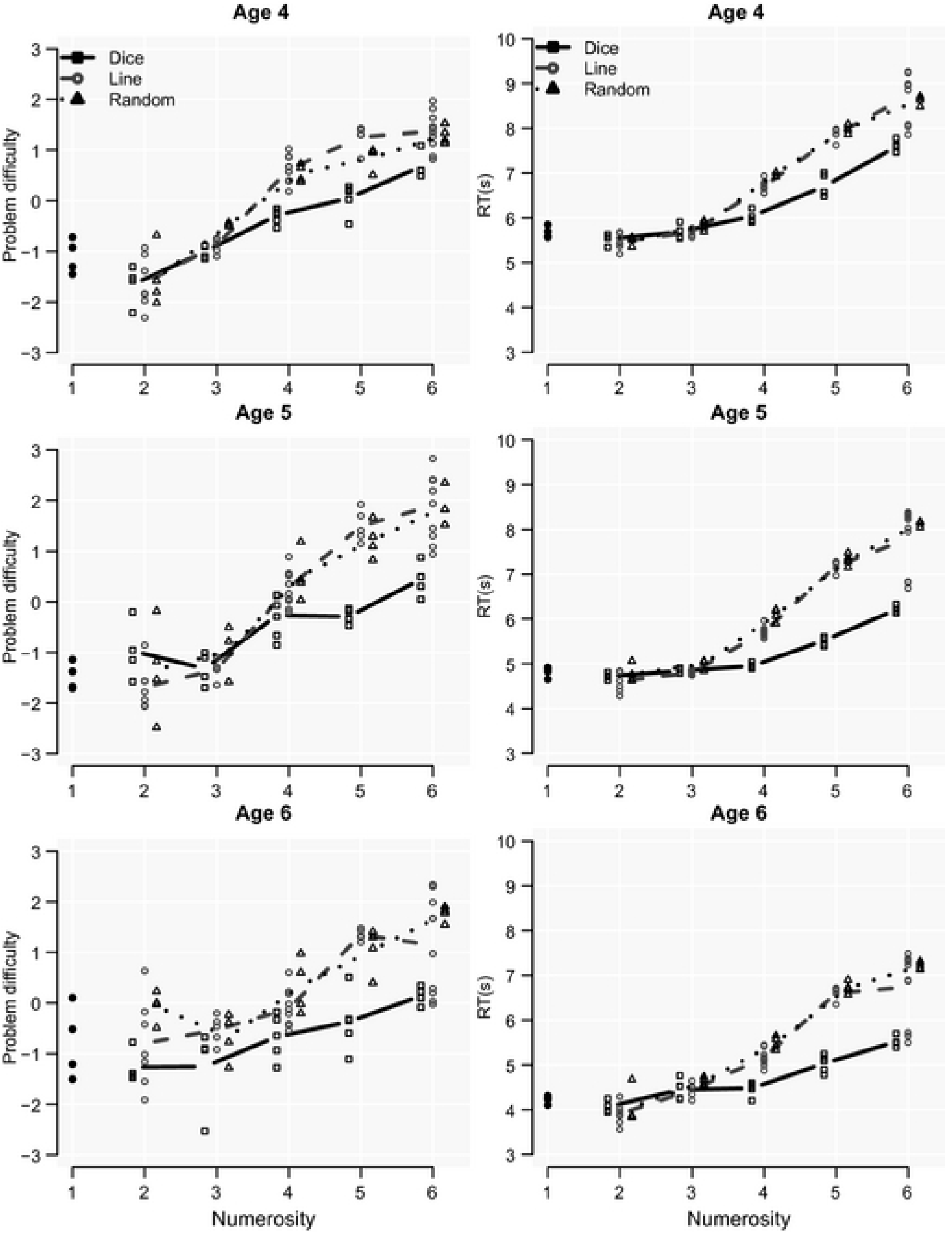


Figure 2.4: Study 2: Problem difficulties and average response times (RTs) in Math Garden by numerosity, configuration, and age group. Left panels show problem difficulties, and right panels show average RTs.

elements and increasing difficulty with each additional element, up to six.

Problem difficulties are the result of a combination of RTs and accuracy. To align with previous studies, an RT analysis is reported next. Error rates are not reported because the adaptive algorithm was designed to keep error rate at a constant value of .25 for as much as possible, for all individuals. Response times

Response times

The right panels of Figure 2.4 show RTs averaged by configuration, numerosity, and age group. A univariate ANOVA with factors configuration, age, and numerosity, and covariate average Euclidean distance was performed with RTs as dependent variable. Average Euclidean distance did not significantly influence RTs, $F(1, 197) = 0.02, p = .892$. The main effect of configuration was significant, $F(2, 197) = 195.63, p < .001$, $\eta_p^2 = .50$: RTs were longer for problems with random as compared to line displays, which were longer than RTs of problems with dice displays, all $p < .001$. The significant main effect of age, $F(2, 197) = 576.36, p < .001$, $\eta_p^2 = .50$ indicated that RTs decreased with age, $p < .001$. Finally, the significant main effect of numerosity, $F(4, 197) = 764.13, p < .001$, $\eta_p^2 = .25$, showed that RT increased with each additional element, all $p < .001$.

Main effects were qualified by significant interactions between configuration and numerosity, $F(8, 197) = 30.21, p < .001$, $\eta_p^2 = .13$, and between age and numerosity, $F(8, 197) = 3.50, p < .001$, $\eta_p^2 = .13$. The interaction between age and configuration was not significant, $F(4, 197) = 2.18, p = .072$, just as the three-way interaction, $F(16, 197) = 0.67, p = .826$. Post-hoc analyses indicated that the configuration effect was not significant for two and three elements, $p \geq 968$, but RTs were lower for problems with dice as compared to random and line displays for four, five, and six elements, $p < .001$. RTs for line and random display problems did not differ significantly from each other, $p \geq .056$. Post-hoc analyses also indicated that in 4- and 5-year-olds, RTs for enumerating two and three elements were equal, $p \geq .708$, but RT increased with each additional element, $p < .001$. In 6-year-olds, however, RT increased with each additional element, $p \leq .002$.

Conclusion

In general, problem difficulties and RTs increased with increasing number of elements, decreased with increasing age, and were lower for problems with dice as opposed to random or line displays, but effects of configuration, age, and numerosity interacted. The configuration effect only occurred when the number of elements was four or higher. In 4-year-olds, problem difficulty increased with each additional element, whereas RTs were equal in the subitizing range, but increased with each additional element in the counting range. In 5-year-olds, both problem difficulties and RTs were equal in the subitizing range and increased with each additional element in the counting range. In 6-year-olds, problem difficulties were equal in the subitizing range and increased with each additional element in the counting range, whereas RTs increased with each additional element in both ranges. Taken together, the results of the analyses of problem difficulties and RTs converged and suggest that manipulating the configuration of elements affected performance for problems in the counting range only and that the maximum number of elements that children could subitize was three, in all three age groups.

2.4 Discussion

In a long-standing discussion, it is debated whether humans use a single or two different processes to precisely enumerate small versus large numbers, referred to as subitizing and counting (Mandler & Shebo, 1982). The ranges in which the processes are used are referred to as the subitizing and the counting range. Here, we follow the argument that the claim of the existence of two distinct processes is supported, when task manipulations have different effects in the subitizing compared with the counting range (Trick, 2008).

In Study 1, a sample of Dutch 4- and 5-year-olds enumerated visually presented dots. Dots were arranged randomly, in a line or in a familiar pattern (dice). Presentation duration was limited or unlimited. Study 2 included an analysis of problem difficulties and RTs, obtained with a computer-adaptive math program (Klinkenberg et al., 2011). Configuration was manipulated as in Study 1. In both studies, task manipulations did not affect performance in the subitizing range, but did affect performance in the counting range. This consistent finding in these markedly different studies supports the claim that subitizing and counting are distinct processes (Mandler & Shebo, 1982; Schleifer & Landerl, 2011; Trick, 2008).

The findings on development in subitizing range diverged. In Study 1, 5-year-olds, but not 4-year-olds, show a clear difference between performance in the subitizing as opposed to the counting range. In Study 2, problem difficulties and RTs complemented each other and showed that 4-, 5-, and 6-year-olds were able to enumerate up to three elements fast and accurately and probably resorted to counting or estimation for larger numerosities.

Results on enumeration of elements in line configurations show that performance on problems with line and random configurations was comparable. If anything, problems with line configurations were easier than problems with random configurations. Line configurations might facilitate counting, as it is easy to move from one to the next element and to remember which elements are already counted. Balakrishnanl and Ashby (1992) solely used linear configurations and found that performance continuously decreased with increasing numerosity, also in the subitizing range. Differences between their study and the present studies demonstrate the necessity of varying the configuration of elements.

Both studies show that arranging elements in dice configurations facilitates enumeration of large numbers of elements compared with random configurations. Performance in the subitizing and counting range was similar when elements were presented in dice configurations (see Mandler & Shebo, 1982, who made a similar observation for adults). These results suggest that pattern recognition can help the enumeration of large numbers. Children with developmental dyscalculia show a deficit in both subitizing and the fast enumeration of elements in the counting range, when presented in familiar patterns. Difficulties in pattern recognition may relate to these deficits (Ashkenazi, Mark-Zigdon, & Henik, 2013). Visuo-spatial working memory (VWM) is a prerequisite for pattern recognition (Ashkenazi et al., 2013). Hence, the hypothesis that subitizing is based on pattern recognition matches the observation that VWM capacity correlates significantly with subitizing capacity (Piazza, Fumarola, Chinello, & Melcher, 2011). Although subitizing may also be the result of the application of a limited number of spatial indexes (Trick & Pylyshyn, 1994), Vetter, Butterworth, and Bahrami (2008) show that it cannot be a pre-attentive process.

In Study 1, performance in the counting range increased when given unlimited observation time as compared to limited time. However, performance for dice patterns was already at ceiling in the limited time condition. Note that the number of possible patterns for numerosities in the subitizing range is limited and that many possible random patterns map the familiar dice patterns. Hence, associating patterns with number words is relatively easy in the subitizing range. The number of possible patterns grows exponentially in the counting range (Benoit et al., 2004), complicating the association between patterns and number words. As a consequence, children in the studied age range may have already learned these associations in the subitizing range, but not yet in the counting range. This sketch of development implies that counting is a prerequisite for subitizing (R. Gelman & Gallistel, 1978).

Mandler and Shebo (1982) and Ashkenazi et al. (2013) already proposed that subitizing is based on acquisition of associations between patterns and number words. The finding that subitizing improves with age (Maylor et al., 2011; Reeve et al., 2012; Starkey & Cooper, 1995; Trick, Enns, & Brodeur, 1996) supports this hypothesis. However, the hypothesis conflicts with results from infant studies (Carey, 2004; Desoete, Ceulemans, Roeyers, & Huylebroeck, 2009; Feigenson, Dehaene, & Spelke, 2004) and studies on number discrimination among people in remote cultures (Dehaene, 1987). Studies including a broad age range, applying a shared paradigm across age groups, may contribute to the discussion on the origins of subitizing.

The current studies are not without limitations. In Study 1, only four repetitions of each number were used in each condition, sample size was small, and distance between elements did not vary randomly across numerosities. Fortunately, Study 2 showed that distance between elements did not influence problem difficulties. In Study 2, administration was unsupervised, and order of presentation of numerosities and displays was uncontrolled. However, the large sample size probably averages out effects of environment and order.

Summarized, the results suggest that children use distinct processes for the enumeration of small and large sets of elements when presented in random patterns. The results concerning enumeration of elements in dice patterns suggest that performance of preschoolers can be explained by three processes: Counting (for large numbers, when given sufficient time), estimation (for large numbers, when time is limited), and subitizing. The latter might be based on pattern recognition as performance for dice patterns is comparable for small and large numbers. If subitizing indeed involves pattern recognition, extension of subitizing to larger numbers may be possible (as the recognition that two dice patterns of six represent 12 elements), comparable to the recognition of complex chess patterns by advanced chess players (De Groot, 1978). Training of pattern recognition (B. Fischer, Köngeter, & Hartnegg, 2008) and presenting elements in fixed patterns may ease number recognition and encourage insight into simple addition and subtraction. After all, enumeration is an important requisite for later math skills (Jordan et al., 2007; Kroesbergen et al., 2009).

3 The Balance-Scale Task Revisited: A Comparison of Statistical Models for Rule-Based and Information-Integration Theories of Proportional Reasoning

This chapter is published as: Hofman, A. D., Visser, I., Jansen B. R. J. & Van der Maas, H. L. J., (2014). The Balance-Scale Task Revisited: A Comparison of Statistical Models for Rule-Based and Information-Integration Theories of Proportional Reasoning. *PloS one 10* (10), e0136449

Abstract

We propose and test three statistical models for the analysis of children's responses to the balance scale task, a seminal task to study proportional reasoning. We use a latent class modelling approach to formulate a rule-based latent class model (RB LCM) following from a rule-based perspective on proportional reasoning and a new statistical model, the Weighted Sum Model, following from an information-integration approach. Moreover, a hybrid LCM using item covariates is proposed, combining aspects of both a rule-based and information-integration perspective.

These models are applied to two different data sets, a standard paper-and-pencil test data set (N = 779), and a data set collected within an online learning environment that included direct feedback, time-pressure, and a reward system (N = 808). For the paper-and-pencil data set the RB LCM resulted in the best fit, whereas for the online data set the hybrid LCM provided the best fit. The standard paper-and-pencil data set yielded more evidence for distinct solution rules than the online data set in which quantitative item characteristics are more prominent in determining responses. These results shed new light on the discussion on sequential rule-based and information-integration perspectives of cognitive development.

3.1 Introduction

Two types of cognitive processing are often considered, and fiercely debated, in theoretical discussions of cognitive development: sequential rule-based processes (RB) versus information-integration (InI) based processes. These two types of processing are also contrasted in other areas in (cognitive) psychology. For example, in the study of information-integration in category learning (Ashby & Maddox, 2011) and in the study of explicit and implicit learning (Shanks, 2010). Moreover, Pothos (2005) provides a discussion of the rules versus similarity distinction in cognition, and Kahneman (2011) provides an broad overview and examples of dual route models, explicit versus implicit, in psychology.

In the study of cognitive development the balance-scale task (Inhelder & Piaget, 1958) is the primary battlefield for this debate and it is the focus of this article. Recent publications (Quinlan, van der Maas, Jansen, Booij, & Rendell, 2007; Schapiro & McClelland, 2009; Shultz & Takane, 2007) attest that this debate is still very much alive. Proponent of the RB perspective, initiated by Klahr and Siegler (1978) and Siegler (1976), state that the cognitive process consists in the sequential comparison of different features of the stimulus. Cognitive development is described as discontinuous jumps between stages characterized by qualitatively different rules, that correspond to the consideration of different stimulus features in different combinations. With age, children acquire new insights that result in the use of more complex rules (Siegler, 1996; Jansen & van der Maas, 2002). From the InI perspective, cognitive processing is based on integrating different features of the stimulus before making a decision (Wilkening & Anderson, 1982, 1991). Knowledge in this perspective is considered graded and implicit in nature, and development is viewed as due to changes in the implicit weights of each dimension (McClelland, 1995; Schapiro & McClelland, 2009).

The cognitive processes used by children on the balance-scale task are especially interesting because their development spans a long period of time. Young children demonstrate interesting types of (erroneous) thinking, and many adults fail to use proportional reasoning to answer balance scale problems correctly. Also, over age, a mixture of developmental patterns seems to occur, ranging from sudden transitions to continuous change see for example Jansen and van der Maas (2002).

Many researchers developed computational models to investigate learning and development on the balance-scale task. Computational models from different research traditions have been proposed: production-rule models (Klahr & Siegler, 1978), decision-tree models (Schmidt & Ling, 1996), connectionist models (McClelland, 1989, 1995; Schapiro & McClelland, 2009; Shultz, 2003; Shultz, Mareschal, & Schmidt, 1994) and ACT-R models (van Rijn, van Someren, & van der Maas, 2003). Although the current models all adopt some characteristics of both theoretical positions, there is still no consensus on the best cognitive processes underlying children's behavior in the balance-scale task (Dandurand & Shultz, 2013).

In our view, this lack of consensus is partly due to the lack of adequate statistical models for the analysis of empirical data. Computational models such as production rule models and connectionist models cannot easily be fitted to data, and the existing models within the RB framework cannot test hypotheses following from the InI perspective. The empirical status of process models differs form traditional descriptive models, and a direct evaluation of these models is difficult since their aims are different (Luce, 1995). In this paper we test empirical predictions that follow from both theoretical perspectives - discussed hereafter. Therefore we develop statistical models for the RB and InI perspective and a hybrid model that combines features of both theories. We apply these models to two different data sets, a paper-and-pencil data set (N = 779) and a data set collected within an online learning environment that includes direct feedback, time-pressure, and reward (N = 808).

The Balance-Scale Task: Two Different Perspectives

In the balance-scale task (Inhelder & Piaget, 1958), children have to predict the movement of a balance-scale (see Figure 3.1), on which the number of blocks on each peg, and the distance between the blocks and the fulcrum are varied. Depending on the number of blocks and the distance between the blocks and the fulcrum on each arm, the beam tilts to one side or remains in balance. Thus, to succeed on the balance-scale task, a child has to identify the relevant task dimensions (number-of-blocks and distance) and to understand their multiplicative relation (Jansen & van der Maas, 2002).

To measure proportional reasoning with the balance-scale task, Siegler (1976) classified items into six item types. There are three simple item types: balance (B) items with an equal number of blocks placed at equal distances from the fulcrum; weight (W) items with

Table 3.1: Siegler’s Rules on the Balance-Scale Task and the Expected Percentage (%) of a Correct Responses

Problem Type	Rule I	Rule II	Rule III	Rule III-Add	Rule IV
Weight-Distance (WD)	100	100	100	100	100
Balance (B)	100	100	100	100	100
Weight (W)	100	100	100	100	100
Distance (D)	0	100	100	100	100
Conflict-Balance-Addition	0	0	33	100	100
Conflict-Weight-Addition	100	100	33	100	100
Conflict-Distance-Addition	0	0	33	100	100
Conflict-Balance	0	0	33	0	100
Conflict-Weight	100	100	33	0	100
Conflict-Distance	0	0	33	0	100
Rule Description	Only weight	Distance when weight is equal	Distance and weight, guess when conflict	Distance and weight, addition when conflict	Distance and weight, product when conflict

Note. Weight = Number of blocks.

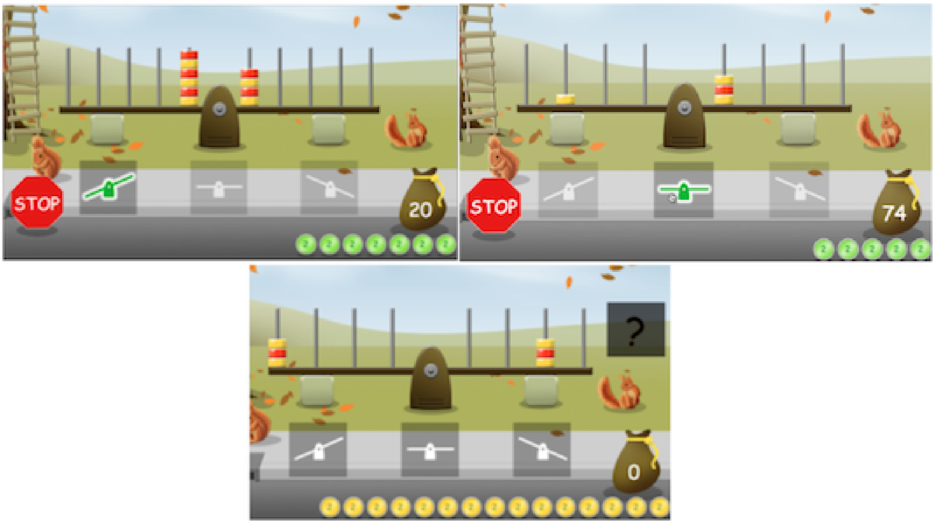


Figure 3.1: Three example items of the balance-scale task as programmed in the Math Garden (upper-left = Distance item; upper-right = Weight item, positive feedback; upper-right = Distance item; lower = Conflict-Balance-Addition item)

a different number of blocks placed at equal distances from the fulcrum, and distance (D) items with the same number of blocks placed at different distances from the fulcrum. We also include weight-distance (WD) items, in which the largest weight is positioned at the largest distance, such that a focus on either weight (i.e. number of blocks) or distance leads to a correct answer. Next to these simple items, there are three conflict item types in which the weight and distance cues conflict: conflict-weight (CW) items, in which the scale tips to the side with the largest weight; conflict distance (CD) items, were the scale tips to the side with the largest distance and conflict-balance (CB) items where the scale stays in balance.

Using these item types Siegler (1976, 1981) differentiated between a postulated series of rules that children might use to solve balance-scale items. A child using Rule I will only consider the number of blocks in the prediction of the movement and disregards the distances - the number of blocks is more dominant than the distance. A child using Rule II does include the distance dimension in the prediction, but only when the number of blocks on each side of the fulcrum is equal. A child using Rule III does know that both the number-of-blocks and the distance dimension are relevant but does not know how to integrate both dimensions. A child using this rule will guess or ’muddle through’ when both dimensions are in conflict. A child using Rule IV compares the torques on each side resulting in correct responses on all problems.

Some studies proposed alternative rules, the main example being the addition-rule (Rule III-ADD; Ferretti, Butterfield, Cahn, & Kerkman, 1985; Normandeau, Larivée, Roulin, & Longeot, 1989; Jansen & van der Maas, 1997, 2002). Children who use the addition-rule compare the sums of the number of blocks and the distance of each side of the fulcrum and predict that the side with the largest sum goes down. Detection of this rule is possible because some conflict items are solvable with the addition rule whereas others are not (see Table 3.1). In this study, we consider conflict items of the type conflict-balance-addition (CBA), conflict-weight-addition (CWA) and conflict-distance-addition (CDA), next to conflict-balance (CB), conflict-distance (CD) and conflict-weight (CW) items. The latter three cannot be solved with the addition rule, whereas the former can be.

In contrast to the RB perspective, according to the InI perspective children use a weighted integration of the number-of-blocks and distance between the blocks and the fulcrum, either based on a sum or a product for each side of the fulcrum and compare these integrations to select their response (Wilkening & Anderson, 1982). Either the number-of-blocks or the distance dimension is more dominant, resulting in a higher weight for one of the dimensions. In this perspective, differences between children are due to the differences in the weights that they apply to either dimension in integrating information. In the statistical extension of the connectionist models introduced in this paper, the weighted integration is only based on the sums and not the products.

Different empirical predictions: Individual Differences and Item Characteristics

The RB and InI perspectives make different predictions about children's behavior in the balance-scale task. Here we discuss the main differences. A first prediction concerns the characterization of individual differences between children. According to the RB perspective, children can be classified into subgroups or classes associated with qualitatively different rules. For example, Jansen and van der Maas (1997) found evidence in agreement with the RB model of Siegler (1976), using latent class models. However, according to the InI perspective, these seemingly qualitative individual differences are due to quantitative differences in integration weights.

A second distinctive prediction concerns responses to different items of the same type. According to the RB perspective, the response probability is solely dependent on the item type. Items of the same item type should have equal response probabilities. This assumption of item homogeneity applies to each rule. For instance, all conflict balance items should have equal response probabilities for all users of Rule I. In the InI perspective, differences in number of blocks and distances between items of the same item type influence the response probabilities. According to Ferretti and Butterfield (1986, 1992); Ferretti et al. (1985), children are more likely to provide correct answers when the difference between the product of the number of blocks and distance, on each side of the scale is larger. Three studies reanalyzed data of Ferretti and Butterfield (1986) and concluded that this was only the case for items with extreme product differences (Jansen & van der Maas, 1997, 2002; van Rijn et al., 2003). Therefore, supporters of the RB perspective have argued that item homogeneity holds.

Statistical Models: Measuring Rules vs Information Integration

As the RB and InI response mechanisms are latent (i.e., unobserved), a measurement model is required to test whether the observed patterns of responses correspond to expected responses following from the different mechanisms. The empirical detection of rules was first conducted by using rule-assessment-methodology (RAM; Siegler, 1976, 1978). RAM was designed to classify children to a set of a-priori defined rules, based on their observed responses instead of their verbal explanations of balance-scale answers. RAM is a two-step procedure. First, based on the set of a-priori defined rules the expected responses to the items are determined for all rules. Second, children are classified to one of the rules based on the best match between their observed responses and the expected responses following from each rule. In this classification some deviation between the observed and expected response pattern is allowed. The degree of deviation allowed depends on the item set. In the InI approach, a comparable rule-assessment method (Wilkening & Anderson, 1982) is used. For some specific choice of weights, expected response patterns are calculated and children are classified as using these particular values based on their observed response pattern.

Although RAM proved to be a valuable method for studying the cognitive processes of children on the balance-scale task, is has two important disadvantages. First, RAM is not based on a statistical model, and as such does not incorporate measurement error. Hence, RAM lacks a statistical test of the fit of the classification of children to rules. As a result, it is problematic to decide on the necessity of incorporating all the rules and to compare competing rule models statistically. Second, by using a priori defined rules one risks overlooking alternative rules (van der Maas & Straatemeier, 2008) and other response mechanisms. These limitations apply to some extent as well to the InI method of detecting integration rules used by (Wilkening & Anderson, 1982).

To overcome these problems latent class analyses (LCA; see McCutcheon (1987), for an introduction) were introduced in the balance-scale literature (Jansen & van der Maas, 1997, 2002; Boom, Hoijtink, & Kunnen, 2001). A latent class model (LCM) is a latent variable model, in which both the manifest (i.e., the item responses left, balance or right) and the latent (i.e., the rules) variables are categorical. Latent variable models are statistical measurement models, which allow for goodness of fit tests and statistical model comparison. It is best seen as a statistically advanced version of RAM. It is important to note that the rule model underlying RAM is in fact an instantiation of a restricted confirmatory LCM with fixed conditional probabilities (van der Maas & Straatemeier, 2008). Recently Dandurand and Shultz (2013) demonstrated in a simulation study that the response probabilities of small classes (N=20) are characterized by high standard errors. This lack of power due to small class probabilities is indeed problematic for parameter estimation in LCMs. Therefore the description and interpretation of small classes should be done with care. However, the simulation study also showed that the LCM correctly recovered the number of classes and the classification of subjects to classes, also for the small classes. To conclude, these difficulties do not outweigh the advantages of LCA over RAM (Quinlan et al., 2007; Shultz & Takane, 2007; van der Maas, Quinlan, & Jansen, 2007; Raijmakers, Jansen, & van der Maas, 2004).

In the next section we describe the RB model and introduce a statistical InI model and a hybrid model based on predictions from both perspectives.

Rule-Based Model In the LCM, both the latent variable and the responses are categorical. Participants are assigned to a latent class, associated with a distinct rule or strategy, based on their observed responses on the balance-scale items - left side down, balance or right side down. Equation 3.1 describes the probability of a response vector **r** in a LCM:

$$P(\mathbf{R} = \mathbf{r}) = \sum_{c=1}^{C} P(C = c) \prod_{i=1}^{I} P(R_i = r_i | C = c), \tag{3.1}$$

where r_i denotes the response to item i and c denotes the latent class. The LCM consists of two parts: the prior (or latent class) probabilities, $P(C = c)$, describing the estimated proportion of children in a given class c, and the conditional response probabilities, $P(R_i = r_i | C = c)$, describing the probabilities of a response to each item given a class. In our formulation, these response probabilities are estimated using a multinomial logit formulation (Bouwmeester, Sijtsma, & Vermunt, 2004). The left response is used as the reference category resulting in two odds-ratios: left versus balance, $log(p(L)/p(B))$, and left versus right, $log(p(L)/p(R))$. The model described in Equation 3.1, is referred to as the exploratory model since no constraints are imposed on the response probabilities between different items.

We also consider a second LCM, in which the response probabilities between items of the same type are not allowed to vary, following the item homogeneity assumption of the RB perspective. The response probabilities can be expressed using the following logit formulation:

$$P(R_i = r_i | C = c) = \frac{e^{\beta_{0rc}}}{1 + \sum_{r=1}^{R-1} e^{\beta_{0rc}}}. \tag{3.2}$$

The response probabilities of all items, of one item type, are modeled as a function of a general intercept β_{0rc} - per odds-ratio, per item type and per class. Hence, in this model, referred to as the item homogeneity model, the response probabilities are constrained to be equal over items of each item type and each latent class. Note that the item type index is missing in Equation 3.2 since the model is fitted seperately to data of each item type.

Information-Integration Model For the InI approach a statistical model is missing. Here, we propose a new measurement model, the Weighted-Sum Model (WSM). According to the InI perspective individuals differ in two respects: a) in the dominance for either the number-of-blocks or the distance dimension and b) in the preference of balance responses. Given these two sources of individual differences the following model for the weighted-addition rule (Wilkening & Anderson, 1982) is proposed:

$$\theta_p = \alpha_p \Delta w_i + (1 - \alpha_p)\Delta d_i, \tag{3.3}$$

$$\begin{gathered}\text{If } \theta_p < -C_p \text{ Then LEFT} \\ \text{If } \theta_p > C_p \text{ Then RIGHT, Else BALANCE,}\end{gathered}$$

where α_p expresses the persons dominance for either the number-of-blocks ($\alpha_p > .5$) or distance ($\alpha_p < .5$) dimension, and Δw_i and Δd_i are defined as respectively the difference between the number of blocks (weights) and distance on both sides. Based on θ_p and a personal threshold, C_p, the observed responses are derived. C_p serves as a boundary between responding either left or right ($|\theta| > C_p$) or balance ($|\theta| < C_p$). A high C_p implies a strong preference for the balance response. The parameters α_p and C_p are estimated per child, based on the likelihood-function of the model (see Appendix to Chapter 3 for a detailed description of the estimation procedure). Since this statistical model is estimated per child, no distributional assumptions about the model parameters are required. According to the InI theory, differences between children are gradual and the distributions of α_p and C_p are assumed to be unimodal. A bi- or multimodal distribution of these parameters provides support for a mixture distribution representing qualitative differences between children, thereby resulting in a hybrid WSM.

Hybrid Models Furthermore, to bridge the gap between the RB and InI perspective, we extend the item homogeneity LCM with item covariates (Huang & Bandeen-Roche, 2004) based on continuous item characteristics. This extension provides a formal measurement of the effect of quantitative item characteristics, such as the product-difference, on the response probabilities, combining the qualitative differences that follow from a RB perspective with quantitative item effects following from an InI perspective.

$$P(R_i = r_i | C = c) = \frac{e^{\beta_{0rc} + \beta_{1rc} \mathrm{x}_i}}{1 + \sum_{r=1}^{R-1} e^{\beta_{0rc} + \beta_{1rc} \mathrm{x}_i}}, \tag{3.4}$$

In this LCM, the item heterogeneity model, a slope parameter β_{1rc} is included allowing for differences in the response probabilities within items of the same item type based on some item characteristic x_i. We focus on the most often used characteristic, the product-difference (PD), the differences between the product of the number of weights and the distance on each side of the fulcrum. To conclude, we present three measurement models: a LCM following from the RB perspective, a WSM following from an InI perspective and a hybrid LCM that combines both RB and InI effects.

Table 3.2: Distribution of Age for the Paper-and-Pencil and Math Garden data set

age in years:	< 6	6-7.99	8-9.99	10-11.99	12-13.99	14-15.99	> 16
Paper-and-Pencil	1	63	148	171	146	147	93
Math Garden	15	209	281	186	41	14	0

3.2 Method

Participants

The paper-and-pencil version of the balance-scale task was administered to 805 children. Responses to the first block and responses from children that did not understand the task or with missing responses (N = 26; hereafter the paper of Jansen and van der Maas (2002) is referred to as JM) were discarded. On average children needed 10 minutes to complete the test (20 seconds per item). Further details on this data set can be found in JM.

The Math Garden data set consists of data of 808 children who completed at least five blocks during the data collection period (between 2011-06-10 and 2011-08-12). In the Math Garden children practiced either during school or outside school hours, resulting in large differences in both the number of items made and in the amount of time spent playing the balance-scale game. On average these five blocks were completed within 8 days (SD = 10.5, range = 0-54). The responses on items of the first block were discarded since children had to get acquainted to the task. Table 3.2 shows the distribution of age of both the paper-and-pencil and the Math Garden data set. Note that older children are somewhat underrepresented in the Math Garden data set compared to the paper-and-pencil data set.

Materials

Paper-and-Pencil The paper-and-pencil version of the balance-scale task consisted of five items of the types W, D, CW, CDA and CBA (see Appendix to Chapter 3 for the item characteristics). Before administration of the task, the experimenter explained that the pegs were placed at equal distances, that all the weights had the same weight, and showed that a pin prevented the scale from tipping. Subsequently, three example items were presented to familiarize the children with the format of the test.

Math Garden In the balance-scale game, children are asked to predict what would happen if the blocks under the balance are removed (see Figure 3.1). The three answer options are displayed below the item. The Math Garden game differs in three respects from the standard paper and pencil test. First, items are presented with a time-limit of twenty seconds. Second, children receive feedback on the accuracy of their response directly after responding. Third, children are rewarded for correct responses and are punished for incorrect responses. The time-limit/pressure is an inherent aspect of the feedback system where size of reward/punishment is positively related to speed (Maris & van der Maas, 2012). If a child has no clue of an answer he or she may press the question mark button. These task elements are designed to keep the task challenging, and enable learning through feedback (see Klinkenberg et al., 2011, for an extended description of the Math Garden system and its rationale).

The original item set consisted of 260 items, divided in twenty blocks of thirteen items of different types. Ten item types are presented in Table 3.1. The remaining item types were items with weights on multiple pegs on one or two arms of the scale. We analyze responses to the four D, CW, CDA and CBA items to increase comparability with the paper-and-pencil and the Math Garden data set (see Appendix to Chapter 3 for the characteristics). In both data sets, for all item types, except CBA items, the quantitative item characteristic of interest was the product-difference. For CBA items we use weight-difference as an alternative (for CBA items the product-difference is zero by definition since the weight- and distance-differences are the same). Although the items were not explicitly constructed to test a quantitative effect, they exhibit sufficient variation in this item characteristic. For both data sets the responses were recoded such that the correct response is the left response for D, CW and CDA items, and such that the largest amount of pegs resides on the left side of the fulcrum for CBA items.

Model Estimation and Comparison

Following the approach of JM, we applied LCA in two consecutive steps. First, the responses per item type were investigated. The number of latent classes was determined (investigating qualitative individual differences) with exploratory LCA (the exploratory model). Thereafter, parameter restrictions, formulated in the item heterogeneity model and the item homogeneity model, were sequentially tested. Second, building on the results of this fitting procedure per item type, response to multiple item types were analyzed with the hybrid LCM (item heterogeneity model; formulated in Equation 3.4). This approach reduces the sparse data problem in LCA when analyzing a large set of variables since it limits the number of estimated parameters compared to exploratory model. Hence the power to detect different classes increases. Third, this item heterogeneity model - the hybrid LCM - is compared with the item homogeneity model - the rule-based model.

For the LCM including all responses, we analyzed the posterior probabilities, $P(C = c|\mathbf{R} = \mathbf{r})$. These probabilities - based on the observed responses of a person and the estimated prior and conditional response probabilities - indicate the classification probabilities of a person to each class. The probabilities are related to the homogeneity of responses of subjects belonging to a certain class and the class separation (Collins & Lanza, 2010). A high (maximum is one) posterior probability implies that the observed response pattern of a subject is well described by the estimated response probabilities of a latent class. A value of one divided by the number of classes indicates that the observed responses pattern cannot be clearly assigned to any latent class. The average (and standard deviation) of the posterior probabilities over subjects assigned to each class is presented. A high mean indicates that subjects can be clearly assigned to this class compared to the other classes.

All RB and hybrid models were estimated with the depmixS4 package (Visser & Speekenbrink, 2010) in R (R Core Team, 2013). For stable model estimation we scaled the product-difference, per item type, such that the mean equals zero. Twenty replications were used with random starting values to prevent solutions based on local optima. All presented models were stable. We used the Bayesian Information Criterion (BIC; Schwarz, 1978) for model selection since this fit measure provides a good balance between goodness-of-fit and parsimony (Raftery, 1995). In addition, BIC-weights, $P(BIC)$, are presented to facilitate the interpretation of BIC differences. BIC-weights are transformed values of the BIC differences to a probability scale representing the probability of each model being the best model given the data and the set of candidate models (Wagenmakers & Farrell, 2004).

For the estimation of the WSM only responses to the conflict items were analyzed since simple items can be solved without the integration of the two dimensions and therefore do not discriminate between differences in the integration strategy (Wilkening & Anderson, 1982).

3.3 Results

To investigate whether children in the Math Garden version of the balance-scale task understood the task we first fitted the exploratory model to WD items. All children should succeed on these items. The LCM with two classes showed the best fit (see Appendix to Chapter 3). Responses of children assigned to the class with high probabilities (N=667) of a correct response (on average 93% correct) were included in further analyses. Of the selected children, 603 played the task before the start of the study, and made on average 800 items (SD = 965, range = 1-7695). Subjects with missing responses were only excluded if the missing response corresponded to the investigated item type, resulting in a different number of children for each analysis. 566 children responded to all selected items. In the next section we compare the results of the exploratory model, item heterogeneity model and the item homogeneity model, per item type in the two data sets.

LCM per Item Type

Distance For the JM data set, the three class item homogeneity model was the best fitting model for D items. The observed response probabilities of each class are presented in Figure 3.2. The three classes resembled respectively children that provided balance responses (Rule I), provided the correct left responses (Rule II or more advance strategies), or predicted that the side with smallest distance goes down. See Appendix to Chapter 3) for the goodness-of-fit statistics of all models. Although JM concluded that the responses of children were best described by four qualitatively different rules, the BIC indicated that the three-class model showed the best fit for the paper-and-pencil data set. This difference results from a different model specification. JM analyzed direct response probabilities,

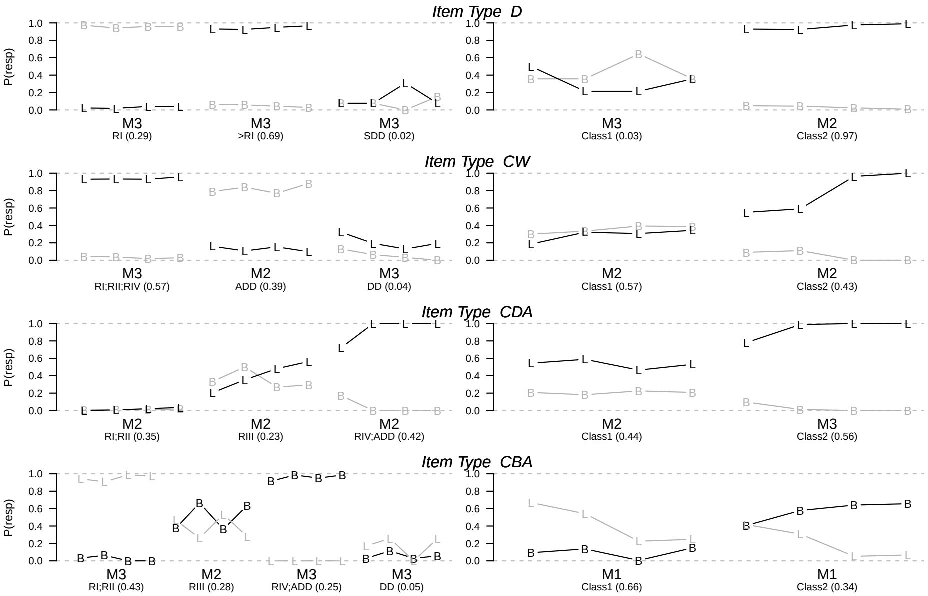


Figure 3.2: The observed response probabilities (y-axis) of the left (L) and balance (B) response of respectively the paper-and-pencil data set (left panels) and the Math Garden data set (right panels), ordered on the product-difference or weight-difference for CBA items (x-axis). M1 (exploratory model), M2 (item heterogeneity model) and M3 (item homogeneity model) indicate which model provided the best fit. The x-axis labels show the class description of JM for the paper-and-pencil data set, and the prior probabilities between brackets. The Small-Distance-Down, Distance-Dominant and Addition Rules are abbreviated as SDD, DD, and ADD

whereas we used a logit transformation of the odds ratios (see Methods section). As a result some conditional probabilities of JM were zero and therefore these parameters did not contribute to the model fit, which is not possible in the logit model specification. For the Math Garden data set, two classes were needed to describe the observed responses. The first class showed an average probability of the correct left response of .36, and the product-difference did not relate to the response probabilities (item homogeneity model). This class is described as guessing behavior. The second class showed a high probability of the correct response indicating that these children use a more advanced rule than Rule I. Furthermore, for this class the probability of a correct response was higher for items with a large product-difference (item heterogeneity model) indicated by an increase in the left-right and left-balance odds ratio. The first latent class found by JM, described as Rule I, was not found in the Math Garden data set.

Conflict-Weight For the JM data set, the three-class model showed the best fit. These latent classes resembled the classes found in JM, described as: a class of children with near perfect responses (Rule I, Rule II or Rule IV), a class of children using the addition rule and a class of children that perceived the distance dimension as the dominant dimension (DD). For each latent class the item homogeneity model resulted in the best fit, i.e., item responses were homogeneous across different product-difference values. For the Math Garden data set, the two-class model showed the best fit. Moreover, the item heterogeneity model fitted better than the exploratory model and the item homogeneity model. The first class showed a low probability of the correct response. The second class showed an overall high probability of the correct response corresponding to Rule I, Rule II or Rule IV (the first class in the paper-and-pencil data set). This indicated that children in the second class perceived the number of blocks as dominant whereas children in the first class perceived distance as dominant. The positive relation between the response probabilities and the product-difference of the item showed that responses of children improved with increasing product-differences.

Conflict-Distance-Addition The LCM for the JM data set resembled the results of JM, and consisted of three classes resembling Rule I or Rule I, Rule III and Rule IV or an addition rule, respectively. Moreover, the item heterogeneity model resulted in the best fit for each latent class. These results correspond to the results of JM, since they also found that the response probabilities of CDA items could not be constrained over items that differed with respect to the product-difference. Even for children using Rule I or Rule II (class 1) the probability of the correct response increased as a function of the product-difference. In the Math Garden data set, the two-class model showed the best fit. In the first class the item heterogeneity model and in the second class the item homogeneity model resulted in the best fit. The first class showed an average probability of the correct response of .5. Children in the second class showed a probability of the correct response of .9.

Conflict-Balance-Addition In the JM data set, the four-class model showed the best fit, resembling the results of JM. Children in the first class had a high probability of the left response (the side with the largest number of blocks), resembling Rule I or Rule II. Moreover, the LCM with a negative effect of the weight-difference in the second latent class (Rule III) resulted in the best fit. For children in this class, the probability of a correct response was smaller for items with a large differences in the number of blocks between the sides of the fulcrum. The response probabilities of the third class are described by JM as produced by children who use Rule IV or the addition rule. For the Math Garden data set, the two-class exploratory model showed the best fit. Hence, the variation in the observed response probabilities cannot be explained by the weight-differences of the items. Also, the LCM did not reveal a class of children with a high performance on CBA items.

Conclusions The LCMs based on the paper-and-pencil data set replicated, in general, the class structure found by JM. In contrast, the models based on the Math Garden data set deviated in number and description of the classes. In eight out of thirteen latent classes in the models for the paper-and-pencil data set, the responses of children were best described by the rule-based item homogeneity model, but this model was the best model in only two out of eight latent classes of the models for the Math Garden data set. In the majority of the classes in Math Garden data set the item heterogeneity model appeared to be the best model.

Mix of Item Types

The following analyses concerned responses to multiple item types. We estimated a second set of hybrid and RB LCMs and applied the WSM to a selection of items of different item types.

LCM In the LCM it is assumed that the responses to items of the same type can be modeled as repeated measures, only allowing variations as a function of the product- or weight-difference of the items. This assumption is not met for the item types where the exploratory model showed the best fit in the previous analysis (see results of the CBA items in the Math Garden data set). Therefore, in the Math Garden data set responses to all D, CW, and CDA items and only the last CBA item were selected and in the paper-and-pencil data set all responses were selected.

Paper-and-Pencil data set We estimated LCMs with one to ten latent classes. As can be seen in Table 3.3, the BIC and *p*(BIC) indicated that the LCM with nine classes showed the best fit. Furthermore, the RB LCM resulted in a better fit than the hybrid LCM (see Table 3.3).

Figure 3.3 shows the response probabilities of the nine classes. The first six classes represented a clear Rule I, Rule II, a small-distance-down (SDD) rule, a distance-dominance (DD) rule, addition (ADD) rule and Rule IV, replicating the findings of JM. Moreover, the average person fit (the posterior probabilities of class membership) of these classes showed that subjects could be rather clearly assigned to most of these classes, respectively .95 (SD=.09), .65 (SD=.18), .77 (SD=.22), .67 (SD=.17), .65 (SD=.15) and .75 (SD=.17). The fourth class, representing the DD rule, was also found in the LCM results per item type. This class was probably not found by the analyses of a mix of item types by JM because of a lack of power. The higher power is achieved by a different item selection and the use of item covariates in the LCM. The sixth class, representing Rule IV, showed perfect performance on all items.

The remaining two classes in JM were interpreted by JM as either Rule III or Rule III/ADD. The current analyses led to three extra classes rather than two, probably as a result of the higher power. The posterior probabilities of the LCM showed that the classification of children to rules was rather ambiguous for these remaining classes, indicated by the high variation and the overall low fit of respectively, .56 (SD=.15), .57 (SD=.19) and .63 (SD=.19), for class 7, 8 and 9 (see Figure 3.4). Hence, the response probabilities cannot be

Table 3.3: Fit Results LCM mix of item types

Paper-and-Pencil data set						Math Garden data set					
Model	NC	NPar	BIC	1:p(BIC)	2:p(BIC)	Model	NC	NPar	BIC	1:p(BIC)	2:p(BIC)
Hybrid	6	101	12357	<.001		Hybrid	2	29	9541	<.001	
Hybrid	7	118	12343	<.001		Hybrid	3	44	9465	.108	
Hybrid	8	135	12336	<.001		Hybrid	4	59	9461	.892	>.999
Hybrid	9	152	12315	>.999	<.001	Hybrid	5	74	9516	<.001	
Hybrid	10	169	12373	<.001		Hybrid	6	89	9565	<.001	
RB	9	88	12152		>.990	RB	4	35	9510		<.001

Note. For the paper-and-pencil (N = 779) and Math Garden (N = 566) data set, responses of 16 and 13 items, respectively, were analyzed; NC = number of latent classes; NPar = number of parameters; p = BIC weight (1) for comparison of models with different number of classes and (2) for the comparison of the hybrid and RB model.

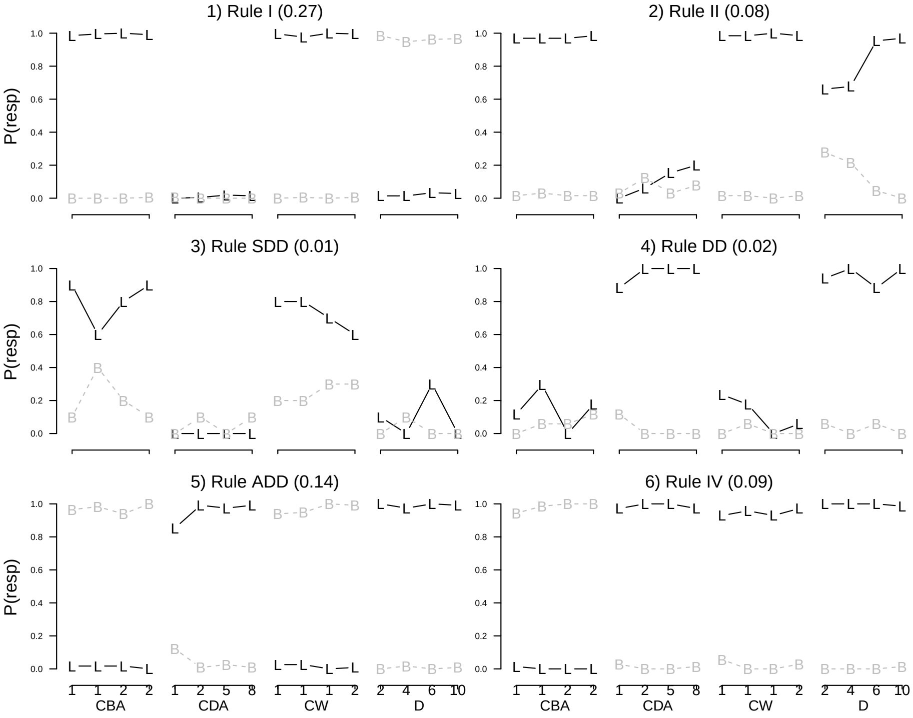


Figure 3.3: Paper-and-Pencil data set. The plots show per class the response probabilities, per item type ordered on the quantitative item effect (on the x-axis).

reliable interpreted as governed by a distinct set of rules. Therefore, these classes are only loosely described as: a distance dominant class providing a lot of balance responses (class seven), a class providing left or right responses (eight) and a class that guessed between the left and balance response (nine).

To conclude, in general, the results of JM are replicated with the new LCM. The gained power to detect individual differences resulted in two additional classes. The person fit indicated that subjects assigned to these latter classes showed a high response variability. Hence, the response patterns were difficult to interpret and could not be ascribed to a clear set of rules.

Math Garden data set For the Math Garden data set, the fit of the sequence of LCMs indicated that four classes were needed to describe the responses, according to the BIC (Table 3.3). Figure 3.5 provides a description of the LCM. The first class (Weight Dominant) had a high probability of the correct response on CW and a low probability on CDA items. Furthermore, the high probability of the left response on the CBA items showed that subjects perceived the number-of-blocks dimension as more dominant. These response probabilities resembled to some extent Rule II. The second class showed high performance

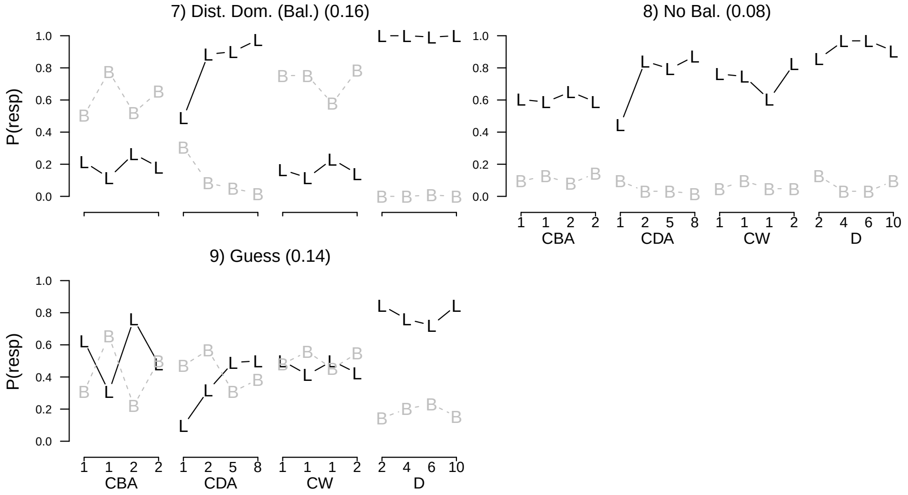


Figure 3.4: Paper-and-Pencil data set: Response probabilities of the LCM on a mix of item types. The plots show per class: the response probabilities, per item type ordered on the quantitative item effect (on the x-axis).

on all item types, except on the CBA item. Again, the high probability of the right response on the CBA item indicated that children in this class perceived the distance dimension as more dominant than the number-of-blocks dimension (for CBA items the right side is the side with the largest distance). In the third class the probability of a correct response was higher on CDA items than on CW items, and highest for D items. Moreover, the high probability of the right response on the CBA item indicated that the distance dimension is perceived as dominant. The forth class mostly resembled the third class, with the addition that the response probabilities for a balance response were considerably lower on CDA, CW and CBA items compared to the third classes.

In general, Figure 3.5 shows that none of these classes resembled Rule I, Rule II, SDD or Rule IV, but rather resembled variations of Rule III. Also, distance-dominant classes were found that have not been reported earlier in paper-and-pencil versions of the balance-scale task. As indicated by the BIC-weight, the response probabilities depended on the product-difference of the item. The probability of the correct response is higher for items with a larger product-difference. Finally, the average posterior probabilities of the LCM, respectively .58 (SD=23), .64 (SD=11), .62 (SD=.19) and .59 (SD=.16) indicates that children could not be clearly ascribed to one of the four classes.

Conclusion A comparison of the results of the LCM of both data sets show that large differences are present in the response mechanism. This is alluded by the better fit of the hybrid LCM in the Math Garden and the rule-based LCM in the paper-and-pencil data set. Moreover, in the paper-and-pencil data set the majority of children could be

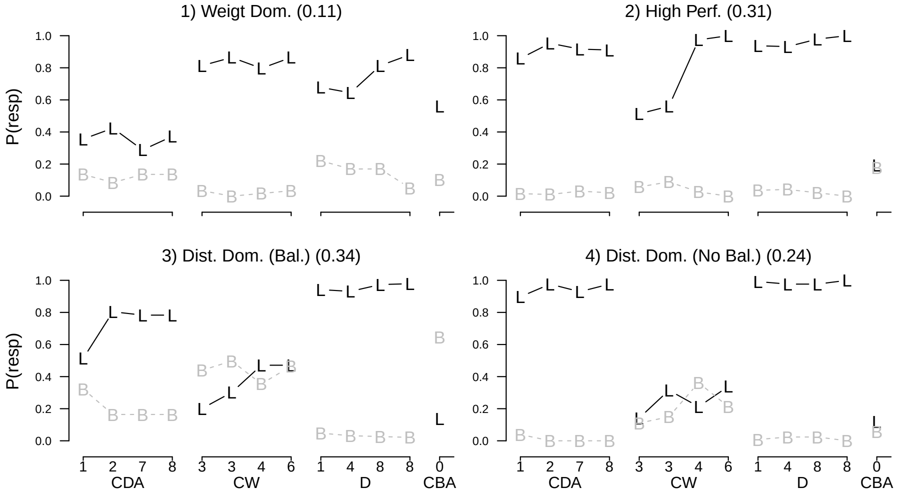


Figure 3.5: Math Garden data set: A description of the four classes of the LCM on a mix of item types. The response probabilities are depicted (on the y-axis), per item type ordered on the quantitative item effect (on the x-axis).

clearly ascribed to latent classes representing qualitative different rules, earlier described by Siegler (1981) and Jansen and van der Maas (2002). In the Math Garden data set the four classes did not resemble any earlier found strategies. Also, the overall lower posterior probabilities showed that differences between the children were more of a quantitative nature when tested in the Math Garden.

Age and Practice Effects JM already showed that large age differences are present between children classified to different classes in the paper-and-pen data set. Using the latent class models introduced in the current paper, we investigated the relation between the dependent variable class membership in the best fitting latent class model (nine classes), and the independent variable age using multinomial regression models. Different models are compared based on the BIC. Results of the paper-and-pencil data again showed large age effects (BIC of model with and without age was respectively 2673 and 3113). In the Math Garden data set, age was not related to class membership (BIC of model with and without age was respectively 1140 and 1125). However, the class membership was related to the amount of practice (BIC of model with and without practice was respectively 1120 and 1125). Practice was defined as the log of the number of items made before the start of the data collection. We use the log function to transform the skewed distribution of the number of item made per child to a normal density. Figure 3.6 shows the predicted probability of a child being assigned to each class as a function of age for the paper-and-pencil data, and as a function of practice for the Math Garden data.

In line with the previous results, large differences are found between both analyzed

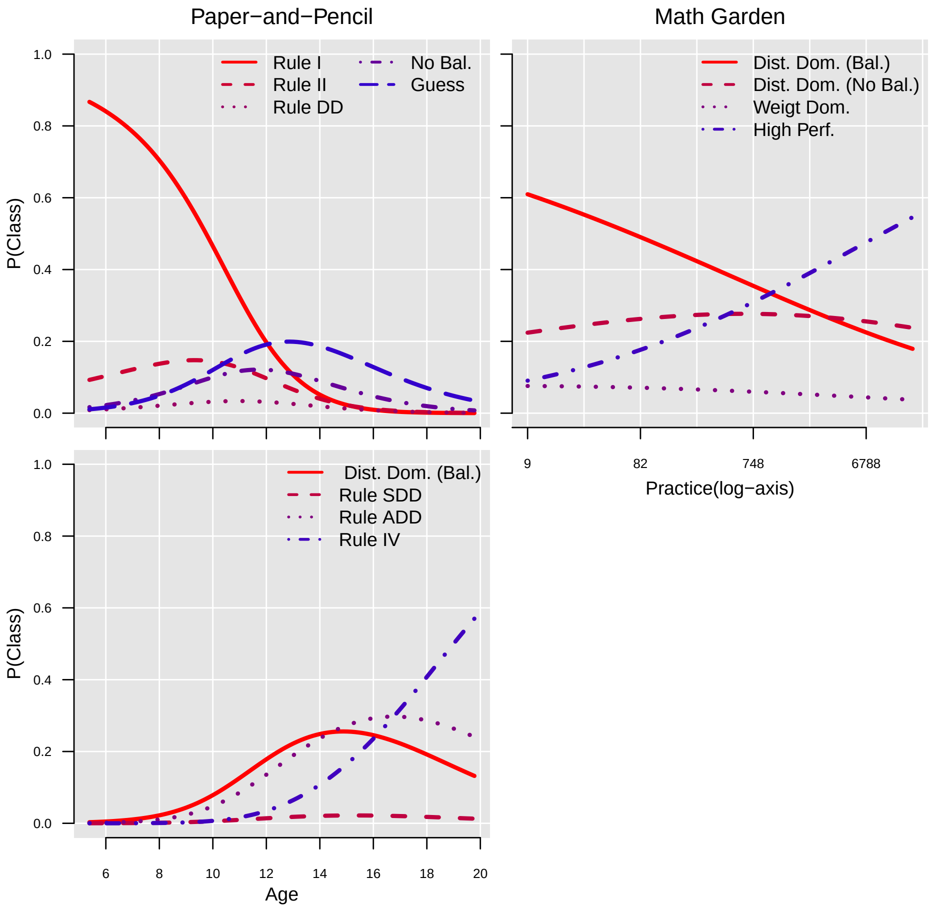


Figure 3.6: The effects of Age in the Paper-and-Pencil data (left) and practice in the Math Garden data (right) on the class membership of the latent class models

data sets. In the paper-and-pencil data a clear developmental change is highlighted by the age effect (further described by JM). In the Math Garden data the developmental pattern is solely based on the amount of practice.

WSM Figure 3.7 shows the distribution of the estimated α and C parameters of the WSM based on responses to the four CW, CDA and CBA items. In the paper-and-pencil data set the distribution of α was clearly not unimodal, and deviated from a normal density as indicated by the Shapiro-Wilk test (Shapiro & Wilk, 1965) (D = .226, p < .001). The large peak at $\alpha = 1$ reflected that some children (N=277, 36%) only responded to the number-of-blocks dimension, including children using Rule I and II (Wilkening & Anderson, 1982). The smaller peak at $\alpha = 0$ indicated that only the distance dimension was reflected in the

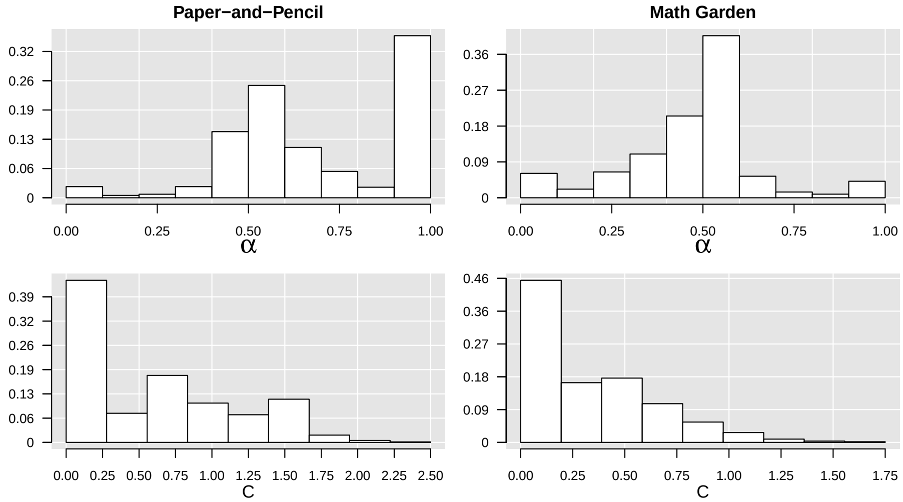


Figure 3.7: Distributions of α, C of the WSM for the paper-and-pencil and Math Garden data set.

responses of 2.4% of the children (N=19). Both values of α indicate that these children did not integrate the information regarding both dimensions. Furthermore, the distribution around $\alpha = .5$ illustrated that the remaining children weighed both dimensions about equally in their responses. The distribution of C showed that 45% of the children already predicted that the scale would tip to a side when their integration of both dimensions resulted in a value just above zero (note that this does not mean they did not provide any balance answers).

In the Math Garden data set, the distribution of α showed a different pattern - again the distribution deviated from a normal density (D = .145, p < .001). In contrast to the paper-and-pencil data set, the peak at $\alpha = 1$ was small (N=23, 4.2%). The large distribution around $\alpha = .5$ showed that the majority of the children weighted both dimensions about equally. However, also a small peak at $\alpha = 0$ was found representing children who only took the distance-dimension into account (N=34, 6.2%). The distribution of C resembled the distribution in the paper-and-pencil data set. The majority of the children decided that the scale would tip if their outcome of the weighted integration of the differences between the arms was higher than zero.

Conclusion The distribution of α and C indicated that also qualitative differences were present since differences between children cannot be described by an unimodal distribution. Moreover, as mentioned previously, a substantial group of children did not integrate information of both dimensions. Hence, a hybrid WSM model is needed to provide a description of the full range of individual differences. However, further developments of the WSM are needed to investigate this. The estimation of the WSM to responses of multiple subjects, and the formulation of the random parameters therein, should provide a test on distribution of these parameters, resulting in a formal test of the InI versus hybrid account. However, a visual inspection of the distribution of the model parameters over persons clearly indicates that a rule-based component is needed to fully explain the observed responses within the WSM framework.

3.4 Discussion

The aim of the paper was to compare a RB and an InI perspective on the cognitive processes used by children to solve balance-scale items, using a new set of statistical models.

According to the LCM analyses aspects of the InI perspective are required to describe the Math Garden data and the CDA items in the paper-and-pencil data set. The results of the WSM, allowing for quantitative (continuous) differences between children in the preference of the number-of-blocks or distance dimension and the preference for balance responses, indicate that quantitative and qualitative differences show up in the inspection of the distribution of the estimated parameters. Hence, results of both statistical models support a hybrid account integrating RB and InI perspectives.

Although we found additional classes in the paper-and-pencil data set, the majority of children can be clearly assigned to one of the rules described by Siegler (1976, 1981) and Jansen and van der Maas (1997, 2002). None of the classes in the Math Garden data set resembles any of these earlier proposed rules. The results indicate that children tested within Math Garden integrate the number-of-blocks and distance dimension to solve balance-scale problems. However, although some children did play the task intensively prior to this study, the LCM did not reveal any children with a perfect integration rule (RIV users). Additionally, whereas Siegler (1976) stated that the number-of-blocks dimension is the dominant dimension, both the LCMs and the WSM reveal that a subset of children perceive the distance dimension as dominant.

In the Math Garden data, the response probabilities are related to differences in the product-difference between items, and to a much smaller extent in the paper-and-pencil data set. This undercuts the conclusions by Jansen and van der Maas (1997) and van Rijn et al. (2003) that this item characteristic was only related to the response probabilities of items with extreme product-differences. Based on a latent-class regression modeling approach resulting in more power to detect an effect of the product-difference, our results indicate that items with a larger product-difference are easier than items with a small product-difference even for items with a reasonably small product-differences. Moreover, the magnitude of this effect differs between both data sets.

Although in both data sets a hybrid account is evident to fully explain the responses of children, differences between both data sets are present as well. In the classical paper-and-pencil version of the task, collected under the standard task demands, cognitive processes are best described by a RB perspective, with the exception of the product-difference effect that follows from a InI perspective. Testing children within the Math Garden, with direct feedback, time-pressure and a rewards system, seems to induce a different cognitive process, providing more evidence for elements of an InI perspective. Where the debate between the RB and InI perspectives in the field of proportional reasoning is concerned with the underlying mechanisms of one cognitive process (or a single response mechanism), the results of this study indicate that the characteristics hereof might depend on the task demands. Positioning the findings based on the Math Garden data alongside the findings of the paper-and-pencil data set suggests that different response mechanisms are at play. This result sheds new light on the debate of RB and InI perspective in the balance-scale literature.

This study was not designed to investigate and isolate the effect of task demands. Also, both age and amount of experience with the task of the tested children differs between both data sets, and have a different relation to the latent classes. Further research is needed to determine which factors influence the response mechanism of children. However, it is surprising that so far, the predictions following from both rule-based and information-integration perspectives on children's knowledge on the balance-scale task, have mainly been tested with only one type of empirical data: responses to a paper-and-pencil test and the computer analogue thereof (Jansen & van der Maas, 1997, 2002; van Maanen, Been, & Sijtsma, 1989). This is even more surprising since Ferretti and Butterfield (1986) already showed that rule assignments differ when children are asked to rebuild one side of the scale instead of predicting the movement.

In other fields of cognitive psychology it is known that task demands influence the type of cognitive processes (or response mechanisms) that are activated or learned. For example, in category learning, differences in the type of task result in the use of qualitatively distinct learning systems (Ashby & Maddox, 2005), and task demands such as time-pressure and feedback have different effects on these distinct learning systems (Maddox, Ashby, & Bohil, 2003; Maddox & David, 2005). Maddox, Bohil, and Ing (2004) show that the performance on a rule-based learning task is impaired when subjects have a short period to process the feedback after a response, while this manipulation did not affect the performance of subjects using information-integration (or similarity) based learning processes.

Therefore, we argue that the differences between the results of both data sets in the present study, are best understood by relating these differences to the differences in the task demands under which children are tested. Based on the described literature, it is expected that the influence of feedback, time-pressure and/or a reward system promotes the usage of different processes. This possible influence of task demands on the response mechanism and an appeal for the integration of RB and InI perspectives in a model of development is already made by (K. W. Fischer & Silvern, 1985, p.626): "under certain conditions of observation and degrees of abstraction, universal stages of cognitive organization can be observed; under others, important individual differences in developmental sequences occur." They conclude that: "What is needed is a view fully grounded in the fact that cognitive development appears diverse under some observational conditions and universal under others." This is also alluded to by McClelland (1995), since he states that rule-like behavior can be induced by different testing situations.

To make the RB perspective compatible with the current results, at least one of the available response mechanisms should be of a more quantitative (similarity-based) nature. The description of Rule III (Siegler, 1976) production model provides such a possibility. Siegler describes children using Rule III as "muddling" through. This strategy could include a mixture of implicit information integration strategies and a preference could be present for either the number-of-blocks or the distance dimension. Moreover, for these children the responses could be based on quantitative item characteristics resulting in the presence of for example a relation between the product-difference and the response probabilities.

To make the InI approach compatible with the current results, it would be necessary to incorporate some qualitative rule-based effects, as found in the LCMs of both the paper-and-pencil and Math Garden data set. The work of Dandurand and Shultz (2009, 2013) already combines RB effects in an InI approach by including an external learning module in which the model is 'taught' RIV - the correct rule where the difference is calculated between multiplication of the weights and distance on each side of the fulcrum. This approach is based on the assumption that children might also learn this rule in an educational setting from instruction instead of from their own experience, which makes it an explicit rule. Such an interpretation of RIV performance fits very well in a rule-based approach. Furthermore, Schapiro and McClelland (2009) also propose a combination of RB and InI processes. They state that: "It is possible that the best account will involve a mixture of explicit and implicit strategies."

To describe the cognitive processes of children used on a proportional reasoning task like the balance-scale task, a model is required that (1) incorporates both a RB and a InI account and (2) specifies in what conditions the behavior is caused by which account. Hybrid models with components relying on rule-based and similarity-based processing of items have become the norm in modeling categorization learning, for example COVIS (Ashby & Alfonso-Reese, 1998) and Atrium (Erickson & Kruschke, 1998). These models can serve as a valuable starting point for including multiple response modes based on different response mechanisms for development of proportional reasoning in general and balance-scale learning specifically under different task demands.

4 Fast and Slow Strategies in Multiplication

This chapter is submitted as: Hofman, A. D., Visser, I., Jansen B. R. J., Marsman M. & Van der Maas, H. L. J., (Submitted). Fast and Slow Strategies in Multiplication. *Learning and Individual Differences*; Preprint at: `10.17605/OSF.IO/AW3QQ`

Abstract

In solving multiplication problems, children use both fast, retrieval-based, processes, and, slower computational processes. In the current study, we explore the possibility of disentangling these strategies using information contained in the observed response latencies using a method that is applicable in large data sets.

We used a tree-based item response-modeling framework (De Boeck & Partchev, 2012) to investigate whether the proposed qualitative distinctions in fast and slow strategies can be detected. We analyzed responses to two sets of multiplication items, totalling more than 500.000 responses, collected with an online computer-adaptive training environment for mathematics.

Results showed qualitative differences between the fast and the slow strategies. Building on these results, both item and person characteristics were differently related to fast and slow processes. These characteristics, resulting from substantive models of multiplication, allowed us to further describe the fast and slow strategies. Results emphasize the quantitative and qualitative differences between strategies used for solving multiplication problems, and provide possibilities for tailored feedback on learning multiplication.

4.1 Introduction

The concept of strategy is central in the study of human problem solving. Important aspects of problem solving behavior such as accuracy, duration, and type of errors, are due to the choice of solution strategy. For instance, in solving arithmetic items, people may use either retrieval from memory or a computational strategy (Dowker, 2005; Ashcraft & Guillaume, 2009; LeFevre et al., 1996), where the former typically requires less time than the latter. In the case of basic multiplication (for example single-digit problems), detailed models for the retrieval process exist (Geary, Widaman, & Little, 1986; Verguts & Fias, 2005), and several models for computational strategies have been developed as well (Lemaire & Siegler, 1995; Imbo, vandierendonck, & Rosseel, 2007). These models make different predictions about item difficulty and solution time (van der Ven et al., 2015).

When measuring arithmetic ability by using psychometric tests, such as in IQ tests, individual differences in strategy choice are usually not taken into account. Arithmetic ability is ultimately tested by counting the number of correct items that participants solve in any particular test (e.g., Liu, Wilson, & Paek, 2008; Aunola, Leskinen, Lerkkanen, & Nurmi, 2004). Different patterns of response times and errors are hence ignored when the aim is to compare individuals on a scale of arithmetic ability. Using the number of correct responses may be warranted when testing and comparing test takers, but may be inappropriate when concerned with studying development and understanding ability differences. In the latter case, different qualitative processes or strategies should be considered.

For example, an important developmental trend in learning arithmetic can be described by changes in strategy choice. Initially children will apply various slower computational strategies (Freudenthal, 1991). Over time, these computations become more sophisticated

(Lemaire & Siegler, 1995). Through practicing multiplication, children will build up a network of associations between numbers. When this network is sufficiently strong, children will be able to confidently retrieve answers to items, and will tend to use faster retrieval from this network instead of a slower computational strategy (Siegler, 1988). Children with learning difficulties do not show this typical transition from computational to retrieval strategies (De Visscher & Noël, 2014; De Smedt, Holloway, & Ansari, 2011). After years of practice, adults will rely predominantly on memory retrieval for single digit multiplication (LeFevre et al., 1996). Hence, the largest divide in strategy choice is whether children and adults use a retrieval strategy or a computational strategy.

In spite of the importance of the strategy concept, detecting strategies is still a major challenge in many areas of cognitive science. Verbal reports and neural imaging features are both correlated with strategy choice (Jost, Beinhoff, Hennighausen, & Rösler, 2004; Tenison, Fincham, & Anderson, 2014; Price, Mazzocco, & Ansari, 2013), but both also have pitfalls as strategy indicators. Verbal reporting, the most commonly accepted method of strategy detection, may interfere with the solution process and bias strategy choice (Kirk & Ashcraft, 2001; Reed, Stevenson, Broens-Paffen, Kirschner, & Jolles, 2015). Another important problem with relying on verbal reporting for detecting strategy choice is that it is time-consuming to apply and thus not feasible in combination with large scale automatic assessment of arithmetic abilities. The latter problem also applies when using neural patterns to identify strategy choice. A third approach, whereby strategies are assessed through latencies combined with accuracy, is more promising in the context of large scale assessment of arithmetic problem solving as retrieval strategies are usually much faster than computational strategies (e.g., LeFevre et al., 1996). Hence, here we explore the possibilities of including response latencies in measurement models of arithmetic performance to disentangle possible qualitative differences between strategies.

In this paper we investigate whether the fast-slow model (Partchev & De Boeck, 2012; DiTrapani, Jeon, De Boeck, & Partchev, 2016) allows for automatic analyses of strategy use in a large scale data set of arithmetic performance in children. In particular, we focus on multiplication problems as this is a well-studied domain. The fast-slow model is based on splitting the data into fast and slow responses and estimating separate abilities for each of the processes. A third process, based on the response latencies, indicates choice for the fast or slow process. The advantage of this type of psychometric model is that item and person effects are easily disentangled. This approach is intermediate between the purely psychometric approach of fitting IRT models to capture multiplication ability on a single latent trait (e.g., Liu et al., 2008; Aunola et al., 2004) and the purely cognitive approach of using computational models to predict response accuracy based on problem characteristics and strategies (partial abilities; e.g., de la Torre & Douglas, 2008).

We will first introduce the fast-slow model, derive predictions for the case of multiplication, and then apply the model to a large data set. This data set includes a large set of responses collected with a popular Dutch online adaptive learning environment for mathematics; the Math Garden (Klinkenberg et al., 2011; Straatemeier, 2014).

The Fast-Slow Model

The fast-slow model is a tree-based item response theory (IRT) model (De Boeck & Partchev, 2012). The rationale of this model is that responses are governed by one of two processes, one fast and one slow, that can be separated by an additional observed variable, in this case the (recoded) response times. The response times are recoded to either fast (1) or slow (0), which serves as an approximation of the underlying process and is modelled as a latent speed dimension. This tree model can be formulated as follows, assuming that a (unidimensional) Rasch model (Rasch, 1960) holds in dimension d, where $d = 1, 2, 3$ denotes the speed-, fast- and slow dimension, respectively. In these dimensions respectively the probability of a fast response, a fast and correct or a slow and correct response are modelled using a Rasch model. In the Rasch model, the probability of a correct (or for the speed dimension a fast) response of a person p on an item i in dimension d is given by the logistic function:

$$P(x_{pid} = 1|\theta_{pd}, \beta_{id}) = \frac{exp(\theta_{pd} + \beta_{id})}{1 + exp(\theta_{pd} + \beta_{id})}, \tag{4.1}$$

where θ_{pd} denotes the ability of person p and β_{id} denotes the easiness of item i on dimension d. Hence, the full model has three sets of person parameters, and three sets of item parameters: θ_{p1} reflects the overall speed of a person, θ_{p2} reflects the ability to give a fast and correct response, and θ_{p3} reflects the ability to give a slow and correct response. Likewise, item easiness parameters correspond to the probability that items are answered fast versus slow (β_{p1}), the probability of a correct response given that the response was fast (β_{p2}), and the probability of a correct response given that the response was slow (β_{p3}). In line with De Boeck (2008), both $\boldsymbol{\theta_p} = (\theta_{p1}, \theta_{p2}, \theta_{p3})$ and $\boldsymbol{\beta_i} = (\beta_{i1}, \beta_{i2}, \beta_{i3})$ are treated as random variables with $\boldsymbol{\theta_p} \sim \mathcal{N}(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$ and $\boldsymbol{\beta_i} \sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta})$, constraining $\boldsymbol{\mu_\theta}$ to zero to identify the model (see Appendix to Chapter 4 for a description of the model estimation procedure).

Empirical Predictions in Relation to Fast versus Slow Multiplication Processes

If fast and slow strategies are found to be qualitatively different, some item and person effects are expected to be differently related to fast and slow strategies. If these effects match common findings in the multiplication literature, the fast-slow model is a useful method to identify strategies at the individual level in a big data set.

Item effects

We focus on three prominent empirical effects; the problem-size effect, the tie-effect and effects of special operands, which are associated with systematic differences in accuracy and response times between items. Models of retrieval and computation strategies in simple multiplication have coined different explanations for these differences.

1) The problem size effect (Ashcraft & Guillaume, 2009) refers to the fact that items with large problem sizes are more difficult than items with smaller problem sizes. According to models of computational strategies this effect is due to the additional steps necessary for computing the answer (van der Ven et al., 2015; LeFevre et al., 1996). In retrieval based models this effect is explained by less frequent practice with items with large operands and therefore a less developed memory network (Ashcraft, 1995). Thus, no differences are expected between fast and slow processes with respect to the problem size effect.

2) The tie-effect (Miller, Perlmutter, & Keating, 1984; De Brauwer, Verguts, & Fias, 2006) implies that ties (items with an equal operand; e.g 7 x 7) are easier than other items. This effect is explained by more practice and easier storage in retrieval based models. Models of computational strategies do not predict a tie-effect since the computations involved in ties are the same as in non-tie items. Hence, a tie-effect is expected in the fast process, which is expected to be associated with retrieval, and no tie-effect is expected in the slow process which is expected to be associated with computational strategies.

3) The special operands effect refers to the finding that items with 1, 2, 5 or 9 as operands are easier than other items (Lemaire & Siegler, 1995). This effect follows from easier computations according to computational accounts, but is not predicted in models of retrieval. Hence, the effect of special operands is expected in the slow but not in the fast process.

Person effects

As explained in the introduction the development of simple multiplication skills involves a shift from computational strategies to retrieval. This shift is expected to be reflected in a higher number of fast responses for older compared to younger children, resulting in an effect of age on the latent speed dimension. A gender effect on speed is expected as well, due to individual differences in response styles. In addition and subtraction problems, boys provided more retrieval responses than girls, while girls were more likely to count with their fingers (Carr & Jessup, 1997). It is expected that boys have a higher probability to respond fast than girls.

4.2 Methods

Data sets: Items and Participants

Data are collected with the website Math Garden. Math Garden is an online adaptive learning environment for learning basic arithmetic, that is currently used by more than 200,000 children involving more than 1,500 schools in the Netherlands (see Appendix to Chapter 4). Math Garden provides a valuable data set, including accuracies and response times of a large group of children, on a large set of multiplication items.

For this study we selected responses of children collected between June 1, 2011 and June 1, 2015 on two subsets of all multiplication items: (1) all responses to items belonging to the

Table 4.1: Data description

Item selection	N responses	N children	N items	% missing
Single digit data set	180,651	3,551	64	21
Most played data set	422,634	7,860	145	63

Note. The number of responses, children, items, and amount of missing data for the different constructed data sets. The missing data is introduced by the adaptive item selection.

multiplication tables from two up to nine (64 items in total), referred to as the single-digit data set, and (2) responses to the 150 most played items, referred to as the most-played data set. This second data set includes some of the items from the first subset and additionally includes multi-digit multiplication items (such as: 1×500, 7×100, 9×12, 803×10 and 80×6000). Items with a minimum of 200 encounters were selected, resulting in 145 items. Through analysing the second data set we investigated whether the results from the first data set can be generalised to a data set including responses to a broader set of items. Also, replicating the initial analyses using this second data set provides a check of the robustness of the results.

We discarded the first 90 responses that each child made to allow children to become acquainted with the task. Furthermore, because data were collected longitudinally and abilities tend to change over time we selected a time frame for a single assessment of a child's ability. This time-frame must contain sufficient data but should also be small enough to ensure a relatively stable ability, and was fixed to one week. Additionally, in order to set a minimum number of responses for this time frame, we selected data of children who completed at least 30 items within one week.[1] Only the child's first response to an item was selected (multiple responses for the same item within the time frame are possible). The total number of responses, children, items and percentage of missing responses for each data set are presented in Table 4.1. Note that the same children can be included in both data sets. Since the data were collected with an adaptive algorithm missing responses are missing by design, and can be seen as missing at random (MAR) since the missingness is conditional on the estimated ability (Rubin, 1976; Eggen & Verhelst, 2011).

In order to apply the model, the response times need to be dichotomized into fast or slow categories. In our analyses, we used three different approaches based on a median split: (1) a split on the overall response times distribution; (2) a within person split allocating 50% of the responses of each person to either fast or slow and (3) a within item split allocating 50% of the responses to each item to either fast or slow. The first split captures both person and item differences in speed, whereas the person (item) split only captures differences between items (persons) in speed respectively. A comparison of the results of each of these split-methods provides information on the robustness of the results (see Appendix to Chapter 4).

[1] It was possible to make different choices for selecting data. However, using different inclusion criteria yielded comparable results, see Appendix to Chapter 4

Model Comparison

Within the fast-slow model, qualitative differences between fast and slow processes would be reflected by a different ordering of the item parameters, person parameters or both, in the fast compared to the slow component of the model. Hence, to test the hypothesis that these differences are present, the full fast-slow model with a set of item parameters for both the fast and the slow part was compared against three constrained versions of the model. This resulted in four different models: (1) the full model, (2) constrained item parameters: i.e., $\beta_{i,fast} = \beta_{i,slow}$, (3) constrained person parameters: i.e., $\theta_{p,fast} = \theta_{p,slow}$, and (4) both constrained item and person parameters. If one, or both, constraints resulted in a worse model fit (in terms of prediction; see next section), this would support the notion that indeed different processes were involved in the fast and the slow responses. However, from a measurement perspective different item parameters do not necessarily suggest that the person abilities are different, since these abilities could be highly correlated (the same holds for item parameters if person parameters are different).

Whenever a constraint was imposed we allowed for a difference in the overall mean and in the variances of the fast and slow item and/or person parameters. This reflects the idea that only a correlation between the fast and slow parameters that is significantly lower than one truly reflects a qualitative different process. For example, if fast retrieval responses are more often correct than slow computational responses it does not necessarily suggest that slow and fast responses have distinct response processes. It may be that for slower responses, retrieval is simply more difficult. However, if for some persons or items the slow responses are more often (in)correct than the fast responses, thereby influencing the correlations of these parameters, this would indeed suggest that a different response process is involved.

Cross-validation was used to assess the models' goodness-of-fit. For each person, data from one response (both the recoded response time and the accuracy) were selected for the test data. The remainder of the data were used to estimate (train) the model parameters, and the estimated models were subsequently used to predict the test data. This approach naturally prevents over-fitting the data with overly-complex models. The test data formed between 1.4% and 3.0% of the total data in the different data sets but was still fairly large as, despite including one response per person, a large number of persons were included (see Table 4.1). Model predictions were based only on accuracy as the models did not differ in their analyses of response times.

Three cross-validation statistics were used, all three based on the deviation between the observed and the predicted response: the prediction accuracy (ACC), the root mean squared error (RMSE) and the log-likelihood LL; Pelánek (2015, see Appendix to Chapter 4 for a detailed description). In both RMSE and LL the continuous prediction of the probability of a correct response is analyzed. This results in a finer model comparison than the ACC, while the ACC provides a simpler interpretation of the goodness-of-fit. When interpreting the ACC and the LL, higher (less negative) values indicate better fit, while for

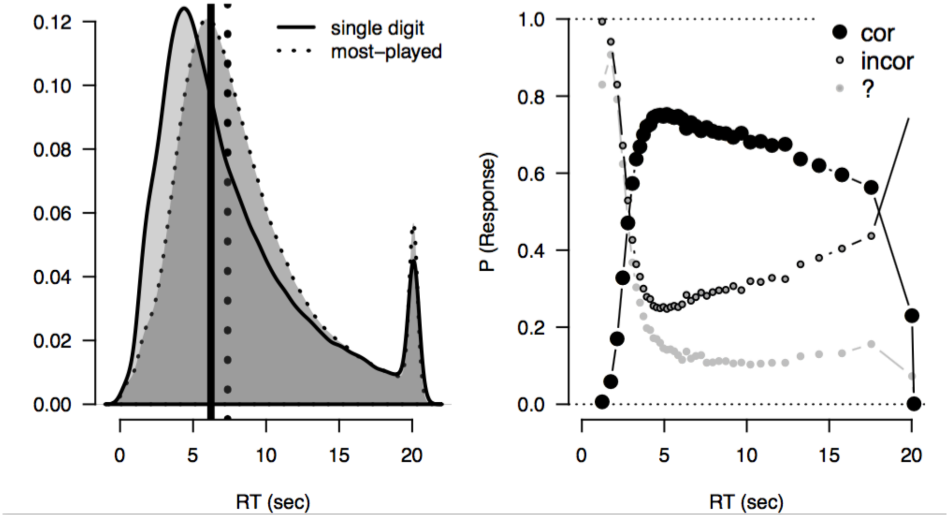


Figure 4.1: Data Description. The left-panel shows the RT distribution for the single-digit and most-played data set. The vertical lines (solid for single-digit and dotted for most-played data set) indicate the median of the RT distribution. The peak around 20 seconds is caused by the deadline in the game. The right-panel describes the proportions of a correct, incorrect and question-mark response for the different observed response times in the single-digit data set.

the RMSE lower values indicate better fit.

4.3 Results

Since the results of the model comparisons were similar across the various dichotomizations, we limit the results section to the analyses from data sets where fast or slow was defined by the overall medium split (see Appendix to Chapter 4).

Data Description

The RT distributions of both data sets are presented in the left-panel of Figure 4.1. For the single-digit data set the median response time (RT) was 6.22 sec. 59% of the fast responses and 62% of the slow responses were correct. The lower percentage for the fast responses was related to the higher proportion of fast question-mark responses: 33% and 11% respectively for fast and slow responses. This is also shown by the relationship between RT and the probability of a question-mark response, plotted in the right-panel of Figure 4.1. In the most-played data set the median RT was 7.36 sec. 72% of the fast and 68% of the slow responses were correct.

Table 4.2: Model fit based on cross-validation of the full and constrained fast-slow models in the single digit and most played item data set.

item selection	model	ACC	RMSE	LL
single digit	full model	**0.777**	**0.391**	**-2416**
	$\beta_{fast} = \beta_{slow}$	0.775	0.397	-2510
	$\theta_{fast} = \theta_{slow}$	0.773	0.398	-2518
	$\beta_{fast} = \beta_{slow}$ and $\theta_{fast} = \theta_{slow}$	0.772	0.397	-2489
most played	full model	**0.750**	**0.416**	**-5239**
	$\beta_{fast} = \beta_{slow}$	0.740	0.422	-5403
	$\theta_{fast} = \theta_{slow}$	0.742	0.420	-5351
	$\beta_{fast} = \beta_{slow}$ and $\theta_{fast} = \theta_{slow}$	0.737	0.421	-5375

Note. Results of the best fitting model are printed in bold.

Model Comparison

To estimate the model parameters we used 1,000 iterations and a burn-in of 100. Since some high auto-correlations were found we used every third iteration for the MAP estimates of the model parameters. Table 4.2 shows the fit measures for the estimated models. In line with our hypothesis, the results indicated that for both the single digit and most-played data set, the model with separate item difficulties and separate person abilities for the fast and slow dimension - the full model - provided a better fit that any of the constrained models in terms of ACC, RMSE and LL (see Table 4.2). This suggested that qualitatively different processes were involved in the fast compared to the slow processes for both the single-digit and the most-played data set.

These results indicate that the response times (split into fast and slow) distinguished between two qualitatively different response processes, both with respect to item and person parameters. In the following sections we will further describe the estimated parameters, and thereby investigate whether differences between the fast and slow strategies can be explained by retrieval and computational models of multiplication.

Fast vs Slow Correlations and Variances

The model comparison indicated that fast and slow item and person parameters are not perfectly correlated since the full model provided a better fit than any of the constrained models. However in both the single-digit and the most-played data set the correlations between β_{fast} and β_{slow} were very high: .969, and .896 respectively for the single-digit and most-played data set. The correlations between θ_{fast} and θ_{slow} were much lower (respectively .778, and .635). The lower correlations between person parameters might be explained by the smaller number of observations for the person parameters compared to the item parameters (which may have created more measurement error). The higher correlations in the single-digit data set compared to the most-played data set can be explained by a more unidimensional process underlying the responses of children in the single digit data set.

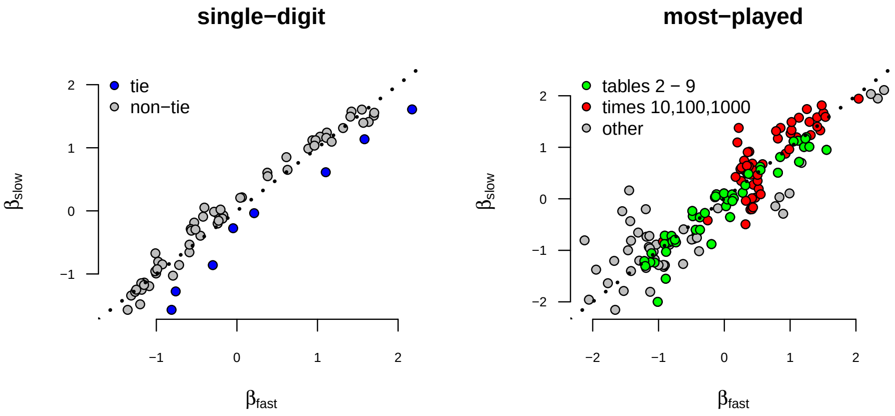


Figure 4.2: Relation between fast and slow item parameters in the single-digit and most played data set

Furthermore, in the single-digit data set, higher variances in β_{fast} compared to β_{slow} were found ($\sigma_{\beta,fast}$ = 1.943 and $\sigma_{\beta,slow}$ = 1.085; Levene's test of equality of variance: F(1, 62) = 30.07, p < .001). This was also the case in the most-played data set, however with smaller differences between fast and slow responses than in the single-digit data set ($\sigma_{\beta,fast}$ = 1.367 and $\sigma_{\beta,slow}$ = .775; Levene's test of equality of variance: $F(1, 145) = 52.58$, p <.001). The lower estimated variance in the slow process could suggest that there is more random variation, compared to structural variance, in the slow responses. This might be caused by a mixture of different slow strategies.

Item Analysis

In the next step in our analyses we regressed the item parameters on different item characteristics for both the slow and fast responses in the single-digit data set. We intended to replicate the effects of problem-size, tie and effects of special operands. Additionally, and most interestingly, here we were able to test for differential effects for fast and slow processing. Finding these differential effects would mean that predictors related to retrieval processes (tie-effect) and/or computational processes (special operands) are differently related to item parameters in fast compared to slow responses. To investigate these interaction effects we imputed the full original data set. To this end we generated a new set of responses based on the model estimated model parameters. We analysed the sum-scores over items for both fast and slow responses. This approach ensured that effects can be directly compared between different nodes.

In separate regression models we predicted the item parameters reflecting the fast and the slow accuracy and the probability of a fast response (speed). We used the BIC (Schwarz, 1978) for model selection, using a backward stepwise procedure.

Table 4.3: Regression of the item easiness parameters for fast and slow processes and speed (reflecting the probability of a fast response) in the single-digit data set.

	Fast			Slow			Speed		
predictor	B	SE	T-value	B	SE	T-value	B	SE	T-value
Intercept	2.851	0.374	7.632**	2.277	0.225	10.139**	0.034	0.167	0.204
problem-size	-0.351	0.032	-11.044**	-0.218	0.019	-11.289**	-0.039	0.013	-3.002*
tie	1.432	0.212	6.765**	ex	ex	ex	0.491	0.104	4.713*
times 2	2.408	0.207	11.607**	1.152	0.125	9.224**	0.757	0.100	7.577**
times 5	1.224	0.168	7.282**	0.762	0.101	7.543**	0.211	0.083	2.538*
times 9	0.817	0.203	4.016**	0.476	0.123	3.876**	ex	ex	ex

Note. * = $p < .05$, ** = $p < .001$; ex = excluded in the stepwise procedure

In line with the predictions, for the fast responses, we found a main effect of: (1) problem-size, indicating that items with large problem size were more difficult than items with small problem size; (2) ties, ties were easier than non-tie items; and (3) problems with special operands two, five and nine which were easier than other problems, see Table 4.3. These main effects explained in total 89.2% of the variance. For the slow responses, the effects of problem-size and the special operands were comparable (resulting in an explained variance of 88.6%). Interestingly, no differences between ties and non tie-items were found in the slow responses. This differential effect indicated that the difference between ties and non-ties was larger in the fast compared to the slow part. This effect is plotted in the left-panel of Figure 4.2, which shows that all tie items are below the diagonal that indicates $\beta_{fast} = \beta_{slow}$. Unexpectedly, effects of items with special operands were not differently related to the fast and the slow item parameters.

For the item speed parameters, a high β_{speed} indicated a high probability of a fast response. Thus Table 4.3 shows that responses to items with large problem sizes were often slow. Also, responses to items belonging to the two and five multiplication tables, and ties were often fast (see Table 4.3) indicating that these items were more often solved by retrieval rather than computational strategies. These effects explained in total 70.6% of the variance in the item speed parameters.

To conclude, the high explained variance indicates that the item difficulties could be largely understood by this set of item features. This supported the reliability of both the data and the model estimation. Moreover, although high correlations between fast and slow item parameters were found, the interaction between tie and node indicates that tie items tap into the differences between fast and slow processes. In line with the results of van der Ven et al. (2015), the tie-effect was more prominent in the fast responses.

Item characteristics were not regressed on item parameters in the multi-digit data set. However, the right-panel of Figure 4.2 clearly shows a positive relation between β_{fast} and β_{slow} in the multi-digit data set. Some items showed higher deviations. An exploratory look at the three items with the highest deviations where either $\beta_{fast} > \beta_{slow}$ or $\beta_{fast} < \beta_{slow}$ showed an interesting pattern. The items 11 x 6, 8 x 8 and 11 x 9 were easier when solved quickly compared to slowly, and the items 80 x 6000, 4 x 108, and 3000 x 80 were easier when solved slowly compared to quickly.

Person Analysis

In the second set of regression models we investigated whether person characteristics were differentially related to fast and slow abilities. For this analyses we only included children between 6 and 11 years old (N=4233; excluded 467), and children for which their age matched their grade (excluded 417 children for which their age deviated more than 1.5 year from the grade average). For the single-digit and multi-digit data sets the average ages of the selected children were 7.86 and 8.42 (SD 1.04 and 1.10) respectively, and 33% and 42% respectively were girls.

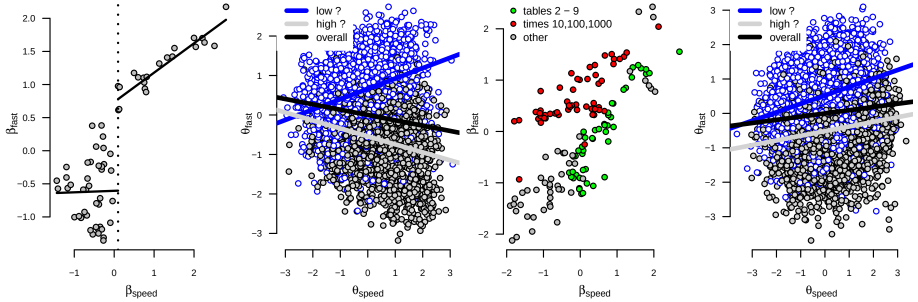


Figure 4.3: Relationship between speed and fast accuracy for items and persons in the single-digit data set (left two panels) and the most-played data set (right two panels). Low and high ? denotes the frequency of question mark usage.

All results, based on a stepwise backward procedure using BIC, are presented in Table 4.4. As expected we found a main effect of age. Older children were more able than young children in both fast and slow abilities. Second, no gender differences were found in both abilities. Third, children with more question-mark responses had a lower ability. However, this effect was smaller with slow compared to fast abilities. This highlights that the differences between the abilities measured by fast and slow responses can partly be explained by differences in how children relate to the question-mark answer option. These effects explain 54.9% and 36.7% of the variance for fast and slow abilities, respectively.

The regression model for differences in speed between children indicated that; (1) older children were faster than younger children, (2) boys were faster than girls and (3) children who provided more question-mark responses were faster than children who did not use the question-mark response as often (see Table 4.4). These effects explain 27.3% of the variance in the abilities between children.

Correlations between Speed and Accuracy

In this last section we explore the relations between speed and accuracy from an item and a person perspective in both the single-digit and most-played data set. We defined speed as the probability of a fast response, based on the overall split in response times.[2]

Item speed and accuracy correlated positively. In the single-digit data set the correlations between β_{speed} and β_{fast} and β_{slow} were .837 and .739. In the most-played data set, these correlations were respectively .694 and .440. The correlations between β_{speed} and β_{fast} are plotted in Figure 4.3.

We observed two interesting results. First, in the single-digit data set, the relationship between speed and accuracy showed an interesting pattern. A regression model with a

[2] The presented results were stable under the different RT splits; the within item split to investigate person speed and the within person split to investigate item speed.

Table 4.4: Regression person ability parameters for fast and slow processes and speed (reflecting the probability of a fast response)

	Fast			Slow			Speed		
predictor	B	SE	T-value	B	SE	T-value	B	SE	T-value
Intercept	-0.038	0.019	-1.971	-0.023	0.012	-1.860	-0.173	0.021	-8.074*
age	0.432	0.020	21.508*	0.267	0.013	20.631*	0.221	0.013	17.077*
gender	ex	ex	ex	ex	ex	ex	0.198	0.026	7.478*
% ?	-1.184	0.020	-58.910*	-0.476	0.013	-36.817*	0.487	0.013	37.786*

Note. Boys are coded as 1 and girls as 0; %? = percentage of question-mark responses; * = $p < .001$; ex = excluded in the stepwise procedure breakpoint resulted in an explained variance of 83.8%, an increase of 13.5% compared to the explained variance of 70.4% of the linear regression model. Furthermore, the breakpoint could be confidently estimated at zero, indicated by a clear peak in explained variance compared to models with differently located non-zero breakpoints. This strongly suggested that, for items that were solved quickly ($\beta_{speed} > 0$), there was a strong relationship between speed and accuracy, whereas for items that were more often solved slowly this relation was absent. In line with results of the model comparison, this result signifies that fast strategies are qualitatively different from slow strategies. Secondly, an exploratory look at the item parameters in the most-played data set showed large differences between items including a times 10, 100 or 1000 operator and the other items, as visualized in Figure 4.3. These items were incorrect more often when answered quickly compared to other items.

For persons, a different pattern was found. In the single-digit data set, negative correlations between person overall speed and fast and slow abilities were found: -.125 and -.033 respectively. Thus, in contrast to expectations, children that were faster were more often incorrect. To test whether the negative correlation was related to differences between children in question-mark usage we calculated separate correlations for children who provided less or more than 13% question-marks (median split). We found a correlation of .306 for children who used fewer question-marks, indicating that for these children, the faster children were more able than the slower children, see the blue line in the second panel of Figure 4.3. This suggested that the negative relation is related to question-mark uses. Furthermore, all correlations found were positive (min = .191 and max = .373) when children were grouped by question mark use from zero to ninety percent in increments of ten percent. The same pattern was found in the most-played data set. To conclude, these results indicate that, when corrected for question-mark usage, children who are faster had higher fast and slow abilities.

4.4 Discussion

In this paper we investigated whether qualitatively different strategies in multiplication can be disentangled using information in observed response times. This approach allowed for an automatic assessment of strategy use appropriate for large-scale data. An application of the fast-slow model on a data set collected with a popular online learning program confirmed that a mixture of different strategies underlies the children's performance on multiplication items. Disentangling fast versus slow strategies improved understanding of children's observed responses.

Building on these results, additional analyses showed that specific item and person characteristics tap into the differences between these strategies. The aim of these analyses was to investigate differential effects between fast and slow strategies, as predicted by fast retrieval versus slow computational processes of multiplication (Siegler, 1988). On the item side, the difference between tie and non-tie items was more prominent in the fast responses compared to the slow responses. Against expectations, no differential effect was found for items of special operands (2, 5 and 9). This could be explained by having used rather crude methods to disentangle strategies. These methods may have allocated some retrieval responses as slow and some computational responses as fast, resulting in a lower power to find these effects. However, varying split methods differing in how responses were categorised as fast or slow show consistent results. Further methodological improvements are possible with developing better ways of splitting response times as the most important one. Ideally, the data itself determines the classification into fast and slow processes, resulting in a more optimal classification of responses to strategies (DiTrapani et al., 2016).

On the person side, older children (who are assumed to have more experience) provided more fast responses. Although older children can be faster in multiple ways, the results indicate that this developmental trend is partly due to a higher probability of a retrieval strategy for older compared to younger children. Additionally, although boys and girls did not differ with respect to both the fast and slow ability - in line with the results of Carr & Jessup, 1997 - boys provided more fast responses than girls. This highlights that, as described by Siegler (1988), individual differences in strategy use are present between children. This difference in strategy use is also reflected by the different relationship between question-mark usage and fast versus slow abilities. Some children are more inclined than others to provide a question-mark response, and furthermore, fast question-marks are governed by a different strategy selection than slow question-marks.

These results confirm that children's strategies for solving mental multiplication items can be disentangled using a split in observed response times. Hence, as described by Siegler (2007) and van der Ven, Kroesbergen, Boom, and Leseman (2012), multiplication ability should be seen as a toolbox of different strategies, where both the ability of each child within a certain strategy and individual differences in strategy selection determine the observed performance. This study indicated that these processes, often studied in smaller and controlled experimental settings, also determine multiplication ability in a large-scale online learning platform, supporting the generalizability of the effects and the validity of the Math Garden.

Future Directions

It should be noted that the mixture of retrieval and computational processes underlying the responses in multiplication will depend on the testing conditions. In the Math Garden items were selected to match children's ability resulting in a mixture of different strategies. Presenting solely easy or hard items will change the mixture of strategies. Additionally, the test conditions were such that children perceived time-pressure. This evokes faster responses, and probably influences the strategies that were used (Hofman, Visser, Jansen, & van der Maas, 2015). Further research should investigate in what manner children's performances in high-stakes tests also depends on multiple processes. Additionally, next to response latencies, error types also contain information about the used strategy (Siegler, 1988). In a first minimal example Coomans, Hofman, Brinkhuis, van der Maas, and Maris (2016) already showed that fast errors in response to multiplication items were different from slow errors. Utilizing both response latency and error types could provide additional confidence in estimating the used strategy.

This line of research will provide applied researchers, teachers and students with valuable information on strategies, without using time-intensive methods such as verbal protocols. First, it sheds light on what cognitive processes are involved in mathematics, and possible many other domains. Second, it enables tailored feedback about proficiency of strategies when learning multiplication, and thereby matches the aims for mathematics education. For instance, in the Netherlands education ultimately aims for understanding multiplication concepts and memorization of the single-digit tables of multiplication (*SLO*, 2009).

5 Tracing the Development of Typewriting Skills in an Adaptive E-Learning Environment

This chapter is published as: van den Bergh, M. Hofman, A. D., Schmittmann, V., van der Maas, H. L. J. (2015) Tracing the Development of Typewriting Skills in an Adaptive E-Learning Environment. *Perceptual & Motor Skills: Learning & Memory*, 121, 3, 1-19.

Abstract

Typewriting studies which compare novice and expert typists have suggested that highly trained typing skills involve cognitive process with an inner and outer loop, which regulate keystrokes and words, respectively. The present study investigates these loops longitudinally, using multi-level modeling of 1,091,707 keystroke latencies from 62 children (Mean age = 12.6 year) following an online typing course. Using finger movement repetition as indicator of the inner loop and words typed as indicator of the outer loop, practicing keystroke latencies resulted in different developmental curves for each loop. Moreover, based on plateaus in the developmental curves, the inner loop seemed to require less practice to develop than the outer loop.

5.1 Introduction

In order to become a skilled typist, one must master a wide variety of motor and cognitive processes, ranging from hand and finger movements to language generation and comprehension (Shaffer, 1976; Rumelhart & Norman, 1982; Salthouse, 1986; Rumelhart & Norman, 1982; John, 1996; Wu & Liu, 2008). Older typing studies primarily focused on developing motor skills. For instance, Swift (1904) measured typewriting skills as number of words typed per hour, and Lashley (1951) focused on optimizing successive keystrokes as a function of speed and accuracy. More recent typing development studies also account for cognitive skill. Logan and Crump (2011) made an explicit distinction between motor-oriented skills and cognitive-based skills, labeled as the inner and outer loops, respectively.

The inner and outer loops are nested feedback loops that serve distinct purposes. The inner loop monitors the immediate goals; e.g., press key T by moving one finger, then press H by moving a second finger, and finally press E with another finger. The outer loop monitors the broader, semantic goals; e.g., what word or sentence is to be typed next.

This two-loop theory for typewriting is supported by several experiments (e.g., Logan & Zbrodoff, 1998; Logan, 2003; Crump & Logan, 2010a, 2010b). For example, Logan and Zbrodoff (1998) showed with a typewritten Stroop task that congruency of the color and the word to be typed affected response times but not the inter-keystroke interval. Hence, the meaning of the word to be typed (outer loop) is influenced by the congruency, but not by the execution of the keystrokes within a word (inner loop). In another study, Logan and Crump (2009) limited the characters to be typed to those that should be typed with one of the hands. This restriction on the inner loop resulted in an increase in errors and decrease in speed. Furthermore, they concluded that the inner loop is largely an unconscious process. For a comprehensive overview of most experiments on the inner and outer loops, see Logan and Crump (2011).

Many studies have provided support for the inner and outer loops by using an expert-novice paradigm; aspects of expert typewriting are compared to those of novices. However, motor and cognitive processes of novice and expert typists are likely to be qualitatively different, as novice typists have developed neither the inner loop nor the outer loop. Furthermore, such a cross-sectional design is of limited use for studying how typewriting develops, or more specifically how the inner and outer loops develop. Instead, a longitudinal approach is more suitable for showing developmental trends of both loops, as the same persons are measured on multiple occasions and hence within-participants differences are also assessed.

Novice typists have to acquire the more motor-oriented inner loop as well as the more cognitive-oriented outer loop. The development of the two loops can only be indirectly inferred from differences in latencies between keystrokes. Novices do not know the layout of the keyboard yet. Hence, they have to search the keyboard for each individual character. If they have to type the same character consecutively, the latency will be smaller, as this character will be typed with the same finger. The difference between the latencies of keystrokes with and without finger repetition can be interpreted as knowledge of the keyboard. Hence, the development of the inner loop (or aspects of it) can be inferred from differences in these keystroke latencies.

The outer loop relates to the meaning of the word or sentence to be typed. Outer loop development can be inferred from differences in inter-keystrokes latencies between typing words and non-words. Indeed, empirical studies show that expert typists with a developed outer loop type words faster than non-words, while novice typists type words and non-words at the same pace (Fendrick, 1937; Shaffer & Hardwick, 1968; Gentner, Larochelle, & Grudin, 1988).

All studies thus far have aggregated latencies beyond individual keystrokes, for instance by measuring the average number of words typed in a certain time interval, the time used for typing words or non-words, or the time used to type a specific number of characters. Such aggregation does not account for parts of the observed variance and thus may considerably bias conclusions (Burstein, 1980). Therefore, this study takes a more statistically sound approach and uses a hierarchical linear model to distinguish within- and between-participants variance components. This is required to test the development of the inner and outer loops (compare with Schwartz & Stone, 1998).

It has been shown consistently that typewriting speed increases with practice (e.g., Hill, Rejall, & Thorndike, 1913) and that typing requires at least two different feedback loops. However, the way in which these loops develop has not been investigated, despite the implications for acquiring typewriting skills. Therefore, the present study traces inner and outer loop development through keystroke latencies in novice typists and assesses the contribution of both loops to overall typing speed development. Additionally, the study will investigate how much practice is necessary to reach (at least temporarily) a plateau in the development of each loop (compare with Buitrago, Schulz, Dichgans, & Luft, 2004; Maniar et al., 2005) and, hence, which loop requires the least practice.

5.2 Method

Participants

Participants were selected for the study from an online typing course (Type Garden; N = 1,226). To ensure that all participants had sufficient practice, only those who had made more than 10,000 keystrokes during the course were selected. Further, to ensure that all participants had progressed sufficiently, only those who scored at the end of the course within the top 25% of all participants in terms of speed and accuracy were selected. This combination of criteria resulted in a selection of 24 boys and 38 girls with a mean age of 12.6 year. (SD=1.6), who in total had 1,091,707 keystroke latencies, with an average of 17,608 keystrokes per participant (range: 10,067-19,999). Only correct keystrokes that were also preceded by a correct keystroke were taken into account. This prevented confounding effects like post-error slowing. The participants agreed with the use of the anonymous data for scientific research when they took a subscription to the Type Garden system.

Measures

Data for this research were obtained through Type Garden, an adaptive e-learning environment that teaches children to touch type. It has eight levels in which keys are introduced progressively. This study used only the first level, which has 270 non-words and 80 words that use the eight keys of the central row of a QWERTY-keyboard (asdfjkl;). This is the simplest level since the fingers do not have to travel over the keyboard. Each item is a letter string consisting of one or more words or non-words shown on screen, and feedback is given by highlighting each letter as it is typed (green for correct, red for error). Each item is to be completed within 20 sec. An item's score depends on both the speed and accuracy of the response. Though the scoring rule has not been evaluated for Type Garden, a similar program called Math Garden has shown it to have excellent psychometric properties (Maris & van der Maas, 2012). Students progress at their own pace, as Type Garden is a computer adaptive program where the difficulty of the next letter string (item) is matched to the participant's current ability (Klinkenberg et al., 2011). Hence, a novice typist will receive mainly easy items, while an expert receives mainly difficult items, which causes participants to practice with items that differ (in the frequency that they are presented). Therefore, two random participants might have the same number of errors, but not the same typing skills. Once a certain level is reached, a student has the opportunity to proceed to the next level. Hence, typewriting skills do not have to be fully developed for a student to progress, as the typewriting skills can still be improved in the next level. The students practiced typing at school, but also had the option of practicing individually at home. Frequent practice was rewarded with digital coins, and 60% of the selected students practiced every other day.

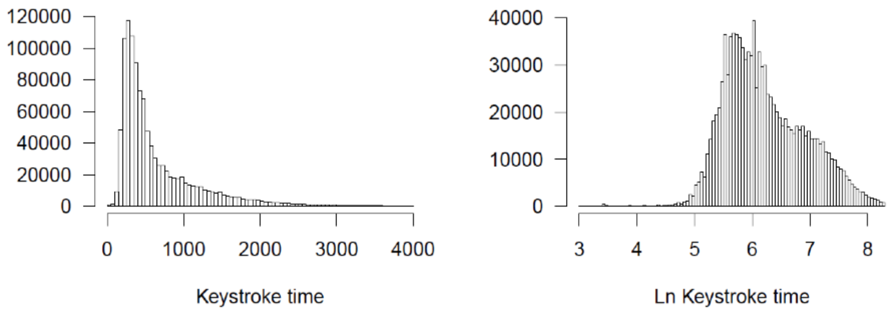


Figure 5.1: The distribution of the reaction times of keystrokes

Analysis

The times between individual keystrokes varied greatly, ranging from 21 to 3,999 msec. (keystrokes outside this range were regarded as outliers and have been removed), with a mean of 656.68 msec. (SD = 559.22). As the data were positively skewed, a natural log transformation was applied to the keystroke latencies to normalize their distribution (see Figure 5.1).

It is well known that even a small distraction can cause the reaction time of a person's individual keystrokes to lengthen (e.g., Strayer & Johnston, 2001). A more reliable measure can be extracted if the data are grouped in fixed segments of 100 successive keystrokes. This way, the number of segments represents the amount of practice. One hundred keystrokes per second was relatively arbitrary, as a segment could also comprise, for instance, one item (a string of keystrokes) or one login session. A number of segment sizes were tested; the authors are convinced that the main results are not influenced by this choice.

The data originate from a complex sample in which observations are nested within individuals. Therefore, a distinction can be made between the variance between participants and the variance between keystrokes within participants. The variance between participants indicates differences in the participants' average successive keystroke times per segment. The variance within participants indicates the difference between keystroke times of different keystrokes for an individual participant in a specific segment. The ratio of the within-participant and between-participant variances per segment is indicative for how well a distinction can be made between participants for a given segment. Generally, there are two indices which are sensitive for this distinction: intraclass correlation (ICC) and reliability (compare with Brennan, 2000).

The change in ICC is shown in the top panel of Figure 5.2. This figure indicates how necessary a multi-level model is (Hox, Moerbeek, & van de Schoot, 2010, p.15). The ICC changes during learning and ranges from .04 in the beginning to .12 in the middle and .04 at the end. Note that small ICC's, or small differences in ICC, can indicate large differences between different typists (compare with Snijders & Bosker, 1999), and can have a great effect on the significance of related parameters (Goldstein, 2011). The bottom panel of

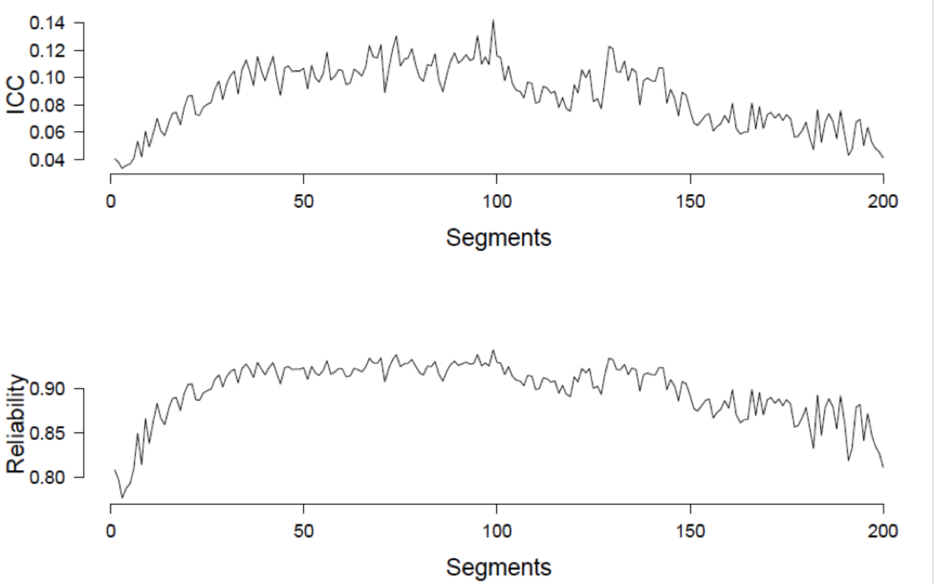


Figure 5.2: ICC and reliability over segments of 100 keystrokes

Figure 5.2 shows the reliability estimates of differences in typing speed of individuals over time. The reliability ranges from .80 to .94, and is on average .90.

Construction of Models

The development of overall typing skills was modeled by fitting subsequent polynomial functions to the data. That is, differences in individuals' keystroke latencies were modeled as a function of powers of practice (i.e., segments of 100 keystrokes). Such polynomials are very flexible functions that can take almost any shape (depending on the order of the polynomial and the value of the individual coefficients). If y_{ij} is the latency on the ith segment (the amount of practice in i times 100 keystrokes) of the jth individual, then a polynomial can be written as: $y_{ij} = function(practice_{ij})$. This function can be written as a regression model, which assumes that the latencies depend on powers of segment:

$$y_{ij} = \beta_{0j} + \beta_{1j} Seg_{ij}^{1} + \beta_{2j} Seg_{ij}^{2} + \ldots$$

As the estimated latencies will never correspond perfectly to the observed latencies, usually the difference is taken into account by an error term. In this case, however, the error term might also depend on practice (e.g., the reliability estimates in Figure 5.2. Therefore, an individual's residuals must be modeled as a function of practice.

The individual regression coefficients ($\beta_{0j}, \beta_{1j}, \beta_{2j}, \ldots$) can be written as deviations from an average of the respective parameter:

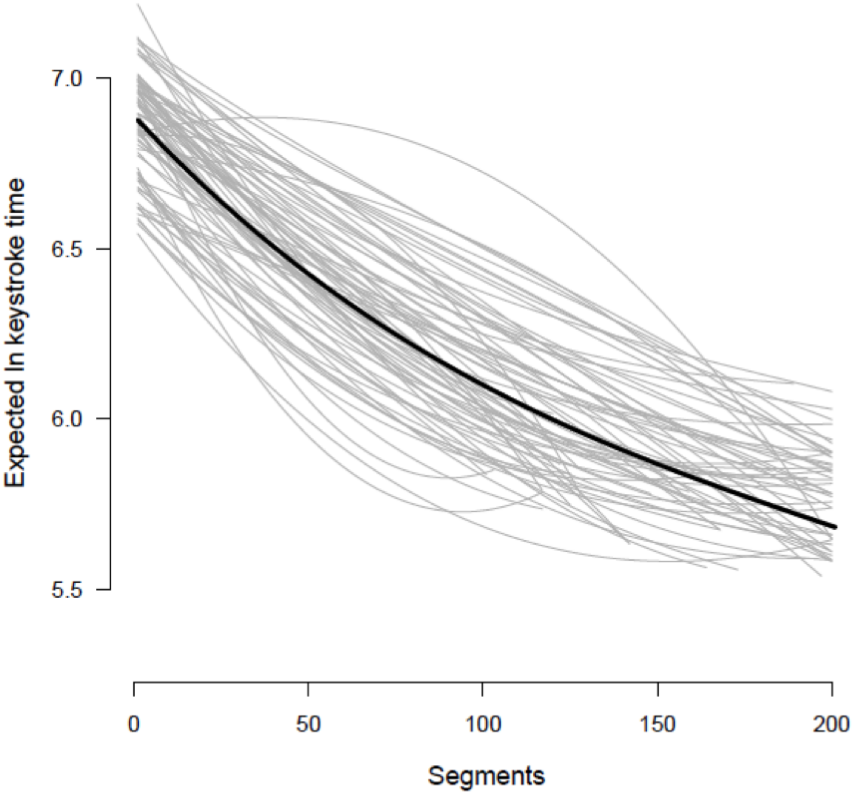


Figure 5.3: The general model of average and individual development of keystroke time over segments

$$\beta_{0j} = \beta_0 + \mu_{0j}\beta_{1j} = \beta_1 + \mu_{1j}\beta_{2j} = \beta_2 + \mu_{2j}...$$

For instance, a second order polynomial can be written as:

$$\begin{aligned} y_{ij} =& \beta_0 + \beta_1 Seg_{ij}^1 + \beta_2 Seg_{ij}^2 \\ & [(e_{0ij} + e_{1ij} Seg_{ij}^1 + e_{2ij} Seg_{ij}^2) \\ & (u_{00j} + u_{10j} Seg_{ij}^1 + u_{20j} Seg_{ij}^2)]. \end{aligned} \tag{5.1}$$

The model, as shown in Equation 5.1, consists of a fixed part and a random part (between square brackets). The fixed part estimates the average change with practice. The first fixed parameter (β_0) represents the average keystroke time for segment 0 (also known as an intercept), the second fixed parameter (β_1) represents the change in keystroke time per segment, and the third fixed parameter (β_2) indicates the extent that the change in keystroke time per segment changes per segment squared. The random part of the model distinguishes between deviations from the average for individuals ($u's$) and deviations of the observations from the individual curves ($e's$). Hence, e_{0ij} represents the deviation of the average keystroke latency of the jth individual. As the within-individuals variance might depend on practice, heteroscedasticity is modeled in terms of the polynomial. It is assumed that all residuals are normally distributed, with an expected value of 0 and a variance of $S^2_{e_{0ij}}, ..., S^2_{u_{20j}}$, respectively. Furthermore, it is assumed that the residuals within and between individuals are uncorrelated ($\rho_{e,u} = 0$). Based on Equation 5.1, the variances within and between individuals are a function of segment. The variance within participants can be approximated as:

$$\begin{aligned} Var(within|Seq = T) =& S^2_{e_{0ij}} + 2Cov(e_{0ij}, e_{0ij})T + S^2_{e_{1ij}}T^2 + \\ & 2Cov(e_{0ij}, e_{2ij})T^2 + ... + S^2_{e_{2ij}}T^4 \end{aligned} \tag{5.2}$$

The variance between individuals can be approximated in the same way (conforms with Goldstein, 2011). Hence, modeling polynomials with a multi-level model allows for accommodating heteroscedasticity of variances.

The order of the polynomial can be seen as an empirical matter. This study chose the polynomial that is most parsimonious and fits the data best according to a likelihood ratio test for subsequent analysis (conforms with van Veen, Evers-Vermeul, Sanders, & van den Bergh, 2013). The model can be extended to include variables indicative for the outer loop (words) or inner loop (finger repetition). Main effects of words or finger repetition indicate that the intercept between outer and inner loops differs from the average, whereas interactions with practice show that the development of both outer and inner loops differs from the average development.

To determine whether the inner loop develops differently from the outer loop, the inflection points (the points where the change in keystroke latencies becomes zero) will be assessed for the four possible circumstances: non-words and no finger repetition (NW-NFR), words and finger repetition (W-NFR), non-words and finger repetition (NW-FR), and words and finger repetition (W-FR). The inflection points will be determined with the first-order derivative, while the second-order derivative will indicate whether an inflection point is a minimum or a maximum. The maxima will not be of interest as they indicate the start of the development. The minima are of interest as they indicate the end of the development. As not every individual's polynomial has to have an inflection point (because not every student has to finish his development), a selection of participants with an inflection point will be made. If an inflection point of one circumstance has a lower segment number (i.e., took less practice to reach) than another circumstance, then the development of the first circumstance finished first, thereby indicating which loop finished developing first. This

Table 5.1: Likelihood ratio test of the general model

Model	No. parameters	-2LL	$\Delta\chi^2$	Δdf	p
$y_{ij} = \beta_{0ij} + [e_{0ij} + u_{00j}]$	3	2,255,410			
$+\beta_1 * Seg_{ij}^1$	4	2,017,726	237,684	1	< 0.001
$+e_{1ij} * Seg_{ij}^1$	6	2,009,725	8,001	2	< 0.001
$+u_{10j} * Seg_{ij}^1$	8	1,994,665	15,060	2	< 0.001
$+\beta_2 * Seg_{ij}^2$	9	1,986,498	8,167	1	< 0.001
$+e_{2ij} * Seg_{ij}^2$	12	1,986,081	417	3	< 0.001
$+u_{20j} * Seg_{ij}^2$	15	1,978,324	7,757	3	< 0.001
$+\beta_3 * Seg_{ij}^3$	16	1,978,144	180	1	< 0.001
$+e_{3ij} * Seg_{ij}^3$	20	1,977,901	243	4	< 0.001
$+u_{30j} * Seg_{ij}^3$	24	1,975,023	2,878	4	< 0.001
$+\beta_4 * Seg_{ij}^4$	21	1,975,021	2	1	0.842

will be done for both the average curve of each circumstance, as well as for the individual polynomials.

5.3 Results

To describe the development in latencies, several models were fitted. Both the fixed and random parts of these models increased in complexity. The fit of each model, along with the difference in fit between consecutive models, is presented in Table 5.1 and expressed by -2 log likelihoods. From the comparison between models it is apparent that a model with a fixed linear component, allowing for differences in keystroke latencies between segments, fitted better to the data than a model with only an intercept ($\Delta\chi^2 = 237,684; df = 1; p < .001$). Allowing the linear component to vary both within and between participants improved the fit, as can be seen in Rows 3 and 4 of Table 5.1 ($\Delta\chi^2 = 23,061; df = 4; p < .001$). The fit of consecutive models increased up to the third order polynomial. The third order term is allowed to differ within individuals and between individuals. As shown in Table 5.1, a fixed quartic term did not improve the model fit ($\Delta\chi^2 = 2.00; df = 1; p = .84$).

Hence, a third order polynomial was necessary to describe the observed average latencies over participants and the changes in variance within and between individuals. The parameter estimates for this model are presented in Table 5.2. As the change per segment directly depends on the scale of the segment variable, this has been centered and runs in 200 steps from -10 to 10. The first column of Table 5.2 shows the parameter estimates for the model of general development. It can be seen that the average time between keystrokes decreases significantly over (the recoded) segments. The average log transformed keystroke time at Segment 100 (keystroke 9,900 until keystroke 10,000; the intercept) is estimated as 6.09 (441 msec.), and changes continuously by -0.057 per segment. As segments have been

Table 5.2: Fixed and random parameters of models to describe the learning of touch typing (for convenience Segment has been recoded from -10 to 10).

	General Development		Word & Finger Repetition	
Fixed Part	Est.	SE	Est.	SE
Seg^0	6.088	0.015	6.152	0.016
Seg^1	-0.057	0.003	-0.056	0.003
$Seg^2 * 10^3$	0.504	0.259	0.141	0.582
$Seg^3 * 10^3$	-0.178	0.035	-0.324	0.058
$Word$			-0.120	0.002
$W * Seg^1 * 10^3$			-0.328	0.546
$W * Seg^2 * 10^3$			0.646	0.046
$W * Seg^3 * 10^3$			-0.071	0.009
FR			-0.472	0.003
$FR * Seg^1$			0.031	0.001
$FR * Seg^2 * 10^3$			-1.635	0.055
$FR * Seg3 * 10^3$			-0.074	0.011
$W * FR$			0.139	0.004
$W * FR * Seg^1$			0.002	0.001
Random Part				
Variance within individuals:				
	Est.		Est.	
$S^2_{e_{0ij}}$	0.342		0.321	
$S^2_{e_{1ij}}$	0.001		0.001	
$S^2_{e_{2ij}} * 10^3$	0.013		0.003	
$S^2_{e_{3ij}} * 10^3$	< 0.001		< 0.001	
Variance between individuals:				
$S^2_{u_{0ij}}$	0.047		0.045	
$S^2_{u_{1ij}}$	0.001		0.001	
$S^2_{u_{2ij}} * 10^3$	0.024		0.032	
$S^2_{u_{3ij}} * 10^3$	< 0.001		< 0.001	

recoded from -10 to 10 in steps of 0.1, this amounts to a change of $-0.057 * 0.1 = -0.0057$ per segment. Simultaneously, there is also an increase with $0.504 * 10^{-3}$ per squared segment and a decrease with $-0.178 * 10^{-3}$ for the cubed segment.[1] Therefore, the expected log transformed successive keystroke time in the first observed segment (with a recoded value of -10) is estimated as 6.786 (885 msec.), while the expected log transformed successive keystroke time for the final segment (10) is estimated as 5.390 (219 msec.). Hence, the average difference in keystroke times between the first and last segment is 666 msec. In Figure 5.3, the average key-stroke time is presented by means of a black solid line.

In this general model, both the fixed parameters and the random parameters have been estimated. The random parameters show differences between participants. The variance of

[1] This shows one of the reasons for recoding the segments. Without recoding, the quadratic and cubic parameters would have been even smaller.

differences between individuals at the intercept, for instance, is estimated as 0.05 (see Table 5.2). Hence, an 80% confidence interval for differences between individuals of segment 0 (i.e., the 100th segment from the start) ranges from 5.73 to 6.37. And the average linear change per segment equals -0.06, but this change differs between participants (variance = 0.001). Thus, an 80% confidence interval of the differences between participants for the linear change with segment ranges from -0.10 to -0.02. That is, for some participants the linear decrease in keystroke time is steeper than for others. The same holds for the quadratic coefficient ($80\%CI = -4.11 * 10^{-3}, 5.12 * 10^{-3}$) and the cubic coefficient ($80\%CI = -0.73 * 10^{-3}, 0.38 * 10^{-3}$).

These random terms can be used to approximate the variance in each segment (Equation 2) as well as the variance within and between participants. For instance, at the intercept the variance within individuals is estimated as 0.34. For the first and last segments, the variance within individuals is approximated as 0.88 and 0.55, respectively. Thus, the difference between keystrokes within participants clearly decreases with practice.

The variance between participants at the intercept is estimated as 0.05. The variance between participants increases significantly with practice; at the first segment this variance is estimated as 0.12, whereas at the last segment the estimate is 0.90. In Figure 5.3, the grey lines represent the estimated polynomials for the individual participants. The average change in keystroke time (on the log scale) is presented by a black solid line.

Words and Finger Repetition

In the next analysis, effects of (non-)words and (no)finger repetition on development of typing skills were compared. A likelihood ratio test showed that the model with a three-way interaction between the linear term, finger repetition, and words provides the best fit at an α level of .05 (Table 5.3).

The estimates of the coefficients of the final model are displayed in Table 5.2. The main effects of words and finger repetition were significant, as well as the interaction between words and finger repetition and the interaction between words, finger repetition, and segment. Hence, the change in latencies with practice when typing words differed from the latencies when typing non-words. For instance, at the intercept words were typed faster than non-words (-0.12). At the end of the study, the average latency in keystroke time in milliseconds for words was 5.152, whereas the average latency for non-words was 5.282.

At the intercept, the difference in average latency due to finger repetition was -0.47, indicating that keystrokes involving finger repetition were faster. At the end of the study, the average latency for items with finger repetition was 4.590, whereas the average for items without finger repetition equaled 5.282.[2] Both the effects for words and finger repetition

[2] Note that this number is the same as the previously mentioned average latency for non-words, because these non-word latencies were estimated for when there was no finger repetition. Hence, 5.282 reflected the average latency for non-words with no finger repetition.

Table 5.3: Likelihood ratio test of the word and finger repetition (FR) effect.

Model	No. parameters	-2LL	$\Delta\chi^2$	Δdf	p
General model	20	1975021			
$+Words$	21	1972762	2215	1	< 0.001
$+Words * Seg^1_{ij}$	22	1972399	400	1	< 0.001
$+Words * Seg^2_{ij}$	23	1972174	273	1	< 0.001
$+Words * Seg^3_{ij}$	24	1972037	118	1	< 0.001
$+FR$	25	1879242	92701	1	< 0.001
$+FR * Seg^1_{ij}$	26	1861802	17473	1	< 0.001
$+FR * Seg^2_{ij}$	27	1860730	1075	1	< 0.001
$+FR * Seg^3_{ij}$	28	1860663	64	1	< 0.001
$+FR * Words$	29	1859542	1129	1	< 0.001
$+FR * Words * Seg^1_{ij}$	30	1859537	4	1	0.040
$+FR * Words * Seg^2_{ij}$	31	1859537	1	1	0.307

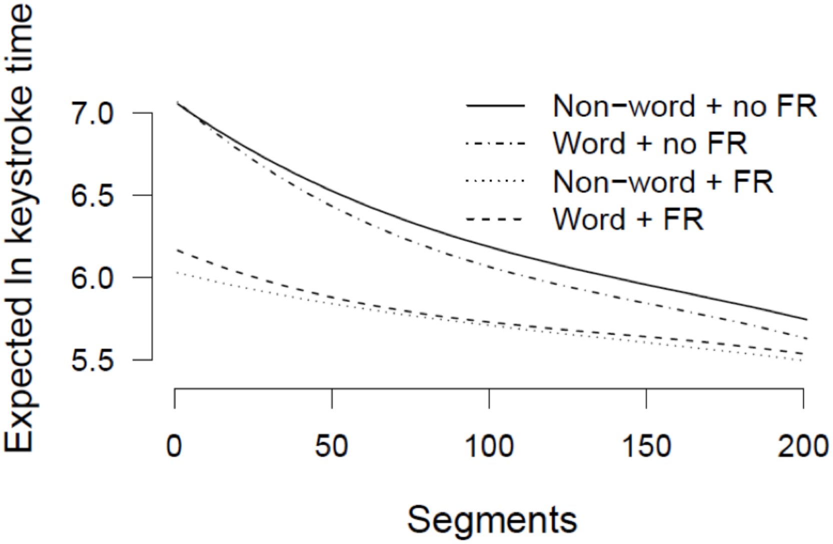


Figure 5.4: Estimations of the combinations of word and finger repetition (FR) effects

changed with practice. Figure 5.4 shows for all four combinations of words and finger repetition the average change with practice. Figure 5.5 shows that without finger repetition the estimated average (ln) keystroke latency initially hardly differed between non-words and words. However, with practice a difference emerged: without finger repetition, words were typed faster than non-words. When finger repetition was present, there was initially a difference between non-words and words, with non-words showing smaller latencies than words. This was probably due to the presence of a set of very easy items with

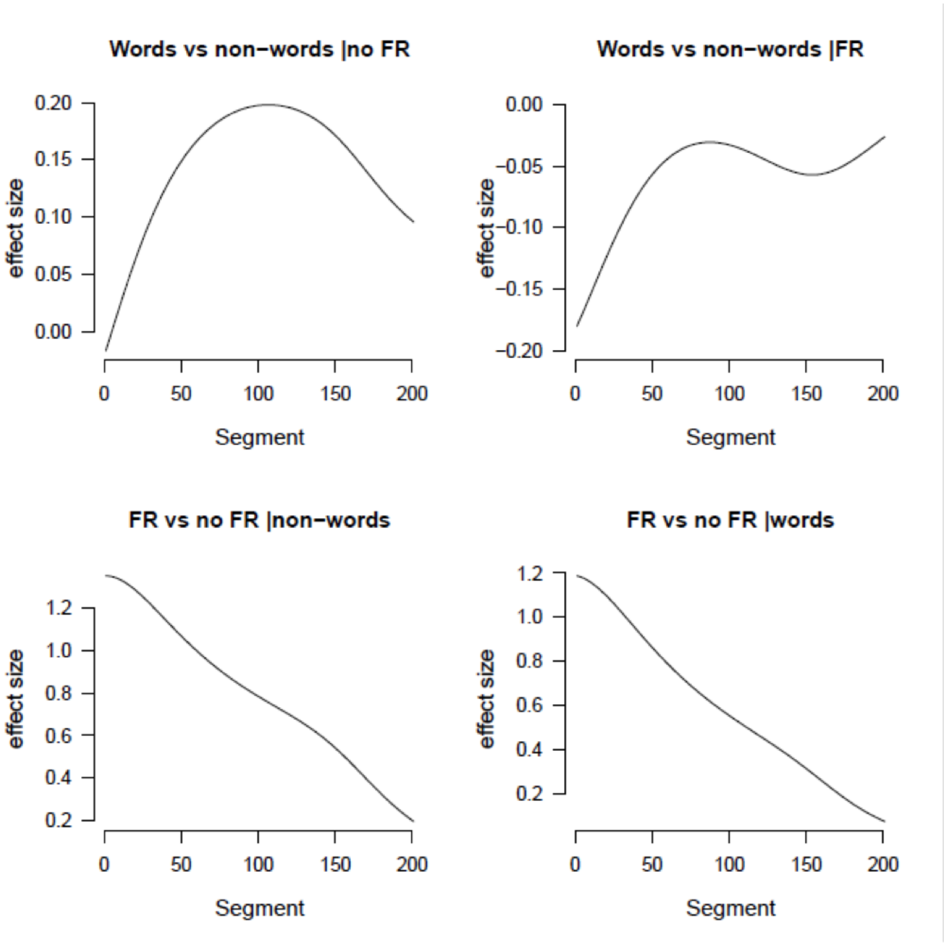


Figure 5.5: Effect sizes of finger repetition and words (Cohen's *d*)

many repetitions (such as fff, aass or aaasssdddfff). The effect of finger repetition existed, since items with finger repetition were typed faster than items without finger repetition. However, this difference between words and non-words was reduced with practice.

Because the sample of this study was large, the power of the significance tests was large. Hence, it is possible that some significant results were only due to very small differences in latencies. Effect sizes show whether the assessed differences were substantial or negligible. Cohen's d (Cohen, 1988) was computed for each segment.[3] The initial effect size for the typing of words without finger repetition equaled -0.02 and increased to a maximum of 0.20 at the 112th segment. Thereafter, the effect size decreased slightly to 0.15 at the final segment (see the top left part of Figure 5.5). This indicates that after some practice words

[3] Effect sizes over 0.8 are considered large, over 0.5 as medium, and over 0.2 as small

Table 5.4: Total number of participants with a (at least temporarily) finished development in the different conditions and the number of participants requiring more and less practice in the different conditions to reach a (temporary) plateau (less above and more below the diagonal).

Condition	NW-NFR	W-NFR	NW-FR	W-FR	Total
NW-NFR		1 (4)	5 (5)	5 (5)	5
W-NFR	2 (4)		4 (4)	4 (4)	4
NW-FR	0 (5)	0 (4)		12 (27)	34
W-FR	0 (5)	0 (4)	14 (27)		31

were typed faster than non-words, but this effect was small at best.

The total effect size for the word effect when finger repetition was present (see the top right part Figure 5.5) had an initial value of -0.19, indicating that in the beginning words were typed slower than non-words. It increased slightly, but can be considered as very small. The initial effect size for finger repetition when non-words were typed was 1.43. This decreased quickly and had a value of 0.32 at the final segment (see the bottom left part of Figure 5.5). The effect size for finger repetition when words were typed showed the same pattern of almost linear decrease, but with smaller absolute values. It started at 1.26 and ended at 0.12 (see the bottom right part of Figure 5.5).

Inflection Points

To assess the moment at which development (at least temporarily) came to a halt, the inflection points of the curves were determined. Such inflection points indicate the segment, or the amount of practice, at which no change in keystroke latencies are expected and a minimum occurs. The average development, as shown in Figure 5.3 and Figure 5.4 shows a continuous decrease in inter-keystroke latencies. Hence, the average development does not show any inflection point; on average the children had not finished developing their skills for typing words and non-words, with and without finger repetition.

Although the average change over time does not show an inflection point, this does not necessarily hold for every individual curve. Comparison of how much practice produces an inflection point in each circumstance (NW-NFR, W-NFR, NW-FR, and W-FR) for every individual allows determination of which loop developed faster. Note that a comparison can only be made if an individual's data actually has an inflection point in both circumstances. If there is no inflection point in a certain circumstance, the development for that circumstance has not been (at least temporarily) finished.

In Table 5.4, the last column shows the total number of inflection points in each of the circumstances. For NW-NFR there were five participants who showed an inflection point, while there were four participants with an inflection point for W-NFR. For the circumstances with finger repetition, NW-FR and W-FR, there were 34 and 31 participants with an inflection point, respectively. There were more participants with an inflection point in the circumstances with finger repetition than those without. This shows that the required practice for development of finger repetition was less than the development of no finger repetition. The effect of words was less obvious. The number of participants with an inflection point with non-words was consistently higher than with words, but the differences were small.

Furthermore, Table 5.4 also shows whether the (temporary) plateau in one of the circumstances preceded that of another circumstance (above the diagonal) or whether the (temporary) plateau occurred later than in another circumstance (below the diagonal). For instance, the first row shows that for one out of four cases the development in the NW-NFR circumstance required less practice than in the word combined with no finger repetition. For five (out of five) participants the NW-FR and W-FR circumstances required less practice to reach a (temporary) plateau than NW-NFR.

The first column of Table 5.4 shows that for two of four participants W-NFR required more practice than NW-NFR. Because the first row already indicated that one participant required less practice, there is one participant who needed the same amount of practice in both circumstances. No participants (out of five) required more practice to reach a (temporary) plateau in NW-FR and W-FR compared to NW-NFR.

The last two cells of the second row and the second column indicate that all four participants with an inflection point in W-NFR and NW-FR or W-FR required less practice in the circumstances with finger repetition. The final comparison between NW-FR and W-FR in the final cell of the last row and column indicated that 12 participants (out of 27) required less practice to reach a (temporary) plateau in the W-FR circumstance, while 14 participants required more practice in the W-FR circumstance. Also, between these circumstances there is one participant who needed similar practice in both circumstances to reach a (temporary) plateau. Hence, the relationship between words (compared to non-words) and typing development does not seem to be straightforward.

5.4 Discussion

This study investigated the typewriting skill development in an adaptive learning environment. The authors analyzed data from the first game of an online course on typewriting in which the eight characters of the home row were learned. Data were collected at the keystroke level; therefore, the number of observations was enormous. Since keystroke latencies were nested within learners, development was modeled using a multi-level approach. This multi-level model was relatively parsimonious, with four fixed parameters for the amount of practice and seven random parameters for the differences between and within individuals. It was shown that the average keystroke latency decreased with practice, but that the learning curves of each individual differed notably. Based on analyzing the inflection points in each individual's polynomial, not all participants had reached their minimal keystroke latencies yet. Hence, their development in this task was not yet completed. This is not surprising, as after reaching a certain level the second game becomes accessible and it is an individual's own choice to continue the first game or to start playing the second game with more letters of the keyboard.

Both the inner loop, indicated by the decreasing finger repetition effect, and the outer loop, indicated by the word effect, appeared to emerge with practice (this conforms with Logan & Crump, 2011). The loops developed differently, and both effects contributed significantly to the model of overall typing development. In general there was no plateau in development (a vanishing rate of change in keystroke latencies), but these plateaus were found in some individual developmental curves. Comparing the amount of practice needed to finish development between the different circumstances of (non-)words and (no) finger repetition for each individual indicated that many more individuals finished their development in the circumstances with finger repetition than in the circumstances without. The results for the (non-)words are not so clear, since some individuals finished developing faster with words while others finished developing faster with non-words. This also made a comparison between the words and finger repetition more difficult, but the strong effect of finger repetition on the development compared to the ambiguous effect of words indicated that the development of the inner loop is finished before the development of the outer loop. This is in concordance with previous findings that while the associations between keys and finger movements are helpful for basic typing, associations between words and letters are required for skilled typing (Yamaguchi & Logan, 2014).

In the present study, the development of average keystroke latencies was analyzed per 100 keystrokes. Such an analysis neglects the natural boundaries between items, which were words or non-words with different number of characters. The proposed model can be expanded to a so called cross-classified model (Goldstein, 2011) in which both the variance between participants and the variance between items are estimated simultaneously. This allows for a more precise analysis of item characteristics. Alas, this was not possible in the present study, as the adaptive nature of the Type Garden allocated the demanded items to the ability of the participant.

Another consequence of the allocation of items was that participants with the same amount of practice did not receive the same items. However, because the presented items depended on the ability of the participants, scaffolding took place for the development of typewriting. Hence, the results should be seen as generalizable for this type of learning. The results of this paper showed how the finger repetition effect disappears and the word effect emerges, indicating the development of the inner and outer loops. The development of the inner loop seems to be finished before the outer loop, as the word effect emerges before the development of the finger repetition is finished. This suggests that the development of the inner and outer loops occur separately.

6 The Dynamics of the Development of Mathematical Ability: A Comparison of *g*-Factor and Mutualistic Network Theories.

This chapter is submitted as: Hofman, A. D., Kievit, R. A., Stevenson, C. E., Molenaar, D., Visser, I., & van der Maas, H. L. J. (2018). The dynamics of the development of mathematics skills: A comparison of theories of developing intelligence. *Developmental Science*; Preprint at: `osf.io/xa2ft`

Abstract

A famous and well replicated finding in psychology, and mathematics as this paper expresses, is that individual differences across a large range of cognitive skills are all positively correlated. In the current study we compared two influential explanations for this so-called positive manifold of correlations: *g*-factor and mutualism theory. We examined a large longitudinal data set ($N \approx 12.000$) that tracked the development of four basic math skills (counting, addition, multiplication and division) for a full school year. We used bivariate latent change score models to investigate whether the *g*-factor or mutualism theory provided a better explanation of the developmental pattern of correlations across math abilities. We found that the correlations between abilities increased during development and that bidirectional mutually beneficial relations occurred over time. Both results support mutualism theory, a dynamic network perspective on the development of cognitive abilities, where, in this case, growth in a particular math subdomain positively influences that of other subdomains. Our results suggest that mutualistic mechanisms may operate not just between cognitive domains, but also within domains. We discuss implications of mutualism theory for understanding the dynamics of learning mathematics.

6.1 Introduction

The ability to do math is essential to daily life and the study of how this develops is of great interest for the fields of education and cognitive development (Siegler & Lortie-Forgues, 2014). Individual differences in mathematical ability relate to a wide set of cognitive abilities (Murnane, John, & Levy, 1995) and lifespan outcomes (Siegler et al., 2012). Understanding how different mathematical abilities co-develop is crucial for understanding learning in general, and mathematics in particular. A large body of studies have focused on the symbiotic development of different cognitive and math skills, focusing particularly on the positive correlations between domains (e.g., Geary, Hoard, Byrd-Craven, & DeSoto, 2004; Halberda, Mazzocco, & Feigenson, 2008).

However, these cross-sectional studies can only provide *indications* of the processes that occur within individuals over time (Tucker-Drob, 2009), and could possibly be misleading (Wohlwill, 1973; P. C. Molenaar, 2004). More direct tests of how the development of different skills unfolds can be better established with a longitudinal approach, see for example the work of Geary, Brown, and Samaranayake (1991) and van der Ven, Kroesbergen, et al. (2012). Currently, the dynamic processes that drive the development of mathematical abilities are unclear, as are possible direct links with the development of other skills. In this paper, we borrow ideas from the expertise in intelligence research to investigate this issue.

The empirical finding of a positive correlational structure of individual differences in abilities across domains abilities is not unique for the field of mathematics. In the field of intelligence this is one of the most famous findings in cognitive psychology, and called the positive manifold (Spearman, 1927). This implies that people who perform well on one cognitive task also tend to score well on other cognitive tasks. In intelligence data, the correlational structure is often explained by *g*, for 'general intelligence', such that the positive correlational structure between the different cognitive test scores is modelled by a latent factor that represents general abilities across domains. Although *g* is a useful construct in predicting educational success and other life outcomes, the presence of such a statistical factor does not necessarily imply a causal role for a single underlying factor across different cognitive tests (van der Maas et al., 2006b; van der Maas, Kan, Hofman, & Raijmakers, 2014; van der Maas, Kan, Marsman, & Stevenson, 2017; Kruis & Maris, 2016). Nor does it imply that *g* necessarily has an ontological status beyond a statistical entity (Borsboom, Mellenbergh, & van Heerden, 2003). This follows from the fact that different hypotheses about the nature of the positive manifold can result in the same cross-sectional correlational data (Bartholomew, Deary, & Lawn, 2009; Anderson, 2017). In order to differentiate between different explanations of the positive manifold, hence to shed light on its true nature, a longitudinal approach is valuable (van der Maas et al., 2006b; P. C. Molenaar, 2004).

In the current paper, following the approach of (Kievit, Lindenberger, et al., 2017), we compare two important theories of the positive manifold based on their main distinguishing feature – the dynamics between different abilities over time. Using this approach, we investigate whether a mathematical *g*-factor (a general math ability) plays a causal role during development or whether bidirectional mutually beneficial relations between domains are sufficient. Therefore we use a large longitudinal data set collected with an online learning program for mathematical ability in schoolchildren (Math Garden; Klinkenberg et al., 2011; Straatemeier, 2014). In this learning program children can log in at any time and play as much items as they want for one or multiple games. The growing popularity of online learning systems and an increasing accessibility to the internet provides large amounts of data on learning. Additionally, this data is collected in a school setting and captures the processes involved in this natural learning setting. This data set provides a unique sample to investigate the developmental patterns involved in learning mathematics.

Few studies have approached the study of the nature of the correlational structure between different skills from the perspective of the positive manifold and compared competing models of development. Even those studies that have focused on longitudinal co-development of cognitive abilities have focused on development *between* domains such as memory, reasoning, vocabulary and perceptual speed (e.g., McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; Kievit, Lindenberger, et al., 2017). No studies to date have examined whether a *g*-factor or interactive account better explains fine-grained development *within* a domain.

Two different theories of cognitive development

The robust finding of the positive manifold in scores on intelligence and mathematical abilities requires an explanation. The most prominent explanation is the *g*-factor theory itself (Jensen, 1998). This theory suggests that the latent factor (for general intelligence or in our case general mathematical ability) is of a natural kind and directly affects the scores on different domains that reflect, to varying extents, the general domain. This *g* is the real variable of interest, an entity that exists independent of the data.

Although the standard *g*-factor theory does not include a developmental perspective (van der Maas et al., 2006b), a natural developmental interpretation of *g* suggests that changes in *g* induce changes in different subdomains. Thus changes in different skills, that function as indicators of *g*, should be caused by developmental processes in *g*.[1] Therefore, it is often assumed that the correlational structure – reflecting the strength of the general factor – is relatively stable over time (Jensen, 1998; Gignac, 2014). Empirical results on these differentiation effects are mixed (D. Molenaar, Dolan, Wicherts, & van der Maas, 2010). Gignac (2014) found that the strength of *g* was relatively stable between ages 2.5 and 90. Whereas McArdle et al. (2002) concluded that the development of a single *g*-factor provides an overly simplistic view, which was based on the analyses of learning curves of 5 to 90 year-olds.

In the field of mathematics such a general latent ability that causes the performance on multiple related mathematical test is often (implicity) assumed. For example, when higher order latent variables are introduced to model individual differences (e.g., Aunola et al., 2004; Muthén, Kao, & Burstein, 1991).

A second, more recent, theory was proposed by van der Maas et al. (2006b) as an alternative explanation for the positive manifold. In mutualism theory, development is seen as a complex system of (positively) interacting processes, where learning one process (skill) supports learning of the other processes in the system. The proposed mutualism model of general intelligence explains the positive manifold by positing mutually beneficial relations between the different abilities during development. The strength of these mutual relations between abilities are captured in the **M** matrix in the mutualism model and can be both unidirectional – as for example in investment theory where growth in crystallized intelligence is fueled by growth in fluid intelligence Cattell (1971) – or bidirectional. These bidirectional facilitating relations have previously been observed between the development of cognitive strategies and short-term memory (Siegler & Alibali, 2005), vocabulary and reading ability (Quinn, Wagner, Petscher, & Lopez, 2015) and between subjective and objective memory in aging (Snitz et al., 2015). In mathematics, van der Ven, Kroesbergen, et al. (2012) found a positive interactions between changes in math skills and changes in working memory, suggesting mutual influences in their development.

[1] A strict interpretation of the *g*-factor model would even predict that development in a lower order factor would not result in any changes in other lower order factors, and would merely result in a larger residual variance (van der Maas, Kan, & Borsboom, 2014).

Contrary to the *g*-factor theory, the mutualism theory predicts a strengthening of the correlational structure between performance on different cognitive tasks during development. Mutualism theory posits that all cognitive processes could initially be unrelated and then become related due to the positive interactions during development.

In this paper we compare both theoretical accounts of the positive manifold with respect to developing mathematical abilities. We investigate the relation between the development of *simple* (counting and addition) and more *advanced* (multiplication and division) mathematical abilities. For both sets of abilities, we examine two competing predictions that follow from each of the two developmental accounts: the cross-sectional correlational structure between domains and longitudinal coupling between domains over time.

First, we tested whether the *g*-factor or the mutualism theory best describes the development of the correlational structure between different math skills. Based on the idea of Gignac (2014), we tracked the development of the correlations between skills over time. In this bivariate approach the straightforward correlation coefficient contains all the necessary information to test whether the strength of the positive manifold is indeed stable or increases during development. Second, using latent change score models[2] (LCSM; McArdle, 2001, 2009; Ferrer & McArdle, 2010; Kievit, Brandmaier, et al., 2017), we compared a set of models associated with the *g*-factor and mutualism accounts of the positive manifold in performance on different mathematics tests. To this end we analyzed data of a large group of children that frequently played different games in Math Garden.

To summarize, the goals of the current study are to examine whether *g*-theory or mutualism theory best explains the positive manifold in mathematical abilities. Mutualism theory predicts (strong) bidirectional relationships between different abilities during development whereas g-theory does not. Testing this prediction requires translating it into a testable statistical hypothesis that can be applied to longitudinal data. The next section provides these methods, as well as a description of the data.

6.2 Method

Modeling Framework: Latent change score models

Latent change score models are structural equation models where the (latent) variables of interest are represented as change scores between time-points. The score on variable *y* of person *p* at time-point *t* is regressed onto the score of *t-1* as follows:

$$y_{pt} = \beta_{t,t-1} y_{pt-1} + \Delta_{pt}.$$

Setting the regression coefficient $\beta_{t,t-1} = 1$ results in a Δ_{pt} that reflects the change in the scores between both time-points:

[2] Sometimes referred to as latent difference score models.

$$\Delta_{pt} = y_{pt} - y_{pt-1}.$$

In the current paper we extend the LCSM to bivariate LCSM (McArdle, 2001) which allows us to investigate the relationship between the development of abilities in two separate domains as follows:

$$\begin{aligned}\Delta_{1,pt} &= \beta_1 y_{1,pt-1} + \gamma_{21} y_{2,pt-1}\\ \Delta_{2,pt} &= \beta_2 y_{2,pt-1} + \gamma_{12} y_{1,pt-1}\end{aligned}$$

In this model a change score is defined for both domains (1,2). We use this framework to model growth using two different components. First, β, is a self-feedback parameter that relates the change between time-points to the scores on the previous time-point. A positive β reflects accelerating growth and a negative β indicates damping, regression to the mean, or ceiling effects. Second, and most important, are the coupling parameters γ. These reflect the effect of the previous score in one domain (i.e., at t-1) on the change in score in another domain. By examining the γ_{21} and γ_{12} parameters, we can determine which domains influence the development of other domains. Both γ and β parameters predict change scores and can therefore capture nonlinear processes (Ghisletta & De Ribaupierre, 2005).

We formulated a set of LSCM's that allow a direct comparison of two developmental theories by focusing on different ways to model change scores, which capture the amount of development between different time-points. We assumed that differences between g-factor and mutualism models (both with uni- and bidirectional effects) would show up in differences in the γ parameters. This is because there is a clear relation between the γ parameters in the LCSM and the **M** matrix in the mutualism model (effect of **M** is also defined on change scores; see van der Maas et al. (2006b). Using LCSM's we formulated different models based on hypothesis of either g-factor or different mutualistic accounts on the development of mathematical abilities, see Figure 6.1. For both models we included covariates of age (at T1) and amount of practice (number of items solved before T1) to account for possible differences between children in the baseline scores (T1). We did not include any covariates on other variables in the model because the developmental processes between the waves should be solely explained by the dynamics defined in the different LCSM's.

First, for the g-factor model, the latent g-factor is the underlying mechanism that steers development (i.e., change) in all domains. Hence, the change processes are defined at this latent level (Δ_{g1} and Δ_{g2}), and the observed scores and changes in both domains are expected to be caused by changes at the higher latent level. This results in a univariate LCSM, since only a single set of change factors at the level of g is required. We imposed

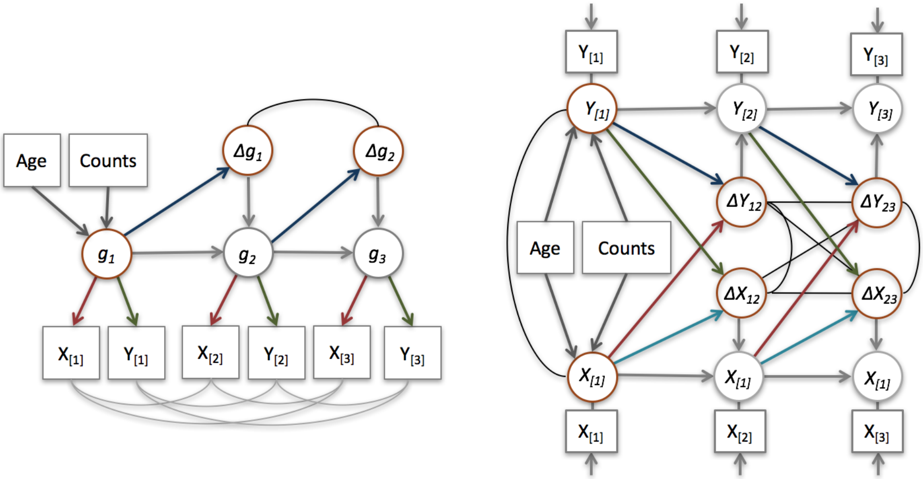


Figure 6.1: Specifications of the latent change score models of *g*-factor and mutualism accounts of cognitive development. Key parameters are color coded, and the same color indicates parameter equalities. Means and variances were estimated for all orange colored latent variables.

measurement invariance over time and allowed the observed scores within each domain to be correlated over time.

Second, for the mutualism model, we formulated a set of bivariate LCSM's with separate change scores for each domain and then successively added coupling parameters to connect the changes (development) across domains. For both the *simple* (addition and counting) and the *advanced* (multiplication and division) pairs of abilities, this resulted in four LCSM's: (1) a model without coupling parameters (see Appendix to Chapter 6 for a description of the differences between this bivariate LCSM without coupling and the univariate LCSM, the *g*-factor model), (2 & 3) two models with unidirectional coupling where only one of the two coupling parameters was present (for example an increase in counting results in growth in addition but not vice versa) or (4) a model with bidirectional coupling parameters (mutualism), where changes in each domain influences development in the other domain. The no-coupling model resembles the correlated growth model of Quinn et al. (2015) and functions as a baseline model which we test against to determine the presence of significant coupling parameters. Also, as indicated by a simulation study (see Appendix to Chapter 6), if the coupling parameters are set to zero the model fit of the no-coupling model resembles the model fit of the *g*-factor model, which further supports its use as a baseline model.

In the bivariate LCSM's the (residual) change factors (ΔX_1, ΔX_2, ΔY_1 and ΔX_2) are allowed to be correlated. These correlations imply that after possible coupling effects, the change factors of different variables and at different time-points are related. These correlations, or 'structured residuals' cf. Curran, Howard, Bainter, Lane, and McGinley (2014) reflect the effects of: (1) a possible larger set of variables that influence growth in both

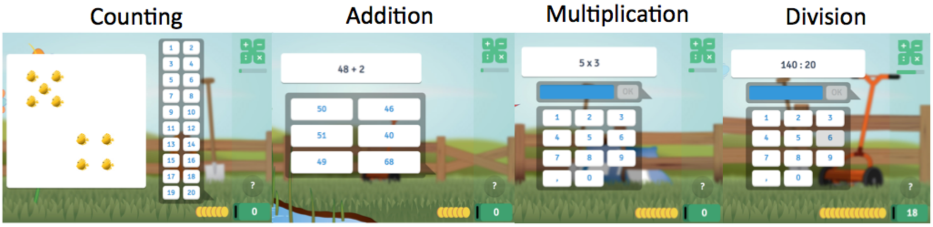


Figure 6.2: Screenshots of the counting, addition, multiplication and division games in the Math Garden online learning platform. On counting and addition tasks children click an answer option to give a response. For multiplication and division children use the numeric keypad to supply their solution. The coins at the bottom disappear gradually, one per second, and represent how the games are scored – one point for each remaining coin after supplying the correct solution.

measured domains, which are not included in the analyzed data set (e.g., working memory, processing speed, reasoning) and/or (2) a mismatch between measurement density during data collection and the time steps with which natural development unfolds (see Appendix to Chapter 6). Both of these causes could be present in our sample.

Both models are formulated based on a data set with three time-points. In constructing a data set based on data from an online learning systems there is a trade-off between either a large number of time-points with fewer children or a smaller number of time-points resulting in a large number of participants. We aimed for the latter and selected three waves. This allows us to test the competing models with sufficient precision, and even allows for extensions such as incorporating the correlations between changes and subsequent changes between and within domains, whilst retaining sufficiently large sample sizes of children who played regularly. Also, the formulation of both models including the correlational structure on the change scores is straightforward, and no higher order change-factors can or have to be specified as for example in Ferrer and McArdle (2003, 2004). Moreover, sampling three full school years allows us to study a sufficient period of learning and development.

Instruments

Data was collected using a popular Dutch online adaptive practice system for mathematics (Math Garden; Straatemeier, 2014; Klinkenberg et al., 2011). The system consists of multiple games that measure different mathematical abilities. In each game a player's ability is re-estimated after every response using a variant of the Elo-algorithm (Elo, 1978; Klinkenberg et al., 2011), which takes both accuracy and response time into account (Maris & van der Maas, 2012). For more details on the Math Garden and its psychometric properties, see Appendix to Chapter 6. In this study we focus on two *basic* skills: counting (Jansen et al., 2014) and addition, and two more *advanced* skills: multiplication (van der Ven et al., 2015; Hofman, Visser, Jansen, Marsman, & van der Maas, Submitted) and division. Each skill is measured by a separate game. Figure 6.2 shows a screen shot of an example item of each game. The estimated ability scores (comparable to factor scores in structural equation models, see Appendix to Chapter 6) were used as input variables in the LCSM's.

Children (or their parents) who indicated that they did not want to participate in scientific research conducted with Math Garden were excluded from the analyses. This research study was approved by the psychology department's Ethics Committee.

Data Selection: Tracking Development

We selected the ability estimates of children who played at least thirty items in one week between the start of the school year in 2013 and the end of the school year in 2016 (covering three full school years). For the investigation of the development of counting and addition we selected children in first grade, born in 2007 , at the start of the data collection (around nine years old at the end of data collection - end of 3rd grade). For the investigation of multiplication and division development we selected children born in 2006 who were on average eight years old at the start of the data collection (beginning of 2nd grade). By selecting children based on birth cohorts we ensured that the childrens' ages increased (linearly) over time, although children could enter and the leave the data collection at different time-points by deciding to start or quit playing a particular game. The same data selection approach was used by Brinkhuis, Bakker, and Maris (2015). We plotted the number of selected children for each week in the upper panel of Figure 6.3. School holidays are denoted by the grey areas. For the analyses on the changes in correlations between the math abilities we only included weeks in which a least 25 children played both games.

Data Selection: Latent Change Score Models

For the estimation of the LSCM's we constructed a data set with three time-points. We selected the ability estimates of children for the three separate years in the first month (2013-09-02 - 2013-10-04; 2014-09-01 - 2014-10-03; 2015-08-31 - 2015-10-02), the middle (2014-01-06 - 2014-02-07; 2015-01-05 - 2015-02-06; 2016-01-04 - 2016-02-05) and at the end of the school year (2014-05-12 - 2014-06-27; 2015-05-11 - 2015-06-26; 2016-05-09 - 2016-06-24) giving us nine ability estimates per math task. We defined these time-points based on school periods to capture our hypothesis that the amount of schooling is the dominant factor in driving developmental change. In addition, only data of children in grades one to six, between four and twelve years old[3] and who solved at least thirty items on the two domains within one or multiple periods were selected. For the analysis based on the counting and addition games we only selected children under ten years old given that the counting game is most relevant for this age group. All scores were standardized based on the mean score and *SD* at T1.

[3] We performed a reliability check on the information about age and grade, where children who deviated more than two standard deviations from the average grade per class were deleted

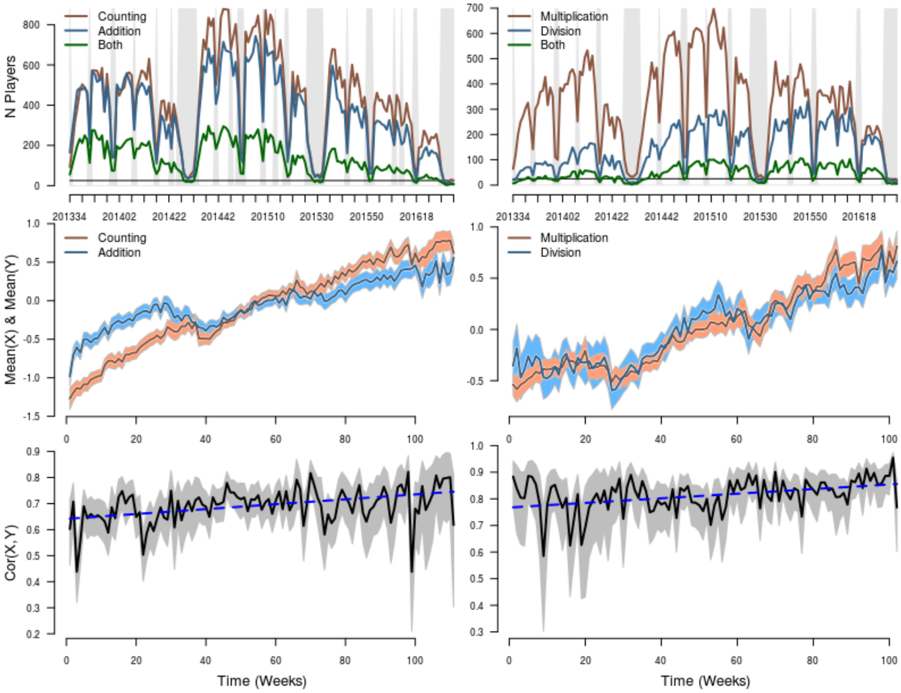


Figure 6.3: The number of selected players in our dataset for each game (upper-panels; gray areas indicate school holidays), the mean ability scores for each game (middle-panels; with the 95% confidence interval) and the correlations between scores for children in a specific birth cohort (children born in 2007 and 2006, respectively). The gray area indicates the 95% confidence interval and the dashed blue line is the regression line which indicates a significant increase in both correlations over time.

Model Estimation and Comparison

LCSM's were estimated in Lavaan (Rosseel, 2012) using Full Information Maximum Likelihood with robust standard errors to account for missings and non-normality. To assess the overall model fit we used the following tests with guidelines based on Schermelleh-Engel, Moosbrugger, and Müller (2003); chi-square test, the CFI (acceptable fit .95 – .97, good fit > 0.97), the RMSEA (acceptable fit < 0.08, good fit < 0.05), the SRMR (acceptable fit .05 – .10, good fit < 0.05). We compared the model fit using information criteria (AIC and BIC) and Akaike Weights (Wagenmakers & Farrell, 2004). These weights express the evidence for each model given the observed data and the set of candidate models.

6.3 Results

Developing Correlations

First, we investigate the development of the correlations between the two math skills in each data set. The upper-panel plots of Figure 6.3 show the number of participants for each week; the green line indicates the number of participants that played both games. The number of participants varied from 25 (minimum) to 250 at the peaks. The division game was the least popular game, with about 100 participants during peak periods. Not surprisingly, in each math domain the average rating clearly increases during the three years of data collection (middle-panel plots), with the exception of the summer holidays where a dip is observed. The lower-panel plots show the correlation between the two math domains for each week (only weeks where >25 children played both games). The grey areas indicate the 95% confidence interval and the blue dashed line shows the predicted values from a weighted linear regression model where the correlation is predicted by time (week number) using the number of observations per week as weights. For both data sets this regression coefficient was positive and significant (1 counting and addition: β = .00094, $t(107)$ = 4.779, p < .001, R^2 = .166; 2 multiplication and division: β = .00087, $t(100)$ = 4.458, p < .001, R^2 = .158)[4]. This clearly indicates that the estimated correlation coefficients increased over time.

This analysis provides indirect evidence that mutualistic interactions drive the development of these math skills, resulting in increased correlations over time. This evidence, however, is inconclusive since the change could be caused by either changing error variances in both scores or changes in the strength of the factor (Gignac, 2014; D. Molenaar et al., 2010). Yet the current result naturally follows from a mutualism perspective, while it is more difficult to reconcile from a g-factor perspective. In the next analyses we provide a more direct comparison of both theoretical accounts explaining the positive manifold, using a set of LCSM's.

Latent Change Score Models

The data selection resulted in 11,980 participants for the counting vs addition data set and 12,368 participants for the multiplication vs division data set (of which 697 and 1,054 cases respectively contained data on both variables at all three time-points). 218 participants from the counting vs addition data set and 8 participants from the multiplication vs division data set were deleted due to outliers (ability score that deviated more than three *SDs* from the mean). Table 6.1 shows the total number of participants included at each time-point, the mean and SD of the unstandardized ratings, and the correlations between the ability scores at each time-point for each data set.

[4] Since some cases might have overlapped for different data-points we used a bootstrap method and performed a permutated null-hypothesis test. Using 50.000 replications, none of the sampled t-statistics exceeded the observed t-statistics, supporting the significant results of the presented tests for both data sets

Table 6.1: Descriptives for Counting and Addition data set (top section) and Multiplication and Division data set (bottom section).

Domains	Variable	Count T1	Count T2	Count T3	Add T1	Add T1	Add T1
Counting	Mean	-5.040	-4.627	-4.482	-9.513	-8.451	-7.638
&	SD	1.353	1.325	1.375	3.410	3.389	3.328
Addition	Count T1	6279	3412	2672	4561	3926	3222
	Count T2	0.795	7580	3952	3644	5549	4505
	Count T3	0.712	0.772	6498	2838	4253	4731
	Add T1	0.782	0.706	0.654	6106	3581	2958
	Add T2	0.764	0.778	0.709	0.881	7862	4356
	Add T3	0.726	0.737	0.759	0.835	0.894	7029
		Mult T1	Mult T2	Mult T3	Div T1	Div T2	Div T3
Multipli-	Mean	-9.892	-8.370	-7.067	-12.220	-11.156	-9.966
cation	SD	3.798	4.215	4.495	5.923	6.374	6.738
&	Mult T1	7104	4583	3649	4254	4842	3986
Division	Mult T2	0.887	8996	5474	3679	6541	5969
	Mult T3	0.810	0.899	7613	2825	4999	6232
	Div T1	0.855	0.817	0.757	5325	3541	2779
	Div T2	0.817	0.869	0.830	0.897	8386	4991
	Div T3	0.773	0.827	0.873	0.834	0.917	8156

textitNote. The lower-diagonal values represent the correlations between the scores on each variable. The upper-diagonal values represent the number of participants with scores on a single (diagonal) or set of variables (off-diagonal). T = time-point, Count = Counting, Add = Addition, Mult = Multiplication, Div = Division.

Table 6.2: Fit statistics for the estimated Latent Change Score models for Counting and Addition (top section) and Multiplication and Division (bottom section).

Domains	Model	Chi	df	CFI	RMSEA	SRMR	AIC	BIC
Counting & Addition	*g*-Factor	571.66	15	0.987	0.056	0.046	142138	142352
	Mutualism	461.48	11	0.989	0.059	0.038	142036	142279
	Uni.(C → A)	475.05	12	0.989	0.057	0.039	142048	142284
	Uni.(C ← A)	470.95	12	0.989	0.057	0.038	142044	142279
	No coupling	485.24	13	0.989	0.056	0.038	142056	142284
Multiplication & Division	*g*-Factor	671.52	15	0.989	0.059	0.030	149304	149520
	Mutualism	517.17	11	0.991	0.061	0.026	149158	149403
	Uni.(M → D)	575.44	12	0.990	0.062	0.027	149214	149452
	Uni.(M ← D)	521.91	12	0.991	0.059	0.026	149161	149398
	No coupling	597.61	13	0.990	0.060	0.026	149234	149464

Note. The the best fitting models are printed in bold. Uni. = Unidirectional Model

We fitted the five different LCSM's on both data sets. Table 6.2 shows the different fit indices for each of the models. For both data sets, all models showed a good fit as indicated by the CFI, RMSEA and SRMR, with only small differences between these three fit indices. The AIC and BIC, which take both fit and model complexity into account, showed that the four bivariate LCSM's (a no-coupling, two unidirectional coupling and a mutualism model with bidirectional coupling) fit better than the *g*-factor model. Furthermore, according to the AIC the mutualism model including two sets of coupling parameters showed a better fit than both unidirectional models and the bivariate model without coupling Using AIC-weights (Wagenmakers & Farrell, 2004), we can represent the conditional probability of each model, given the observed data, within our set of candidate models.

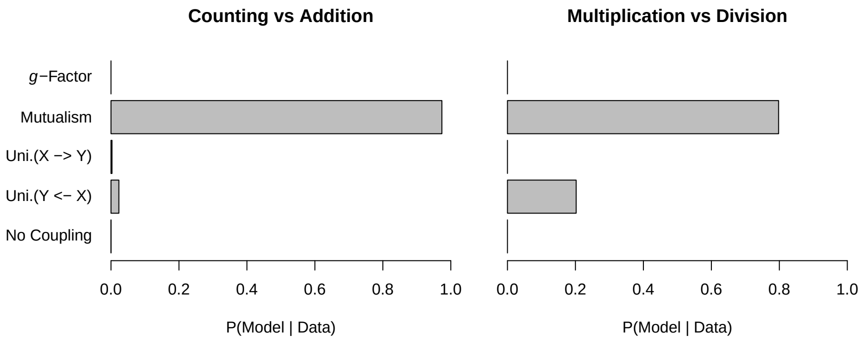


Figure 6.4: The AIC-weights of all models for both data sets, indicating the evidence for each model given the observed data and the set of candidate models.

Figure 6.4 shows these weights for each model and clearly indicates that the mutualism model outperforms all other candidate models in the counting vs addition data set. This differences is less prominent for multiplication vs division data set. For this data set the fit of the mutualism model only slightly outperforms the unidirectional model including a coupling parameter from division to multiplication. According to the BIC this unidirectional model provides a better fit, while comparison of the likelihoods indicates significantly better fit for the mutualism model ($X^2(1) = 4.746, p = .029$). Both models clearly outperform the *g*-factor account. Although the evidence of bidirectional coupling is less clear, the mutualism model provides the best description of the data.[5]

In the following section we investigate the parameter estimates of the mutualism model in more detail. Figure 6.5 shows all parameter estimates for the mutualism model fitted

[5] For the multiplication vs division data set, the results were comparable if no covariates were included in the model and the mutualism model clearly outperformed the other models. For the counting vs addition data set this was also the case when only Age was included as a covariate. Excluding Age resulted in a small variance in one of the change factors (no difference from zero). This suggested that no individual differences were present in this change-score. This caused problems for the estimation of the covariance matrix of the change scores, resulting in unreliable model comparison.

to the counting and addition data set. Since the model parameters for the counting and addition data set were very similar to those of the multiplication and division data set, we first describe the general model details and then go on to describe the main differences between the results for the two data sets (see Appendix to Chapter 6 for all model parameters of the multiplication and division data set).

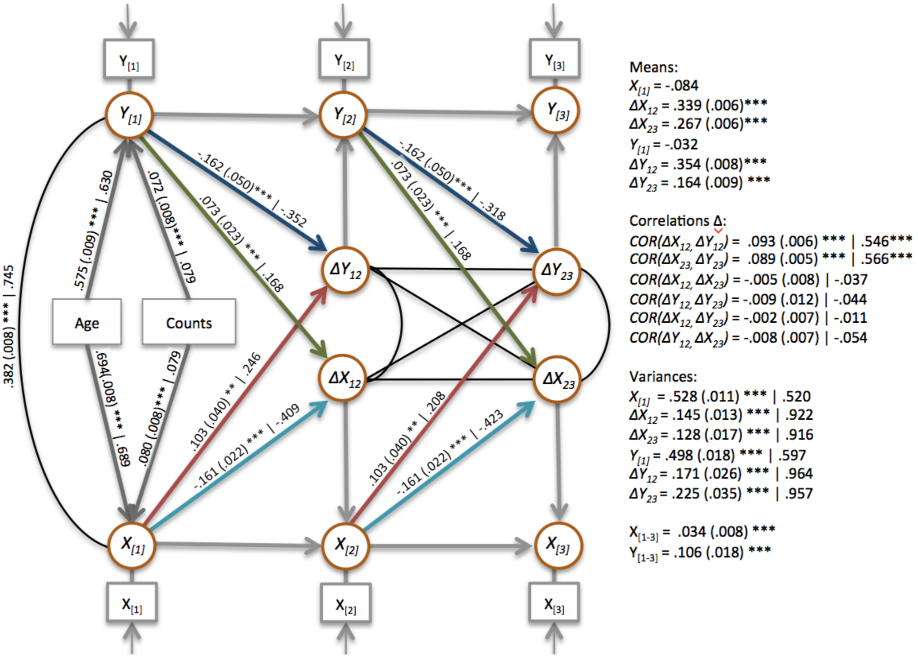

$Y_{[1]}$
$Y_{[2]}$
$Y_{[3]}$
$Y_{[1]}$
$Y_{[2]}$
$Y_{[3]}$
-.162 (.050)*** | -.352
.073 (.023) *** | .168
-.162 (.050)*** | -.318
.073 (.023) *** | .168
.575 (.009) *** | .630
.072 (.008)*** | .079
.382 (.008) *** | .745
ΔY_{12}
ΔY_{23}
Age
Counts
ΔX_{12}
ΔX_{23}
.694(.008) *** | .689
.080 (.008)*** | .079
.103 (.040) ** | .246
-.161 (.022) *** | -.409
.103 (.040) ** | .208
-.161 (.022) *** | -.423
$X_{[1]}$
$X_{[2]}$
$X_{[3]}$
$X_{[1]}$
$X_{[2]}$
$X_{[3]}$
Means:
$X_{[1]}$ = -.084
ΔX_{12} = .339 (.006)***
ΔX_{23} = .267 (.006)***
$Y_{[1]}$ = -.032
ΔY_{12} = .354 (.008)***
ΔY_{23} = .164 (.009) ***
Correlations Δ:
$COR(\Delta X_{12,} \Delta Y_{12})$ = .093 (.006) *** | .546***
$COR(\Delta X_{23,} \Delta Y_{23})$ = .089 (.005) *** | .566***
$COR(\Delta X_{12,} \Delta X_{23})$ = -.005 (.008) | -.037
$COR(\Delta Y_{12,} \Delta Y_{23})$ = -.009 (.012) | -.044
$COR(\Delta X_{12,} \Delta Y_{23})$ = -.002 (.007) | -.011
$COR(\Delta Y_{12,} \Delta X_{23})$ = -.008 (.007) | -.054
Variances:
$X_{[1]}$ = .528 (.011) *** | .520
ΔX_{12} = .145 (.013) *** | .922
ΔX_{23} = .128 (.017) *** | .916
$Y_{[1]}$ = .498 (.018) *** | .597
ΔY_{12} = .171 (.026) *** | .964
ΔY_{23} = .225 (.035) *** | .957
$X_{[1-3]}$ = .034 (.008) ***
$Y_{[1-3]}$ = .106 (.018) ***

Figure 6.5: The estimated model parameters for the best fitting mutualism model for the counting and addition data set. The first value denotes the unstandardized coefficient, with the standard error between brackets, and the value after the vertical bar denotes the standardized coefficient. The stars indicate the significant levels (* = p <.05, ** = p < 0.01 and *** = p < .001). The grey lines indicate fixed parameters. The observed scores (X and Y) are the latent trait estimates provided by Math Garden.

Both age and the number of items solved before T1 (counts) have large effects on the scores at T1, and the corrected scores in both domains are, as expected, still highly correlated. The means of the change scores indicate significant positive changes in ability from one time-point to the next. The self-feedback parameters are negative, which indicates that participants with high scores have lower change scores than participants with lower scores. These negative effects are often found, both in longitudinal data (e.g., Kievit, Lindenberger, et al., 2017) and in dynamic testing (e.g., Stevenson, Hickendorff, Resing, Heiser, & de Boeck, 2013). This could indicate regression toward the mean, as well as reflecting a deceleration in mathematical development over time. In principle these negative effects can also be caused by ceiling effects, but these were not present in the data. More importantly, the coupling parameters are positive, indicating that high scores in one domain result in more change in the other domain, as is predicted by the mutualism model. The self-feedback and coupling parameters explained 7.8% and 8.4% of the variance respectively for the first and second change score of addition (X) and 3.6% and 4.3% respectively for the first and second change score of counting (Y). For the multiplication and division data set these explained variances were 3.1% and 2.6% for the first and second change score of multiplication (X) and 1.9% and 0.4% for first and second change score of division (Y). Although these values differ, in general they indicate that a non-trivial part of the variance in the change scores can be explained by the effects included in the model.

We found significant correlations between the latent change scores of the different domains at the same time-points (respectively, .546 and .566 for the first and second change scores). These correlations show that after including the coupling parameters the residual changes are still positively correlated. This suggests that, unsurprisingly, it is likely that other cognitive or developmental factors influence change rates in both domains (see Appendix to Chapter 6) or that the analyzed data does not include enough waves to perfectly describe the true developmental processes (see Appendix to Chapter 6). Possible other cognitive candidates include growth of working memory, processing speed and attention factors. For the multiplication vs division data set these correlations were respectively .638 and .657. The remaining correlations between the change scores (correlations between ΔX_{12} and ΔX_{23}, ΔY_{12} and ΔY_{23}, ΔX_{12} and ΔY_{23} and between ΔY_{12} and ΔX_{23}) did not differ from zero in the counting vs addition data set. In the multiplication vs division data set these correlations were non-zero, varying between .104 and .199. The remaining residual correlations between the change factors $\Delta_{1,tp}$ and $\Delta_{2,tp}$, after including the direct links between learning processes, could be caused by unobserved coupling effects (see Appendix to Chapter 6), but also allows for an additional g-factor effect. Hence, these results are inconclusive about the explanation of the remaining correlations structure of the change factors.

To conclude, consistent with the conclusion from the analyses of the developing correlations, the comparison of different LCSM's indicated that, in line with the mutualism model, bidirectional coupling is present between the development of counting and addition, as well as the development of multiplication and division. A final alternative account to explain the large positive correlational structure between change scores can again posit a g-factor to explain these correlations between change-scores. Hence these results indicate that a non-trivial proportion of developmental processes can be ascribed to mutualistic processes, but a considerable proportion of the shared variance remains to be explain, either by a g-factor account or by incorporating more dynamically coupled cognitive domains.

6.4 Discussion

We compared two leading theories on the development of intelligence, g-factor and mutualism, using a large longitudinal dataset of primary school children's developing mathematics skills. The main finding is that a g-factor account of mathematics learning provides a too simplistic account of the developmental processes. Instead, to understand how children learn mathematics, a dynamic mutualism account, where learning one skill benefits learning another, and vice versa, needs to be added to the equation.

We found significant coupling effects during the development of counting and addition as well as multiplication and division. These mutually beneficial interactions could emerge through mechanisms in cognitive development and/or as a result of the educational processes Kievit, Lindenberger, et al. (2017). From a cognitive development perspective, it is possible that, on the one hand, learning to count is a prerequisite for solving addition problems and, on the other hand, some counting tasks are easier to solve using advanced addition strategies (counting sets of elements Carpenter & Moser, 1984). If so, we would observe direct coupling between overlapping cognitive skills, which could possibly be due to test items with overlapping skills (this hypothesis corresponds with the bonds model of Thomson (1951)). Since the transfer effects of learning between different tasks are usually not very strong (Barnett & Ceci, 2002), a second indirect coupling is also likely to occur via the (learning or school) environment of students. From an educational perspective, indirect coupling could, for example, be due to selection effects. The skills under study are relatively easy to observe in classrooms and this would allow teachers to intervene based on individual differences they see between children. An education system where more skilled children are selected for more challenging academic environments would result in an indirect coupling between learning different skills (see for example the proposed mutualism model of intelligence by p. 101 Kan (2012) where both direct and indirect environmental effects are included). In such an environment an excellent student in, for example, multiplication would be earlier provided with more challenging division items compared to his peers. These environmental effects are, for example, present in the relative age effect in education (Verachtert, De Fraine, Onghena, & Ghesquière, 2010) which can be explained by selection effects (Musch & Grondin, 2001).

Another explanation of the coupling effects we found could be based on developments in the modeling procedure. That is, it is likely that individual differences between children are present in the coupling between learning different skills. van der Maas et al. (2006b) showed that individual differences in coupling effects leads to vastly different learning curves between children, and also the complex factor structure often found in real data. For example, Ferrer, Shaywitz, Holahan, Marchione, and Shaywitz (2010) found that differences in coupling parameters between IQ and reading development were present between typically developing readers and two groups of children with reading difficulties (compensated and poor readers). The authors concluded that the differences in the dynamics between developmental groups accounted for the differences in reading development. If substantial individual differences in coupling parameters are indeed present, these could provide a new window for understanding development and this would allow for early detection of abnormal developmental trajectories.

Although we analyzed 'big' longitudinal data collected in a natural learning setting, the sample has some shortcomings. First, the children selected the games they wanted to play.

Most likely higher-ability children chose to play more difficult games earlier than lower-ability children; this would mean that we compared the scores of higher-ability children at earlier time-points with those of lower-ability children at later time-points. This self-selection bias would likely underestimate the average level of change between time-points. However, since this selection mechanism likely underestimates rather than overestimates average changes in ability, our findings remain valid. Second, the the density of data collection determines the level of detail we see in the window of development, which of course should mimic (to a reasonable extent) real developmental processes. But, coupling effects that occur on a different time-scale than the detail of data collection could influence the correlational structure of the change scores. The time-points we used, three within a full school year, probably resulted in an underestimation of the true coupling effects. This is also supported by the high correlations between the change factor of different skills at the same time-points. Third, as with other longitudinal data analyses, practice effects could have influenced our results (Salthouse & Tucker-Drob, 2008). However, as explained by Kievit, Lindenberger, et al. (2017), the presence of practice effects would only underestimate mutualistic coupling effects, because ability changes caused by practice effects would be less likely to result in coupling with other skills. For example, Lövdén, Ghisletta, and Lindenberger (2004) showed that, within a multilevel model framework, coupling effects increased when the models accounted for retest effects. Fourth, the estimated ability scores are based on responses in a low-stakes testing environment and children might be distracted or not motivated to do their best. However, we do not expect that this caused biases in the data for two reasons. On the one hand, we analyzed a large group of children, which means that possible differences between motivational differences between children are likely to cancel each other out. On the other hand, the motivational effects probably did not differ between time-points.

Despite these limitations, our findings have important implications for understanding the processes involved in the development of mathematics abilities. The mutualistic effects found in this study imply that during development different skills become more intertwined and possibly, at some point, become so strongly connected that they can be seen as a higher-level unidimensional skill. Hence, these higher-level factors emerge from direct links during development. These mutualistic effects imply accelerated growth during this developmental period and differences in the learning speed should be determined by the strength and number of coupled skills. If the presented results replicate, future research should study large-scale multivariate longitudinal data of important skills during development to get a more complete picture of the processes involved in cognitive development, and mathematics learning in particular.

7 Learning Analytics: Examining Processes in Time-Series of Responses to Single Items

This chapter is published as: Hofman, A. D., Jansen, B R. J., de Mooij, S M. M., Stevenson, C. E. & van der Maas, H. L. J. (2018) A Solution to the Measurement Problem in the Idiographic Approach Using Computer Adaptive Practicing. *Journal of Intelligence*, 6 (1), 14

Abstract

Molenaar's manifesto on psychology as idiographic science (P. C. Molenaar, 2004) brought the $N = 1$ times series perspective firmly to the attention of developmental scientists. The rich intraindividual variation in complex developmental processes requires the study of these processes at the level of the individual. Yet, the idiographic approach is all but easy in practical research. One major limitation is the collection of short interval times series of high quality data on developmental processes. In this paper we present a novel measurement approach to this problem. We developed an online practice and monitoring system which is now used by thousands of Dutch primary school children on a daily or weekly basis, providing a new window on cognitive development. We will introduce the origin of this new instrument, called Math Garden, explain its setup, and present and discuss ways to analyze children's individual developmental pathways.

7.1 Introduction

Mathematical proficiency is essential for functioning in today's society. Higher proficiency levels are associated with higher levels of employment (Hoyles, Wolf, Molyneux-Hodgson, & Kent, 2002; Finnie & Meng, 2006) and are, for example, necessary for making well-informed choices about health and health care (Reyna & Brainerd, 2007). Despite the importance of mathematics, relatively little is known about the development of mathematical abilities (van der Ven, Kroesbergen, et al., 2012). One main reason for this is the lack of an intra-individual perspective and the scarcity of large longitudinal data set collection (P. C. Molenaar, 2004).

In 2007, an adaptive online learning system for mathematical abilities was launched (Math Garden; Straatemeier, 2014; Klinkenberg et al., 2011) with the aim to collect the longitudinal data required for the study of developmental patterns in learning mathematics. Math Garden and other similar online learning systems can provide recommendations to students, predict their future performance, provide feedback for teachers and facilitate the development of cognitive models of student behaviour and learning processes (Romero & Ventura, 2010; Koedinger, D'Mello, McLaughlin, Pardos, & Rosé, 2015). The emergence of such large-scale online learning systems (Castro, Vellido, Nebot, & Mugica, 2007) have resulted in the availability of a new type of data set to study learning. In particular, such systems can provide high-frequency measurements of single individuals performing on single items, and such data have great potential in elucidating the processes of learning at the lowest possible level of study: the process of learning a single item.

In the current study we explore how to visualize, describe, and analyze data at this detailed level of an individual's responses to single items. Example response data are shown in Figure 7.1. Before delving into these data, we briefly introduce the Math Garden, which was used to collect the time-series data, in the next section.

Math Garden

Math Garden is a computer adaptive practice (CAP) environment for mathematics. In 2017 Math Garden collected around 500 million responses from 409,000 children on more than 22,000 items. In Math Garden, with the use of computerized adaptive testing (CAT), items are matched to children in such a way that each player has a fixed probability of answering an item correctly. This fixed probability corresponds to the chosen difficulty level. These difficulty levels - easy ($P(+) \approx 90\%$), medium ($P(+) \approx 75\%$) or hard ($P(+) \approx 60\%$) - can be selected by the players (Jansen, Hofman, Savi, Visser, & van der Maas, 2016). If players don't know the answer they can press the question mark button. After every incorrect response feedback about the correct response is presented to the player. The feedback provided helps facilitate learning (Van der Kleij, Feskens, & Eggen, 2015).

To match items to players, Math Garden uses an extension of classic computer adaptive testing (CAT). The extended CAT method Math Garden uses is based on two psychometric innovations. First, it implements an explicit scoring rule that incorporates both accuracy and response time. It is called the Signed Residual Time (*SRT*) scoring rule and was introduced by Maris and van der Maas (2012). The scoring rule discourages fast guessing and it makes the speed-accuracy trade-off explicit. Second, Math Garden uses an Elo estimation algorithm based on the Elo Rating System (ERS) that originates from chess competitions (Elo, 1978). Within each game, the Elo estimation is used to track the person ability and the item difficulty after each answered item (Klinkenberg et al., 2011).

This extended CAT method is based on two assumptions. First, the processes involved in both the accuracy and the speed of the response can be captured with a single latent trait. Although, recent work of Rijn and Ali (2017) showed that this assumption seems tenable, while Coomans et al. (2016) and Hofman et al. (Submitted) propose different approaches relaxing this assumption using an extended measurement model. Second, it is assumed that responses to multiple items within one game (e.g., the addition game) are due to a single unidimensional ability. This means that the responses of a person with a specific ability level to all items within a given domain are conditionally independent for that domain. If this assumption holds it should not be possible to find systematic differences in responses to items within each game for users with the same ability.

The Addition and Multiplication Games

For this research a subset of the addition and the multiplication data will be used. The data is from children that have visited Math Garden almost daily and who played frequently for prolonged periods. The data consists of a large set of person-by-item time-series, which are time-series of responses of a single child to a single item. The large amount of data on learning of individual children that is unlocked by these time-series is illustrated by Figure 7.1.

This player starts (week one) with correct and some incorrect responses to easy multiplication items. Items are ordered by item difficulty. Easier items are displayed at the

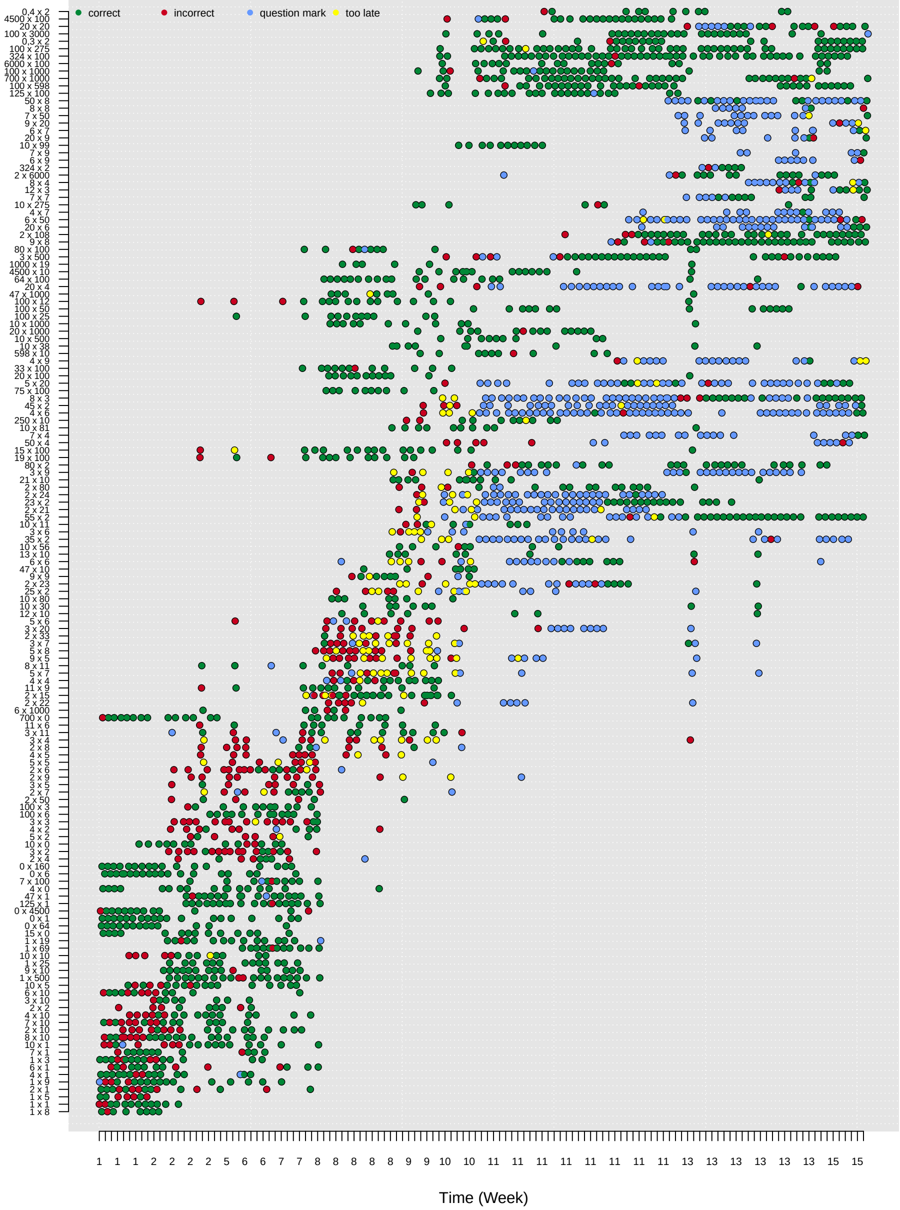


Figure 7.1: One player's development in learning to solve multiplication problems correctly. The colors refer to correct (green), incorrect (red), question mark (blue) or responses that were too late (yellow). The minimum number of responses for each time-series was 5. The items are sorted by item difficulty (low = easy and high = difficult). Plots for other players, providing different patterns, are available on `www.abehofman.com/papers`

lower end of the y-axis and more difficult items are placed at the higher end. Due to the CAT routine at each time-point only a subset of all items are presented. Around week two this player seems to learn the items with a times 10 operator. As a consequence his rating increases and more difficult items are presented. Responses to this more difficult set of items (e.g., 2×6 and 5×5) are more often incorrect, as is predicted from the measurement model. Around weeks eight, nine, and ten even more difficult items are presented (e.g., 2×33 and 19×100). Here the observed responses seem to deviate from the expected probabilities. These deviations are especially prominent for the long series of only correct responses to more difficult items (e.g., 64×100 and 100×12). These differences between observed and expected responses are again present in weeks thirteen and fifteen. The most difficult items (at the top of the figure) are almost always solved correctly, while the easier items - including some items that belong to the the standard multiplication tables - are still solved incorrectly.

Furthermore, the individual item patterns show very different patterns of longitudinal changes. Some items seem to be learned slowly (e.g., 7×10), whereas other item types seem to go through abrupt changes (e.g., 55×2). For some items it is even unclear whether the are learned at all, since the player continues to switch between correct and incorrect responses (e.g., 2×1). Also, for some items it seem that despite not being learned they are no longer presented (e.g., 2×9).

The visual inspection of these figures highlights many interesting patterns. However, the large number of users in these systems, combined with the fast rate with which responses are collected (about a million a day), makes it impossible to inspect these plots for all users. Hence, learning analytics are needed to characterize different patterns of learning, to highlight users that show interesting (deviating) developmental patterns, and in the end to use such analytics to provide users with person-specific support in their learning process.

Research Questions

We developed and investigated learning analytics to characterize different learning patterns.[1] These analytics are aimed to describe per item: (1) whether the child learned the item; (2) what the learning pattern is; (3) what the stability and variability of responses are over time. We focus on learning analytics that are feasible in a big data setting, such that they can also be of a practical value for online learning systems such as Math Garden. The learning analytics should thus be fairly simple and easy to compute. The aims of this study are twofold: (1) To shed light on the processes of learning mathematics and (2) To collect learning analytics to improve the system of Math Garden.

In Study 1 different learning analytics will be analyzed. In a follow-up project (Study 2) we investigate the dimensionality of the data using different analytics based on an item clustering approach. This clustering approach (based on an extend measurement

[1] We thank Timo Fernhout for his contributions to parts of this chapter.

model; Pelánek et al., 2016) is aimed to classify items into subsets of related mathematical constructs. These subsets are defined by stronger correlations within the item sets and weaker correlations between items in other sets. These sets possibly relate to different skills involved within the games and therefore allow a more detailed description of the individual differences between children.

7.2 Study 1: Learning analytics

Method

Data Selection

We collected responses of frequent players from the addition and multiplication games between 2013-09-01 and 2017-07-01. To this end, we first selected players with more than 1500 responses (N is 5339 and 4714, respectively). In both games so-called mirror items exist. Mirror items are items that only differ in the order of the operands (e.g., $1 + 2$ and $2+1$). Since these mirror items are closely related, responses to both mirrors were combined within a single time-series. In a second step, we omitted all responses to (mirror) items with less than five observations and only included players when they provided at least 250 responses to 25 different items.

The data selection procedure resulted in a data set of 3,801 and 3,711 players for the domains of addition and multiplication respectively. The data included in total 2,140,431 and 2,841,067 responses to 1169 and 741 mirror items. As expected from the results of (Jansen et al., 2016), these responses were mainly collected for children who played at the hard difficulty level. The probabilities of a correct response per difficulty level (see Table 7.1) were lower than expected based on the item selection procedure, but match the probability described by (Jansen et al., 2016).

Table 7.1: The descriptives of the response probabilities and times-series per difficulty level.

	Addition			Multiplication		
	Easy	Medium	Hard	Easy	Medium	Hard
$P(+)$	0.79	0.69	0.54	0.80	0.67	0.53
$P(-)$	0.14	0.19	0.22	0.11	0.15	0.15
$P(?)$	0.03	0.05	0.13	0.06	0.12	0.26
responses	700523	399587	1040321	395509	354736	2090822
series	66279	43983	121013	32754	31405	154258
mean length	10.57	9.09	8.60	12.08	11.30	13.55
max length	165	360	249	210	180	439

Note. $P(+)$, $P(-)$ and $P(?)$ are respectively the probability of correct, incorrect and question mark responses.

Learning analytics

To characterize individual learning curves, a number of candidate learning statistics were computed for each person-by-item time-series. These were the following (statistics were computed for correct and incorrect responses only, including question mark and too late responses as incorrect):

1. Response probabilities of the last two responses.

2. Transition probability matrix of correct and incorrect responses.

3. Coefficients of logistic regression model.

The percentage correct responses in the last two responses informs us whether users are able to answer an item correct at the end of a time-series. The transition probability matrix indicates how persistent switches from incorrect to correct responses are; this is a 2 by 2 transition matrix, where the probability of switching from an incorrect to a correct response and the probability of staying at a correct response are particularly informative. Since, the other two transition probabilities are complementary this information is redundant and will be omitted. Parameters of the logistic regression provide information on the person-by-item learning curve. The logistic function is:

$$P(x = 1|t) = 1/(1 + \exp^{-\beta_0 + \beta_1 x_t})$$

where β_0 is the intercept and β_1 is the steepness of the curve. We used the position in the time-series as the explanatory variable (x_t). The steepness of the curve reflects the speed of learning. A flat curve indicates that an item is already learned at the start of the data collection, or that an item is not learned during the period of data collection, depending on the value of the intercept (β_0).

We use Bayesian logistic regression instead of regular logistic regression, because the latter cannot handle complete separation (A. Gelman, Jakulin, Pittau, & Su, 2008). Complete separation occurs when a developmental trajectory involves a perfect step-like function between different states (Adolph et al., 2008). The models were fit using the *arm* package (A. Gelman et al., 2009) in *R* (R Core Team, 2013) using default priors. BIC differences were calculated between models with and without a slope parameter to compare the contribution of this parameter to model fit.

Results

The distribution of the length of the person-by-item time-series is shown in Figure 7.2. As can be seen, a high proportion of the person-by-item time-series are rather short. For the addition games these series were on average shorter than for the multiplication game. These differences can be due to (1) different numbers of items in the item banks, (2) possibly relate to violations of the unidimensionality assumption within games (see Study 2), or

(3) result from possible intrinsic differences between learning to add or to multiply. If indeed learning to multiply is more difficult than learning to add this might result in a slower progression through all items in the item bank and therefore in more observations of particular items.

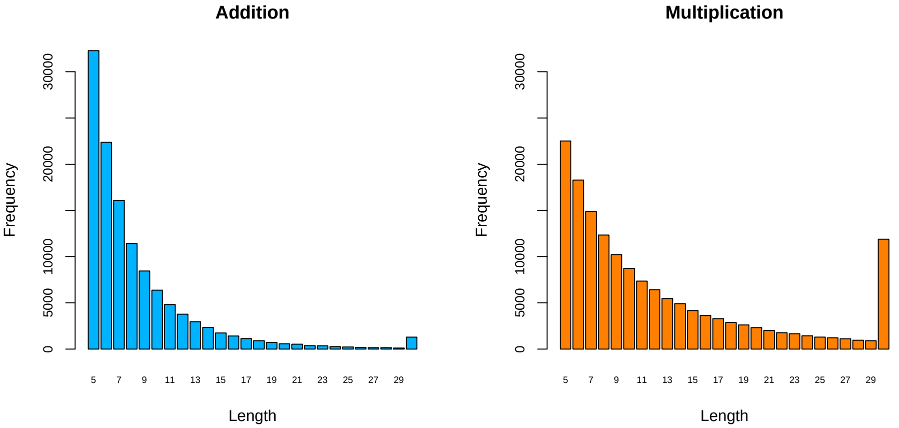


Figure 7.2: The distribution of the length of the collected time-series for the Addition (left) and Multiplication (right) games.

Stability, Change and Learning

First, we examined the proportion of correct responses for the last two or last five responses for each series. These proportions indicate whether the player learned the item at the end of the series. In the addition data set 35% of the series ended with two correct responses. This percentage is higher than the expected probability of .29 (0.54^2) based on the average response probability over all collected responses (see Table 7.1) and close to the measurement model implied average probability of .36 ($.6^2$).[2] For the multiplication data set 47% of the series ended with two correct responses, higher than the expected probability.

Second, the transition probabilities reflect the amount of switching between correct and incorrect responses (including question mark responses). These probabilities indicate whether an observed correct response reflects a truly learned state or whether switching is more accidental. Figure 7.3 shows a histogram of the switching probabilities between an incorrect (0) and correct response (1) (*learning probability*), indicating learning, and the

[2] For many results in this section, formal significance tests could be provided. However, we refrain from doing so because: (1) no clear a priori hypothesis could be formulated and (2) due to the large data sets very small an uninteresting results become significant in a null-hypothesis testing framework. For example, the 35% did significantly differ from the expected 36% according to a simple proportion test ($X^2(1) = 58.60, p = 1.936e - 14$)

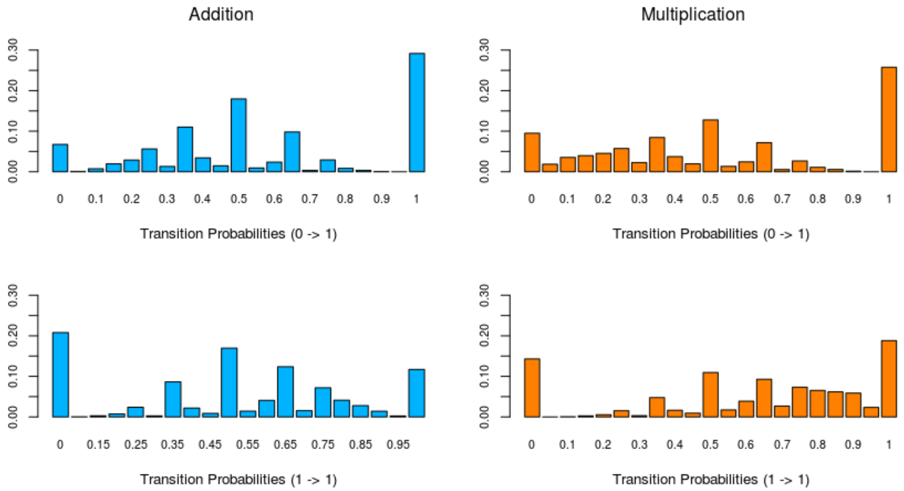


Figure 7.3: The distribution of the transition probabilities of switching for incorrect to correct (top) and staying at an correct response (bottom) for addition (left) and multiplication (right) for all collected time-series.

probabilities of *staying* at a correct response for each domain.[3] For both domains there is a clear peak at one (bar at the right), indicating that in about 30% of the addition series and 25% of the multiplication series the *learning probability* is one. For the staying probabilities a large peak is observed at zero, indicating that in 20% for addition and 15% of the series for multiplication a correct response is always followed by an incorrect response. Furthermore, a comparison of the probabilities of staying in the learned state (1 → 1) between the addition and multiplication data shows that for addition these probabilities are lower than for multiplication. In general, these results imply that the switch from incorrect to correct responses is not very stable.

Third, we investigate the evidence for learning in the time-series by fitting learning curves with logistic regression models. To explore the fit of these models to the observed time-series we plotted the observed and predicted responses of three series of the same player in Figure 7.4.

For the addition data set 19% of the models fitted to the time-series included significant slope parameters (as indicated by the BIC difference between the model with only an intercept and the model with both a slope and an intercept parameter). Of these time-series that included a significant slope, 70% of the slopes were positive. For the multiplication data set 36% of the series were best described with a model including a slope parameter, and 86% of these were positive. The slope parameters for multiplication were higher ($\overline{\beta_1}$ = .25) than the slope parameters for addition ($\overline{\beta_1}$ = .08), see also the left panel of Figure

[3] The incorrect to incorrect and correct to incorrect are not depicted since these are simply the inverse of the presented probabilities.

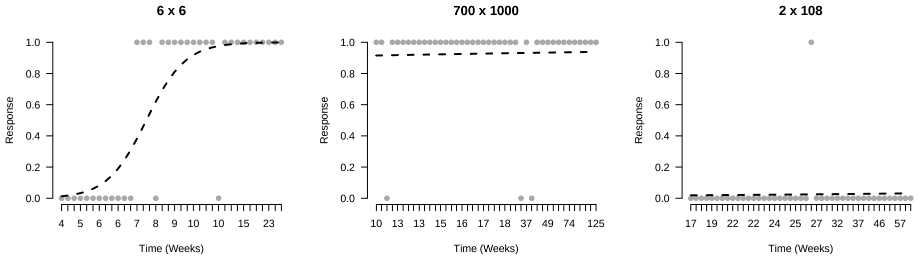


Figure 7.4: An example of three different developmental patterns of responses to different items by the same player. The left panel shows a time-series with a clear increase in the probability of a correct response. The middle and right panel respectively show a series of a previously learned item and a series that indicates no learning.

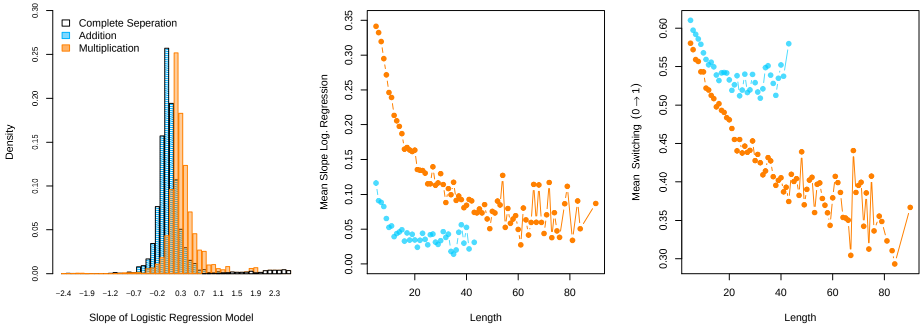


Figure 7.5: The distribution of the estimated slope parameters for both data sets (left-panel), and the relation between the length of the series and the slope parameter (middle-panel) and the switching probability from an incorrect to a correct response (right-panel).

7.5. Furthermore, for both data sets the time-series length was negatively correlated with slopes and *learning probability* in the transition matrix.

For a better understanding of these learning curves, we investigated differences between the estimated slopes of children. First, a positive correlation was found between the average learning between the domains addition and multiplication ($\rho(826) = .345, p < .001$; only players with more than five time-series in both domains were included). Second, we investigated the correlations between the slopes on different items within the multiplication domain. Based on the patterns of Figure 1.1 and the results of van der Ven et al. (2015) two different item clusters can be defined: items than belong to multiplication tables 2 through 9 (*table* items) and items with a times 10, 100 or 1000 operator (*time 100* items). Within the multiplication domain no significant correlation was found between these two item clusters ($\rho(185) = -.003, p = .964$). We tested whether the correlation between these two sets of multiplication items is indeed lower compared to correlations based on more similar items within the domain. To this end, we used a permutation test[4] to calculate the correlation between learning curves of two random sets of *table* and *times 100*. The average (within cluster) correlation for *table* items was .471 (SD = .067) and for *time 100* items was .420 (SD = .050). These results indicate that players who show steeper learning curves on *table* (*time100*) items also show steeper learning curves on other *table* (*time 100*) items, but learning between items sets is unrelated. These results are surprising and will be further explored in Study 2.

To conclude, the results about learning, stability, and change show that only for a small set of the series significant learning seems to be present in the data. Although large variations in learning can be expected (Siegler, 2006), some variability is caused by the manner of data collection in an adaptive learning environment. Items are selected with a constant fixed expected probability correct. In these systems, items that are too easy or too difficult for children are not presented, hence more variability is expected on the observed time-series. A second explanation for the small learning curves is provided by possible violations of unidimensionality. This will be the topic of the next section.

7.3 Study 2: Exploring Dimensionality

In Study 2 we investigate possible violations of unidimensionality of the measurement model currently used by Math Garden (see Klinkenberg et al., 2011; Maris & van der Maas, 2012). Such violations, resulting from possibly different types of heterogeneous response behaviour, could result in lower model fit and thereby worse predictions of the expected score based on the estimated person's ability and item difficulty parameters. One indication of such a violation is already provided by the zero correlation between the learning curves of *table* and *time100* items. First, differences between these two sets of predefined item types are further investigated. Second, using an extended measurement model we examined the presence of item clusters in both the addition and multiplication data set.

Methods

Analytics to Investigate Dimensionality

In order to find deviations from the expected patterns based on the Math Garden model, the following statistics of person-by-item time-series were collected:

1. Percentage correct (question-mark responses were labeled as incorrect) in the last 10 responses (omitting shorter time-series).

2. Percentages question mark, correct responses, and incorrect responses.

[4] For the permutation test we used 5000 replications where in each replication all items were randomly distributed between two sets. For each set we calculated each player's average slopes and the correlation between both of these averages values.

3. The mean difference between observed and expected score according to the *SRT* scoring rule.

Using these statistics, the most important deviation that will be tested is the violation of the assumption of unidimensionality with respect to different sets of items. In particular between items that belong to the table two through nine (*table* items) and items that include a times 100 or 1000 operator (*time100* items). This first analysis allows us to compare multiple predefined item sets. In a second approach, we will explore the possibility of multiple item clusters that, if present, violate the unidimensionality assumption.

Model Extension and Item Clusters

In a recent paper by Pelánek et al. (2016), different extensions of Elo models are presented. One of these extensions, called the *networked* model, is especially suited for estimation of item clusters. In the *networked* model the expected score is based on an weighted sum of the global skill (θ_{global}) and the local skill (θ_{local}):

$$P(x = 1|\theta_{global|p}, \theta_{local|p}, \beta_i) = 1/(1 + e^{-(w_1\theta_{global|p}w_2\theta_{local|p}-\beta_i)}),$$

where the configurations of the weights (w_1 and w_2) were both set to .5. In our case we estimated a *local* skill for each individual item. In a second step the clustering of item – using the *local* skill estimates – was explored using a hierarchical clustering algorithm.[5] In this model we only focus on the response accuracy. The model parameters were updated in the following order:

$$\begin{aligned}\theta_{global|new} &= \theta_{global|old} + K_p * (S - E(S))\\ \beta_{new|i} &= \beta_{old|i} - K_i * (S - E(S))\\ \theta_{local|new} &= \theta_{local|old} + K_p * (S - E(S)),\end{aligned}$$

where K_p was set to .25 and K_i to .01.

Data Selection

Two different data sets were analyzed for both the addition and the multiplication game. First, for the computation of the analytics on dimensionality the data of Study 1 was used. Second, for the estimation of the *networked* model, we selected the responses (accuracies) of the 200 most played items of players who completed at least 20 sessions of 15 responses between ′2014-09-01′ and ′2017-06-01′. This resulted in 5,144 and 8,180 users providing in total 2,708,027 and 4,557,333 responses for the addition and the multiplication game, respectively.

[5] These values could be optimized with different cross-validation procedures on training data.

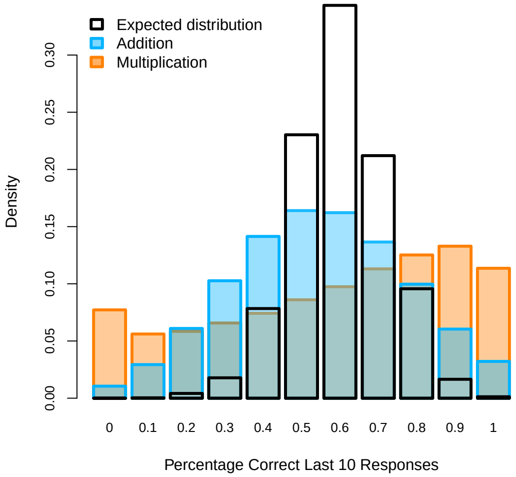


Figure 7.6: The distribution of the average response probability of a correct response per time-series of the Addition and the Multiplication game and based on the expected response probability according the the CAT algorithm.

Results

Analytics

The distribution of percentage correct, $P(+)$, for the last 10 responses to an item for all time-series is plotted in Figure 7.6. Included in this figure is the expected distribution. The expected distribution was based on samples from a binomial distribution with a $P(+)$ of .6 (corresponding to the hard difficulty level)[6] and the number of samples that corresponds to the length of the collected time-series. For both the addition and the multiplication data, the variance of the distribution of the average $P(+)$ was higher than the distribution of the expected $P(+)$. For the multiplication data set there was even a large density observed at the tails of the distribution. This reflects that (contrary to the model expectancy) in these series either only incorrect or correct responses were observed. To further examine whether this large variance is due to a violation of unidimensionality, we investigated differences between the addition and the multiplication items and between the *table* and *times 100* subsets of the multiplication items.

The mean percentages correct per item did not differ between the addition and multiplication games. However, within the multiplication game differences were present between the two sets of items. The left panel of Figure 7.7 shows that the *time 100* items had a higher

[6] Note that also a $P(+)$ of .54 could be used based on the overall observed response probabilities, see Table 7.1. Although this would change the location and the variance of the distribution slightly, the observed variances were also much larger than the expected variances based on this $P(+)$

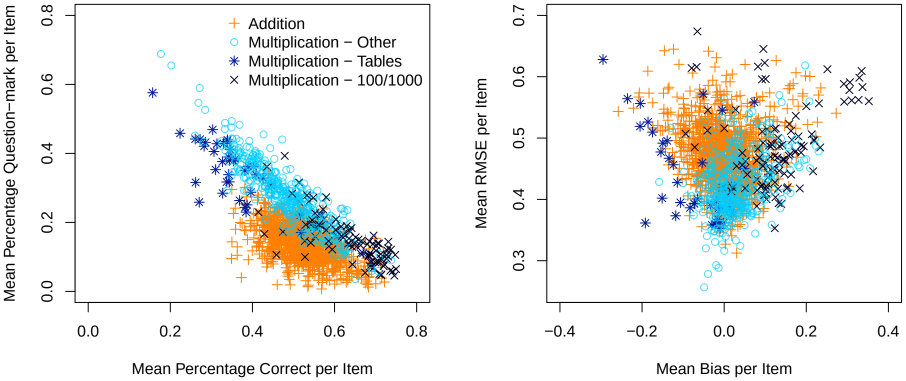


Figure 7.7: The average response probabilities (accuracy and question mark; left-panel) and the bias and root square mean error (RSME; right-panel) per item.

probability of a correct response and a lower probability of a question mark response compared to the *table* items. These differences between these sets of items were also present in the model fit - the difference between the observed and the expected score according to the *SRT* model. The right-panel of Figure 7.7 shows the bias (the average difference between the observed and the expected score) and the root mean square error (RMSE) per item, taking both the accuracy and response time into account. For both the addition and the multiplication game, as expected, the item bias was generally centered around zero. However, the *time 100* items had a negative bias whereas the *table* items showed a positive bias.

Next, we found a negative relation between the average response probabilities of users for the *table* and the *time 100* items ($\rho(185) = -.416, p < .001$; we only included users with time-series of a minimal length of 10 to more than five different items).[7] This indicates that for a subgroup of users *table* items were easier than predicted while the *time 100* items were more difficult than predicted, and one group of users for whom it was the other way around.[8] This further supports the results following from the item perspective, that

[7] We compared this correlation coefficient with the estimated correlations between the average response probabilities of users based on two random sets of items within the *table* or *time 100* set (using a permutation test with 5000 replications). The average correlation between two random sets of *table* items was .815 (SD = .030) and between sets of *time 100* item was .727 (SD = .070). Hence, users that score higher than expected on one set of items, score lower than expected on the other set of items.

[8] It should be noted that the negative relationship between the percentages correct for the 10, 100 and 1000 operators and for the regular multiplication tables is a local effect due to the adaptive nature of Math Garden. This negative relationship holds for users who have roughly the same ability estimate for multiplication, because only users with roughly the same ability estimate will make the same items. Thus, it does not mean that the same negative relationship would hold if users would be presented with all items in a non-adaptive test.

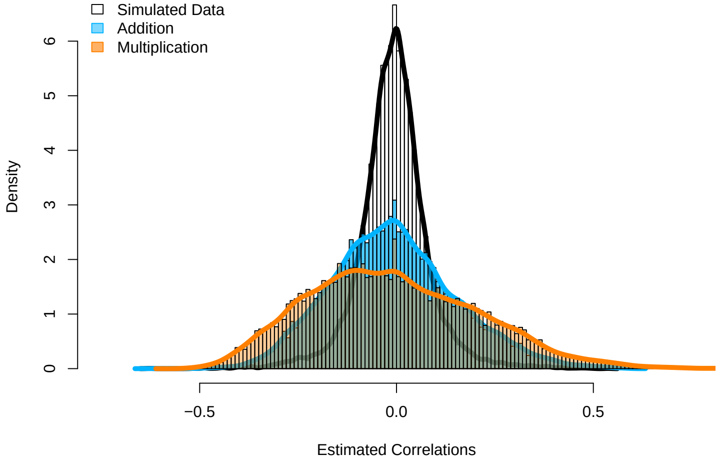


Figure 7.8: The distributions of the estimated correlations between the local skills in the addition, multiplication and simulated data. Gray squares indicate missing values in the correlation matrix, resulting from the adaptive item selection.

indeed violations of unidimensionality seem present between the *table* and *time 100* item sets.

To conclude, the presented analyses on dimensionality suggest that within both the addition and the multiplication domain some violations of the unidimensionality assumptions occur. Furthermore, in the multiplication domain these problems seem to be more severe, and related to differences between items of the 2-9 tables and items with a time 100 operator. In the next section we explore whether more item clusters are present in the multiplication game and whether any clusters can be found in the addition game.

Item Clusters

We visualized the correlation matrix of the estimated local skills (θ_{local}) to explore patterns in this correlation matrix. Therefore we used a heatmap based on a hierarchical clustering procedure using the *clustR*-package (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2017). In this clustering procedure initially every local ability was assigned to its own cluster. In each iteration the two most similar clusters were merged based on the sum of the absolute differences of the estimated correlations. This step was repeated until all local abilities were collapsed into a single cluster.

To interpret the results we first compared the heatmaps based on the addition and the multiplication data set to a heatmap constructed on simulated data. For this simulation we generated responses with an unidimensional *SRT* model and fitted the *networked* model to

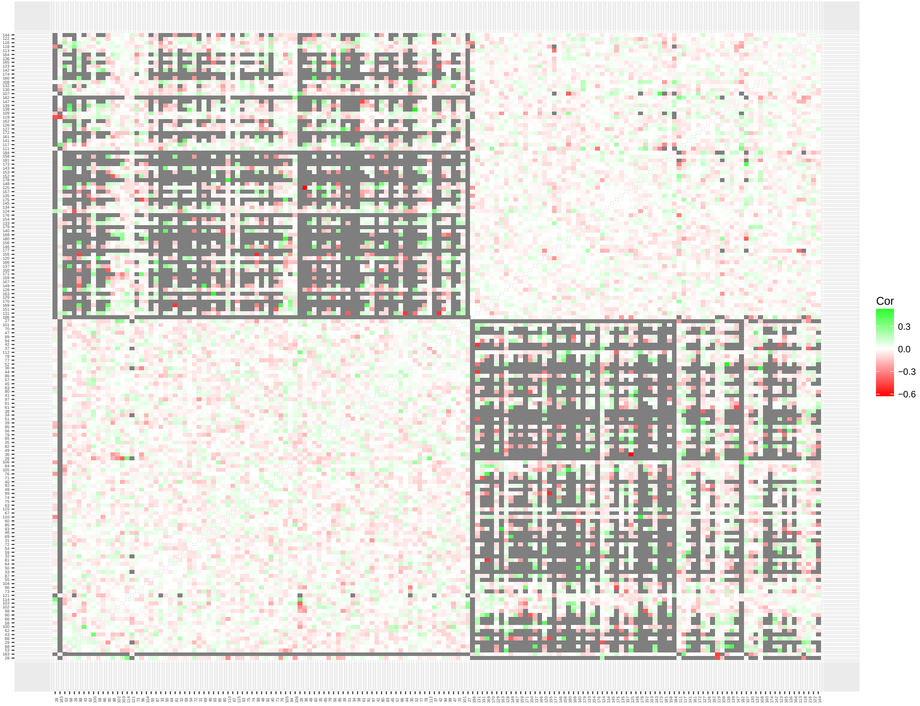


Figure 7.9: A heatmap based on the correlation matrix of the local user abilities estimated on a simulated data set based on the unidimensional *SRT* model. Gray squares indicate missing values in the correlation matrix, resulting from the adaptive item selection. The figure can also be found on `www.abehofman.com/papers` allowing a more detailed inspection.

the generated accuracy data. In the generated data no local skills were present. Hence, extending to model by estimating these local skills would merely result in capturing random fluctuations (error) in the generated data. Hence, the correlation matrix of the local skills will in this case be centered around zero and no specific patterns between items are expected. As shown by Figure 7.8, the estimated correlations are indeed around zero. The heatmap based on this correlation matrix shows no clear patterns (see Figure 7.9). Although the clustering metric seems to be sensitive to the missing data patterns, no clear clustering structure is found on the estimated correlations.

A comparison of the density of the estimated correlations between the local skill of both collected data sets and the simulated data set showed large discrepancies, as indicated by the large tails in both distributions in Figure 7.8. Furthermore, the heatmaps based on both the multiplication data (see Figure 7.10) and the addition data (see Figure 7.11) showed clear item clusters of items with related content. First, replicating the results of the previous section on Analytics, a large cluster of items that involve a times 10, 100 or 1000 operator was found. As expected, the local skills within this cluster had a strong

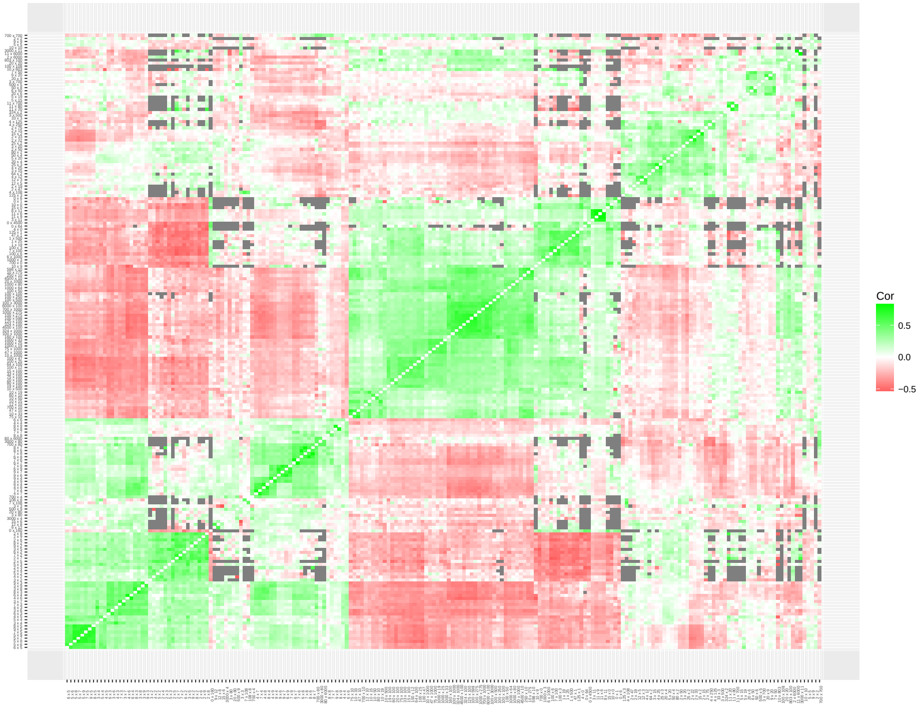


Figure 7.10: A heatmap based on the correlations matrix of the local user abilities of 200 multiplication items. Gray squares indicate missing values in the correlation matrix, resulting from the adaptive item selection. The figure can also be found on `www.abehofman.com/papers` allowing a more detailed inspection.

negative correlation with the items of the tables 2 through 9 (*table* items). The *table* items are not clearly represented in a single cluster. However, based on content of the items the clustering solutions seems clearly interpretable. For example the items 700 × 80, 3000 × 80 and 80 × 6000 are placed close to the items 8 × 7 and 8 × 6. A third weaker cluster seems present that included items that involve larger, more complicated, calculations without any times 10, 100, or 1000 operator.

For the addition data set the cluster solution seem less prominent, however certain item clusters do allow for an obvious interpretation. First, a cluster of addition items with relative small solutions is present (add 2, 3, 4 or 5), with negative correlations between items with large solutions that involve adding tens (e.g., 40 + 10 and 6 + 90). A clear third small cluster is present with items that have an add zero term. However, a large set of addition items cannot be clearly assigned to a cluster. This indicates that for the addition data set the violations of unidimensionality seem less severe than for the multiplication game.

To conclude, the different analytics show that *table* and *time 100* items are best described

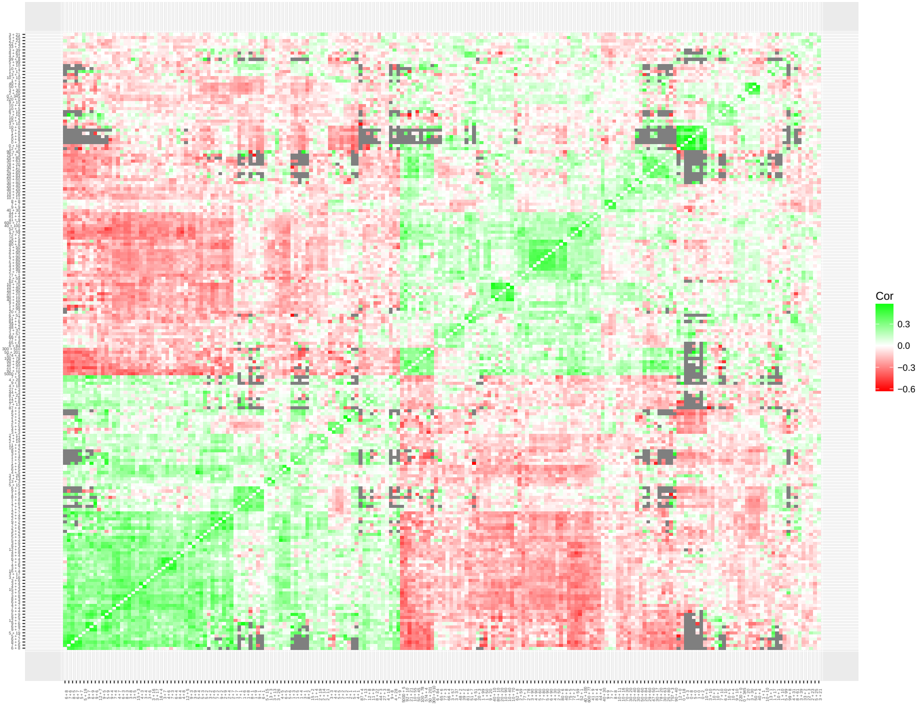


Figure 7.11: A heatmap based on the correlations matrix of the local user abilities of 200 addition items. Gray squares indicate missing values in the correlation matrix. See `www.abehofman.com/papers` for a downloadable version.

by two different skills, resulting in heterogeneous response patters. Furthermore, the results based on the *networked* model show that local skills can be estimated that provide an important addition to the global skill currently used in the Math Garden. Furthermore, the correlations between these local skills should be described by multiple clusters that could be interpreted based on the item content. Follow-up analyses using cross-validation techniques should determine both the weights in the expected score formula and the optimal number of clusters.

7.4 Discussion

In the current paper we collected and explored different analytics based on time-series data of responses of children to single items. The results of Study 1 shows that the learning analytics on the collected time-series often fail to provide support for learning at the item level. In Study 2, different analyses indicated that multiple item clusters are present, possibly caused by qualitative different solutions strategies in both the addition and the multiplication game.

Since the data was collected adaptively the learning patterns were expected to show more variability than in non-adaptive environments. To better investigate different learning curves, the *CAP* system could be modified to collect more interesting time-series. The current item selection procedure could be changed such that with a certain probability one of a set of pre-selected items is selected. Although this would result in a less optimal item selection for practicing, it would result in the collection of valuable time-series data to further study development. Siegler and Crowley (1991) already stressed the importance of such frequent sampling to collect longitudinal data as exemplified in his microgenetic method. The key properties of this method are: observations span the full change process, the density of observations is high relative to the rate of change, and the analysis is aimed to infer the process that gives rise to both quantitative and qualitative changes (Siegler & Crowley, 1991). Since it is very time intensive to collect this type of data using more traditional approaches *CAP* systems - possibly with some extensions - do seem a promising way to collect this type of data.

One of the main findings of microgenetic research is that children's cognitive development is highly variable; the results of Study 1 are in line with this observation. Another main finding based on the microgenetic method is that strategy-usage is highly variable within children when a problem is repeatedly presented close in time (Siegler, 2006). According to Siegler and Crowley (1991) multiple strategies are available to an individual and new more advanced strategies are not consistently applied. Over time the more advanced strategy will replace the older less adequate strategies. Hence, developmental change is not sudden, from strategy A to strategy B, but characterized by continuous shifts in the distribution of use of multiple (in)correct solution strategies (Kuhn et al., 1995), which Siegler refers to as the overlapping waves theory (Siegler, 1996). Moreover, this variability in strategy-use is often found in mathematics learning (Lemaire, 2010; Ambrose, Baek, & Carpenter, 2003). For example, children start with mostly simple counting strategies (Dowker, 2005; Ashcraft & Guillaume, 2009) and after they gain experience these will be replaced with more complex strategies, such as repeated addition for solving single digit multiplication (van der Ven, Boom, Kroesbergen, & Leseman, 2012). In an experiment done by Lemaire and Siegler (1995) children progressed to using more complex strategies more often, but at each time point children used a mixture of strategies.

The results of Study 2 supports this view on the development in mathematics. Although each game consists of items that belong to a clearly defined domain (e.g., multiplication), we found multiple indications of severe multidimensionality. In the CAP system of Math Garden this results in errors in predicted scores, and larger variances in the difference between the observed and the expected scores. As Straatemeier (2014; p.174) proposed, ideally ability estimates would be based on unidimensional small item clusters within large item banks. We expect that separate ability estimates for these clusters will provide detailed insights into students' skills.

Currently Oefenweb is implementing a new feature that allows items to be classified into learning goals as defined by a Dutch centre of curricula development *SLO* (2009). Based on the approach of Study 2, a more data driven modeling framework could be used to construct these learning goals, using the detected item clusters. Hence, with an extended version of the current model that includes ability estimates of item clusters within a domain (Pelánek et al., 2016), we could capture multidimensionality and provide teachers with additional information about erroneous strategies or subsets of items for which a student lacks knowledge or skills. Additionally, Oefenweb is developing a learning module where students can select which learning goals they want to practice. This new way of focused practice aimed at certain misconceptions in well defined item clusters seems a very promising next step for the Oefenweb systems.

8 General Discussion

8.1 Summary of Main Findings

In this dissertation entitled "Psychometric Analyses of Computer Adaptive Practice Data: A New Window on Cognitive Development" we investigated a wide set of cognitive tasks that are either directly or indirectly related to an educational setting. These tasks ranged from general abilities such as proportional reasoning to specific skills such as learning to touch type. The majority of the data was collected with adaptive web-based games from Oefenweb, like Math Garden. In the first part of this dissertation we analyzed cross-sectional data to study between person differences, such as the cognitive strategies involved in the different tasks (Chapters 2-4). In the second part of this dissertation we used a longitudinal approach to uncover within person dynamics (Chapters 5-7).

In Chapter 2 we compared a Math Garden data set with a data set from a controlled experiment. The task was for children to determine the number of elements in a visual display, which could be achieved by counting, subitizing or a combination of strategies. We manipulated the configurations of the elements (random, line or dice patterns) during data collection and investigated whether this influenced the children's enumeration strategies. The manipulation affected the performance in the counting range (4-6 elements), but not in the subitizing range (1-3 elements); this supports the idea of two distinct strategies. Since we found similar results on the items with dice patterns for the subitizing and counting range, this might indicate that subitizing is best seen as a pattern recognition process.

Based on the analyses of two cross-sectional data sets collected with the balance-scale task (Inhelder & Piaget, 1958; Jansen & van der Maas, 2002), we found in Chapter 3 that responses in the Math Garden data set were best described by more implicit information-integration approaches, whereas responses in the paper-and-pencil data set were best described by explicit rule-based strategies. This discrepancy shows that the results based on traditional paper-and-pencil data cannot be generalized to those collected with online learning environments such as Math Garden.

In Chapter 4, cross-sectional data from the multiplication game in Math Garden were analyzed. Chapter 4 illustrates, by using a novel psychometric approach, that fast processes are qualitatively distinct from slow processes in basic multiplication. These results support the idea that multiplication involves both faster fact-retrieval processes and slower back-up strategies. Allowing for these differences resulted in a more detailed and better fitting measurement model compared to a simpler unidimensional model.

In Chapter 5 we analyzed two longitudinal samples from Math Garden. One sample concerned data from the counting and addition games whereas the other concerned from the multiplication and division games. It was investigated whether a one-factor latent variable model (as implied by the g-factor theory) or a mutualistic network model best describes developmental patterns and the interrelations between two related domains. A latent-change score specification of both models allowed for a direct comparison of the models. The mutualistic network model fit the data better for both the counting and addition data set and the multiplication and division data set. This showed that, although cross-sectionally the patterns can be described by a one-factor model, the one-factor model is too simplistic when using a developmental perspective and describing the dynamics in the developmental system.

In Chapter 6 we analyzed longitudinal data from the typing game in Type Garden. Cross-sectional studies show qualitative differences between novice and experts and the aim of Chapter 6 was to investigate how these differences might emerge. To this end we analyzed the development of learning to type from keystroke data provided by a group of children who completed a typing course in Type Garden. The longitudinal results supported the two loop theory that predicts qualitatively different strategies between novice and expert typists. We were able to show that the developmental patterns in each loop were different, despite large individual differences.

In Chapter 7 longitudinal data were analyzed from the addition and multiplication games in Math Garden. We used a data-driven approach to classify patterns in time-series of individuals on single items. The clustering of specific item abilities indicated that multiple small clusters are present within the large item banks of both the addition and the multiplication games. In both games, estimating the ability for each cluster of items refines the assessment of children's skills.

To summarize, the results of the dissertation corroborate earlier findings in cognitive development and educational psychology and also provide new insights. The dissertation illustrates the range of possibilities of the data from the computer adaptive games in Oefenweb for studying development and learning. The data also pose important psychometric challenges that were identified and addressed. In the next section we discuss the challenges that will be the subject of future research of the psychometric research group at the University of Amsterdam (UvA) and of the development team at Oefenweb. This discussion concludes with a short reflection on psychometrics, cognitive development and educational practice.

8.2 Psychometrics in Math Garden Systems: The Next Steps

The research in this thesis has brought to the forefront many aspects of Math Garden system that are as yet unstudied. In what follows a number of these aspects are discussed along with suggestions for directions of research to further understand and improve adaptive practice systems.

Estimating the Dimensionality of Games First, within each game a single ability is estimated based on the responses to all items in the item bank. This approach assumes unidimensionality: a single ability underlies the individual differences between children on all items. Although each game consists of items that belong to a clearly defined domain (e.g., multiplication), Chapter 7 shows that within a domain different clusters of items can be present. The presence of item clusters cannot be captured by Math Garden's current IRT model; this results in errors in predicted scores and larger variances in the difference between the observed and the expected scores. In the multiplication game, the items in a cluster have similar content (e.g., basic multiplication problems versus problems with an operator of 100 or 1000) and require the same multiplication strategies (Pelánek et al., 2016). Using learning analytics, these clusters were detected and could be clearly interpreted, thus supporting their validity.

As Straatemeier (2014, p. 174) already proposed, ideally ability estimates are based on small item clusters within the large item banks. Item cluster abilities could provide detailed insights into students' skills. Oefenweb is currently implementing a new feature that allows items to be classified into learning goals as defined by *SLO* (2009). Alternatively, the presented research in Chapter 7 provides a more data driven modeling framework for detecting such item clusters. Learning goals can be formulated for the detected item clusters. Hence, with an extended version of the current model that includes ability estimates of item clusters within a domain (Pelánek et al., 2016), we could capture multidimensionality and provide teachers with additional information about erroneous strategies or knowledge that is lacking for particular subsets of items. Oefenweb is also developing a learning module where students can select which learning goals they want to practice. This new way of focused practice aimed at certain misconceptions in well defined item clusters seems a very promising next step for the Oefenweb systems.

Accuracy and Speed The second issue is also related to a unidimensionality assumption in the current model, but in this case concerns the accuracy and speed of the responses. It is clear that response times entail important information about the solution strategies in mathematics (e.g., Campbell & Austin, 2002), in addition to accuracy. However, different choices can be made on combining these sources of information in a measurement model. First, the *SRT* model introduced by Maris and van der Maas (2012), collapses information about accuracy and response times into a single score to estimate the ability of students, and thereby imposes a one-to-one relation between accuracy and speed. Second, a more flexible approach is introduced in the seminal work of van der Linden (2007). In this measurement model (*Van der Linden* model)) a correlation is introduced between two sets of latent abilities: one related to the accuracy of the responses and one related to the (log of the) response times. Using this approach the information in the response times is used to optimize the precision of the ability estimates based on accuracy. Rijn and Ali (2017) showed in a computerized adaptive testing context that the reliability of the estimated parameters was higher for the *SRT* model than for the model by *Van der Linden*. Third, Partchev and De Boeck (2012) propose an even more flexible approach by separating fast and slow responses (*fast-slow* model). Coomans et al. (2016) compared each of these frameworks on different Math Garden data sets and concluded that the most flexible *fast-slow* model is needed to capture all patterns in the data (these results correspond with the results in Chapter 4).

Future psychometric work entails the implementation and evaluation of different modeling approaches in Oefenweb games to deal with the imperfect correlation between accuracy and response times. First, Hofman et al. (Submitted) and Savi, Ruijs, Maris, and van der Maas (2017) show that there is a relatively large proportion of fast question mark responses. These fast question mark responses violate the assumption in the *SRT* model, and are also undesirable from an educational perspective. An A/B test was implemented (Savi, van der Maas, & Maris, 2015; Savi, Williams, Maris, & van der Maas, 2017) to test the effects of delaying the possibility to provide a question-mark response (Savi, Ruijs, et al., 2017). Results indicated that, as expected, fewer question mark responses were provided. Although direct effects on the model fit were not tested, these results will have a positive effect on the fit of the *SRT* model.

A second approach to circumventing the problem of the imperfect relation between accuracy and response times involves slight variations of the *SRT* scoring rule. A scoring rule that does not punish fast errors more than slow errors (i.e., a penalty independent of response time) seems very promising for two reasons. First, such an adaptation of the *SRT* rule allows for removing punishment, which can be encouraging for insecure students. Second, this variation of the *SRT* model implies a different relation between speed and accuracy, possibly resulting in a better description of the data.

A third approach to take into account the imperfect relation between accuracy and response times in a measurement model involves the incorporation of a complete new, state-of-the-art tracking system. Current work of Gunter Maris (and others) is based on an urning-scheme (see next section). One of the possibilities of using this approach is tracking two sets of abilities based on either accuracy or speed, and the possibility to provide (on-the-fly) significance testing of whether the abilities can be collapsed.

Optimizing Prediction and Tracking A third psychometric issue concerns the optimization of the estimation of abilities and difficulties. In Math Garden, abilities of children and difficulties of items are estimated continuously. Each response is used to update both. Although abilities may be stable in the short term, they will grow in the long term as practice enables learning. In contrast, item difficulties are assumed to be stable in the short term and only minor changes in the long term.

Ability estimates must keep up with the growing abilities that result from the learning process. However, responses are noisy and ability estimates may fluctuate too much if the updating process is too sensitive, introducing variance and lowering the reliability of estimates. Hence, there is always a trade-off between bias (estimates do not grow) and variance (estimates fluctuate too much) in optimizing the estimation process.[1]

In Elo systems, a K-parameter is introduced to optimize the bias-variance trade-off. In general, large (small) values of K provide a large (small) weight to new responses, and thereby introduce a high (low) variance and a low (high) bias in the system; the Elo update rules for person abilities and item difficulties are:

$$\theta_{new|p} = \theta_{old|p} + K_p * (S - E(S))$$

$$\beta_{new|i} = \beta_{old|i} - K_i * (S - E(S))$$

Two approaches are currently available to determine the value of K_p and K_i in order to optimize the bias-variance trade-off. A first approach is based on the off-line re-calculation of all model parameters, using different values of K. Metrics based on differences between observed and expected scores help in determining the value of K. A disadvantage of this approach is the off-line character. A second approach is to determine K on the outcomes of a large A/B test. Such an approach has been implemented recently. In ten different settings, different values of K and of related parameters are used. These ten conditions are based on the initial results of the first approach. Several metrics showing differences between observed and expected scores are again used to decide on the value of K. The advantage of this second approach is its empirical basis.

An alternative approach for optimizing the estimation process comes from the urning updating scheme based on Pólya urn models (e.g., Mahmoud, 2008), that is currently under construction. In the urning updating scheme, an ability is expressed as the proportion of successes (an item is solved correctly) and failures (an item is solved incorrectly), which are represented by marbles in an urn. Likewise, items are represented by urns with marbles representing the successes and failures on the item. The proportion of these marbles indicates the item difficulty. Updating of abilities and difficulties takes place by randomly removing one of the marbles in the ability urn and the item urn and replacing it with the marble that represents the current success or failure. Urn sizes determine the precision of estimates and speed of updating the estimates. Although the bias-variance trade-off also needs to be optimized in this system by determining the urn sizes, the urning-system has the advantage that the distribution of the model parameters is known, which is not the case for the Elo-system. One of many possible extensions using the urning-scheme is the possibility to calculate confidence intervals around parameters of interest. These intervals can, for example, be used for inferences about individual differences as well as about individual developmental growth.

To conclude, a computer-adaptive learning environment requires a firm psychometric foundation. In Oefenweb this is accomplished through the close connection between Oe-

[1] The same problem is for example present in a model that uses predictions based on moving averages.

fenweb and the psychometrics research group at the University of Amsterdam.[2] Their largely overlapping research agendas result in an iterative process where insights from fundamental research are applied to the Oefenweb learning systems and problems that arise in these systems are worked on (and occasionally solved) by researchers of the psychometrics group. A good example is the more frequent use of A/B tests to optimize both the psychometric groundwork and the learning environment of students, as illustrated above. In the next section we zoom out even further and describe the crucial links between psychometrics, cognitive development and the educational practice.

8.3 Bridging Psychometrics, Cognitive Development and Education

Math Garden was originally developed to collect data for the study of the complex dynamics of the development of math skills. This resulted in a set-up with three different goals. First, at the student level we aimed to develop an engaging adaptive learning environment. The following quotes from children who played in Math Garden show their opinion on Math Garden.

> ‘Rekenen ga ik nooit leuk vinden maar door Rekentuin gaat het gewoon beter’. *I will never like maths, but with Math Garden I just do better.* Roos (student with dyscalculia)

> ‘Door het vaker te doen werd het [rekenen] leuker, door de goede resultaten te zien werd ik beter’. *By practicing more I liked maths more, and because of the positive feedback I learned more.* Jens (student)

Second, at the teacher level we aimed to free teachers from correcting workbooks and to provide them with detailed feedback on the learning process. The quotes below illustrate teachers’ evaluation of Math Garden.

> ’Wat ik een meerwaarde vind aan dit programma [Rekentuin] is de mogelijkheid om meerdere zintuigen tegelijkertijd te gebruiken en dat het zich aanpast aan het niveau van de leerlingen.’ *An advantage of Math Garden is that it can trigger different senses at the same time, and that it adapts to the level of the students.* (Teacher at the Frans Jozef Pryor school in Suriname)

> ’We houden iedere week bij waar de kinderen op vooruit zijn gegaan of eventueel op achteruit zijn gegaan en dat zie je in één oogopslag en dat is echt de kern. Daar kun je je les op inrichten.’ *With a quick look we can track the progress and decline of student on a weekly basis - that is the core. We can configure the next lessons based on this.* Egon Stroop (teacher of grade 5 (7 in the Dutch schoolsystem), De Zonnewijzer)

[2] See for example the interests of the different participating researchers at `https://www.oefenweb.nl/wetenschappelijke-partners/`

> 'Math Garden creates insight and urgency in their [students'] learning process.'
> Victor Ogola (research intern who explored the implementation of Oefenweb systems in public schools in Uganda)

Finally, at the researcher level we aimed to unlock a large time-intensive database to study learning and cognitive processes (Chapter 7 Straatemeier, 2014).

The citations of teachers and students and the popularity of Math Garden, used by about 2000 schools, suggest that the first two goals are met. The many research papers and this dissertation (the third on Math Garden) demonstrate the success in meeting the research goal. However, an overarching aim has emerged: the outcomes of research should ultimately result in improvements of the feedback to teachers and the learning environments for students. Ideally, teacher and student experiences again feed both fundamental and applied research. In the next section, we will highlight two different links between psychometrics, cognitive development, and education based on results of this dissertation.

Classification of (erroneous) strategies within domains Chapter 4 provides an example of a direct link between psychometrics and cognitive development. Based on the literature on multiplication we expected that both fast (retrieval strategies) and slow (back-up strategies, such a counting) would be used by children solving multiplication items. The use of the extended psychometric approach in this chapter allowed us to disentangle these processes. The results enable tailored feedback on proficiency of both the slow and the fast strategies when students learn multiplication. This matches the aims of Dutch education to both understand multiplication concepts (slow process) and to memorize the single-digit tables of multiplication (fast process) (*SLO*, 2009). An implementation of this extended measurement model in the Oefenweb systems will provide teachers information on, for example, the ability of children in both processes. This would allow teachers to act on this information by, for example, providing practice in automatizing the solution for specific item clusters (see for example Chapter 7).

Understanding the connection between learning in different domains Cognitive development can be conceptualized as a complex system of connected abilities (van der Maas et al., 2006a; van der Maas, Kan, Hofman, & Raijmakers, 2014). The results of Chapter 5 support the presence of positive connections between the development of different mathematical abilities measured by different games in Math Garden. Further explorations of connections are aimed at: (1) an even larger set of abilities, (2) testing differences in strength over time and (3) examining individual differences in the strength of the connections. Using this approach Ferrer et al. (2010) found that children with dyslexia have weaker connections between reading development and the development of IQ compared to typically developing peers. These results suggest that the strength of connections between developing domains can provide important information about the developmental trajectories of individual children.

To summarize, the present dissertation shows examples of the application of psychometric innovations that improve insights in cognitive development, learning processes, the learning environment of the student and information for teachers. Practice shows that teachers' wishes and viewpoints as well as children's learning behavior again require new psychometric innovations, closing the circle between psychometrics, cognitive development, and educational practice.

A Appendix to Chapter 3: The Balance-Scale Task Revisited

Estimation of the weighted-sum rule

The estimation of the weighted-sum rule is based on the optimization of the following set of equations, for each subject. First, the person vector θ_p is based on the implicit weight for the number-of-blocks ($\alpha_p > .5$) or distance ($\alpha_p < .5$) dimension and on the characteristics of the item set:

$$\begin{aligned} `_{\mathbf{p}} &= (\alpha_p \mathrm{wl} + (1-\alpha_p)\mathrm{dl}) - (\alpha_p \mathrm{wr} + (1-\alpha_p)\mathrm{dr}) \\ &= \alpha_p \Delta w_i + (1-\alpha_p)\Delta d_i . \end{aligned}$$

Based on the scaled $`_{\mathbf{p}}$ ($\sigma = 1$), and assuming a normal density to express the response probabilities, the log-likelihood of the response vector can be expressed as follows (person subscripts are dropped):

$$\begin{aligned} \log L(`, C) = \sum_{i=1}^{I} (R_i = l) \log \int_{-\infty}^{-C-\theta_i} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta_i)}{4}} dx + \\ (R_i = b) \log \int_{\infty}^{C-\theta_i} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta_i)}{4}} dx - \log \int_{\infty}^{-C-\theta_i} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta_i)}{4}} dx + \\ (R_i = r) \log \int_{C-\theta_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta_i)}{4}} dx, \end{aligned}$$

where the indicator terms, $R_i = l, b, r$, are one if the response is respectively left, balance or right and zero otherwise.

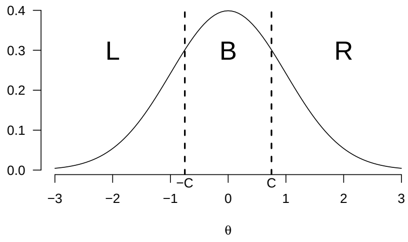


Figure A.1: The response probabilities expressed by the weighted-sum rule.

Figure A.1 shows a visual representation of these response probabilities. The set of functions is optimized with respect to α and C, using a constrained optimization implemented with the optim-function in Cran-R (R Core Team, 2013). α is constrained between zero and one, and C higher than zero. Note that (Wilkening & Anderson, 1982) also propose an implicit multiplication rule that can capture RIV. However it is not possible to estimate the parameters of this rule since the likelihood function is zero if α is zero, hence this rule will not be further studied.

Description of the selected items of the Math Garden data set

Table A shows the item description of the selected items of the Math Garden data set. We also included W, CWA, CD, CB and multiple-pin items of the types B, W and D - for multiple pin items the weights are placed on two pins on each side of the fulcrum. Responses on these item types are not included in the results for this paper. Items where constructed such that the variation of the product-differences within the set of items of the same type is comparable between the different types.

	Paper-and-Pencil						Math Garden					
Type	WL	DL	WR	DR	PD	Cor	WL	DL	WR	DR	PD	Cor
CBA	3	4	4	3	0	B	1	3	3	1	0	B
CBA	1	2	2	1	0	B	5	6	6	5	0	B
CBA	2	4	4	2	0	B	2	1	1	2	0	B
CBA	1	3	3	1	0	B	1	5	5	1	0	B
CDA	1	4	3	1	1	L	4	1	1	5	1	R
CDA	1	4	2	1	2	L	2	1	1	4	2	R
CDA	2	4	3	1	5	L	4	2	3	5	7	R
CDA	4	1	3	4	8	R	5	4	6	2	8	L
CW	1	3	2	2	1	R	3	3	1	6	3	L
CW	2	2	1	3	1	L	5	3	2	6	3	L
CW	3	3	2	4	1	R	2	5	1	6	4	L
CW	1	4	2	3	2	L	3	4	1	6	6	L
D	2	4	2	3	2	L	1	3	1	4	1	R
D	4	4	4	3	4	L	4	4	4	3	4	L
D	3	2	3	4	6	R	2	2	2	6	8	R
D	5	2	5	4	10	R	2	6	2	2	8	L

Note. The name of the item type refers to the dimension that determines the correct response; for example, in distance items, the beam goes down to the side with the largest distance between the peg with the weights and the fulcrum; WL, WR, DL and DR refer to respectively the number of blocks on the left and right side and the distance between the blocks and the fulcrum on the left and right side; PD = product-difference; Cor = correct response.

Table A.1: Item Characteristics of the Math Garden and Paper-and-Pencil dataset

Description of the LCM per item type

Type	Model	NC	BIC	1:p(BIC)	2:p(BIC)
D	Exploratory	2	2766	<.001	
		3	2661	.999	<.001
		4	2674	.001	
	Item Heterogeneity	3	2607		>.999
	Item Homogeneity	3(1)	2601		
		3(2)	2595		
		3(3)	2598		
	Final	3(1,2,3)	2581		
CW	Exploratory	2	4013	<.001	
		3	3899	>.999	<.001
		4	3934	<.001	
	Item Heterogeneity	3	3839		>.999
	Item Homogeneity	3(1)	3834		
		3(2)	3831		
		3(3)	3840		
	Final	3(1,2)	3826		
CDA	Exploratory	2	4047	<.001	
		3	3831	>.999	0.048
		4	3856	<.001	
	Item Heterogeneity	3	3825		0.952
	Item Homogeneity	3(1)	3837		
		3(2)	3864		
		3(3)	3925		
CBA	Exploratory	3	4280	<.001	
		4	4253	>.999	<.001
		5	4294	<.001	
	Item Heterogeneity	4	4233		>.999
	Item Homogeneity	4(1)	4221		
		4(2)	4224		
		4(3)	4240		
		4(4)	4227		
	Final	4(1,2,4)	4208		

Note. NC = the number of latent classes, the number between brackets refers to the class with constraints; p = BIC model probability of (1) select number of classes (2) select constraints; The best fitting model is underlined.

Table A.2: Paper-and-Pencil: Fit Results LCM per Item Type

Type	Model	NC	BIC	1:p(BIC)	2:p(BIC)
WD	Exploratory	1	2839	<.001	
		2	2749	>.999	
		3	2784	<.001	
D	Exploratory	1	1330	0.018	
		2	1322	0.982	<.001
		3	1367	<.001	
	Item Heterogeneity	2	1303		>.999
	Item Homogeneity	2(1)	1293		
		2(2)	1304		
CW	Exploratory	1	4906	<.001	
		2	4785	0.999	0.165
		3	4798	0.001	
	Item Heterogeneity	2	4782		0.835
	Item Homogeneity	2(1)	4793		
		2(2)	4840		
CDA	Exploratory	1	3247	<.001	
		2	3209	0.992	0.003
		3	3219	0.008	
	Item Heterogeneity	2	3197		0.997
	Item Homogeneity	2(1)	3186		
		2(2)	3213		
CBA	Exploratory	1	4817	<.001	
		2	4763	0.920	>.999
		3	4768	0.080	
	Item Heterogeneity	2	4858		<.001
	Item Homogeneity	–	–		

Note. NC = the number of latent classes, the number between brackets refers to the class with constraints; p = BIC weights (1) select number of classes (2) select constraints; The best fitting model is underlined.

Table A.3: Math Garden: Fit Results LCM per Item Type

B Appendix to Chapter 4: Fast versus Slow Multiplication

Simulating from the full-conditional distributions

The full-conditional distribution of the between dimension person covariance matrix $\mathbf{\Sigma}_\theta$ is easily sampled from:

$$f(\mathbf{\Sigma}_\theta) \propto \text{Inverse–Wishart}_{n-1}(\mathbf{S}_\theta),$$

where n refers to the sample size and $\mathbf{S}_\theta$ to the 'sample' covariance matrix $Cov(\boldsymbol{\theta})$. Similarly, we find that the full-conditional distributions for $\{\mu_{\beta,\,d}, \sigma^2_{\beta,\,d}\}$ are easy to sample from:

$$f(\mu_{\beta,\,d} \mid \sigma_{\beta,\,d}, \boldsymbol{\beta}_d) \propto \mathcal{N}(\bar{\boldsymbol{\beta}}_d, \sigma^2_{\beta,\,d}/k)$$
$$f(\sigma_{\beta,\,d} \mid \boldsymbol{\beta}_d) \propto \text{Inverse-}\chi^2(k-1, \sum_{i=1}^{k} \beta^2_{id}/(k-1)),$$

where k refers to the number of items in our analyses.

Unfortunately, the full-conditional distributions of the person and the item parameters are not readily sampled from. Standard approaches, such as the Metropolis within Gibbs approach of (Patz & Junker, 1999a, 1999b), are difficult to apply here due to the need of non-trivial fine-tuning that is required for each of the $n\times3$ person and $k\times3$ item parameters. This fine-tuning is particularly problematic as each of the persons responds to a possibly different set of items, and, similarly, each of the items has been responded to by a different set of persons.

To sample from the full-conditional distributions of the person and the item parameters we therefore utilize an independence chain Metropolis algorithm that was proposed by (Marsman, Maris, Bechger, & Glas, 2015). Their approach is particularly efficient when applied to the Rasch model and is simple to use with incomplete designs. [1]

Robustness Analysis

To investigate the stability of the comparison of the full fast-slow model with the more constrained versions of the model we constructed multiple data sets and replicated the analyses presented in the paper.

Data Selection

For the single-digit items, we constructed two data sets based on the selection of children that completed at least thirty items within one day *or* within one week. For the most-played items we selected data of children that completed at least 30 items within one day *or* one week *or* sixty items within one week. These choices resulted in a total of five different data sets. Within each data set, we selected items with a minimum of 200 responses, and looked at the child's first response to an item (multiple responses can be given to the same

[1] Details about this algorithm as applied to the Rasch model can be found in (pages 85–88 Marsman, 2014).

items	time	min items	N responses	N children	N items	% mis
single-digit	day	30	51,284	1,164	64	31
	week	30	180,651	3,551	64	21
most-played	day	30	387,882	7,403	135	61
	week	30	422,634	7,860	145	63
	week	60	490,874	4,813	147	31

Table B.1: Data description. The number of responses, children, items, and amount of missing data in the different constructed data sets.

item within a set of 30 or 60 items). The total number of responses, children, items and percentage of missing responses for each data set are presented in Table B.1

For each of the five data sets the response times were split using the overall median RT, within-person median RT and the within-item median RT. This resulted in a total of fifteen model comparisons.

Model Comparison

item selection	time frame	min items	RT split	model	ACC	RMSE	LL
single digit	day	30	overall median	full model	0.788	0.374	-657
				$\beta_{fast} = \beta_{slow}$	0.789	0.381	-684
				$\theta_{fast} = \theta_{slow}$	0.786	0.378	-674
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.786	0.380	-676
			within persons	full model	0.783	0.391	-725
				$\beta_{fast} = \beta_{slow}$	0.783	0.401	-788
				$\theta_{fast} = \theta_{slow}$	0.769	0.405	-783
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.783	0.398	-750
			within items	full model	0.804	0.374	-656
				$\beta_{fast} = \beta_{slow}$	0.802	0.377	-690
				$\theta_{fast} = \theta_{slow}$	0.783	0.382	-685
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.784	0.383	-685
	week	30	overall median	full model	0.777	0.391	-2416
				$\beta_{fast} = \beta_{slow}$	0.775	0.397	-2510
				$\theta_{fast} = \theta_{slow}$	0.773	0.398	-2518
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.772	0.397	-2489
			within persons	full model	0.776	0.387	-2353
				$\beta_{fast} = \beta_{slow}$	0.764	0.400	-2493
				$\theta_{fast} = \theta_{slow}$	0.756	0.402	-2513
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.769	0.397	-2465
			within items	full model	0.783	0.391	-2415
				$\beta_{fast} = \beta_{slow}$	0.778	0.397	-2519
				$\theta_{fast} = \theta_{slow}$	0.772	0.397	-2498
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.775	0.396	-2473

Table B.2: Model fit based on cross-validation of the full and three constrained fast-slow models in the single digit data set.

item selection	time frame	min items	RT split	model	ACC	RMSE	LL
most played	day	30	overall median	full model	0.760	0.410	-5131
				$\beta_{fast} = \beta_{slow}$	0.740	0.425	-5489
				$\theta_{fast} = \theta_{slow}$	0.750	0.417	-5287
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.747	0.417	-5286
			within persons	full model	0.766	0.404	-5025
				$\beta_{fast} = \beta_{slow}$	0.761	0.409	-5150
				$\theta_{fast} = \theta_{slow}$	0.752	0.415	-5283
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.750	0.417	-5328
			within items	full model	0.761	0.409	-5141
				$\beta_{fast} = \beta_{slow}$	0.757	0.415	-5342
				$\theta_{fast} = \theta_{slow}$	0.756	0.414	-5263
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.753	0.415	-5268
	week	30	overall median	full model	0.750	0.416	-5239
				$\beta_{fast} = \beta_{slow}$	0.740	0.422	-5403
				$\theta_{fast} = \theta_{slow}$	0.742	0.420	-5351
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.737	0.421	-5375
			within persons	full model	0.750	0.413	-5168
				$\beta_{fast} = \beta_{slow}$	0.748	0.415	-5225
				$\theta_{fast} = \theta_{slow}$	0.736	0.421	-5363
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.730	0.425	-5449
			within items	full model	0.747	0.417	-5262
				$\beta_{fast} = \beta_{slow}$	0.734	0.425	-5436
				$\theta_{fast} = \theta_{slow}$	0.744	0.422	-5385
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.740	0.422	-5379
	week	60	overall median	full model	0.777	0.391	-2416
				$\beta_{fast} = \beta_{slow}$	0.775	0.397	-2510
				$\theta_{fast} = \theta_{slow}$	0.773	0.398	-2518
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.772	0.397	-2489
			within persons	full model	0.776	0.387	-2353
				$\beta_{fast} = \beta_{slow}$	0.764	0.400	-2493
				$\theta_{fast} = \theta_{slow}$	0.756	0.402	-2513
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.769	0.397	-2465
			within items	full model	0.783	0.391	-2415
				$\beta_{fast} = \beta_{slow}$	0.778	0.397	-2519
				$\theta_{fast} = \theta_{slow}$	0.772	0.397	-2498
				$\beta_{fast} = \beta_{slow}$ & $\theta_{fast} = \theta_{slow}$	0.775	0.396	-2473

Table B.3: Model fit based on cross-validation of the full and three constrained fast-slow models in the most played items data set.

C Appendix to Chapter 6: The Dynamics of Development

Description of Math Garden

Data was collected with an adaptive learning program for mathematics called Math Garden (www.mathsgarden.com; Klinkenberg et al., 2011; Straatemeier, 2014). After logging in, children arrive at a page showing a garden, where each plant represents a game that covers a particular math domain. A game starts after selecting the corresponding plant. Children are given 15 items to solve. They respond by clicking on a response option (counting and addition) or by using the virtual numeric keypad (multiplication and division). Each item is presented with a time limit of 20 seconds. The time is visualized by disappearing coins (one is lost each second that a response is not provided). If a correct response is given the coins are added to a money bag, but they are subtracted if the response is incorrect. The scoring rule rewards fast correct responses, but also punishes fast incorrect responses. The score (S) that follows from this model is defined as follows:

$$S = (2X_{pi} - 1)(d - T_{pi}), \tag{C.1}$$

with the following expected scores, based on the current θ en β estimates:

$$E(S|\theta,\beta) = d\frac{exp(2d(\theta-\beta))+1}{exp(2d(\theta-\beta))-1} - \frac{1}{\theta-\beta} \tag{C.2}$$

This explicit scoring-rule – called the High Speed High Stakes (HSHS) scoring rule (Klinkenberg et al., 2011; Maris & van der Maas, 2012) – informs players on how to weigh speed and accuracy. This ensures that children perceive some time-pressure and are motivated to provide fast responses. But, they are also discouraged from guessing due to the penalty on a fast incorrect response. When a child does not know the answer (s)he can best wait the full 20 seconds. To prevent such waiting times the child can also use the question-mark button, in which case (s)he does not win or lose any coins.

With the HSHS scoring rule in the Math Garden the estimates of both the user ability and the item difficulty can be updated after each response using an Elo updating scheme (Elo, 1978):

$$\theta_p = \theta_p + K_p(S_{pi} - E(S)_{pi}) \tag{C.3}$$

$$\beta_i = \beta_i - K_i(S_{pi} - E(S)_{pi}). \tag{C.4}$$

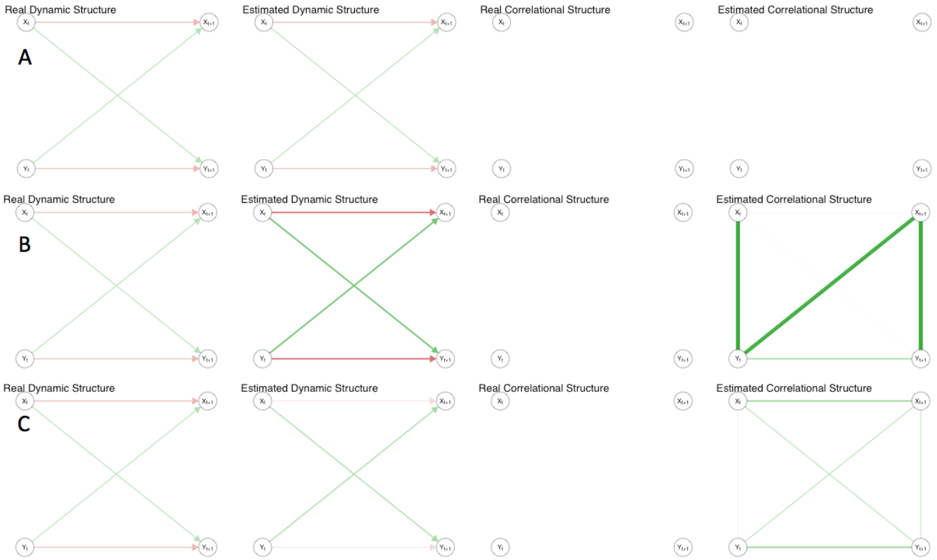
This can be seen as a dynamic system where both θ and β change can change over time. K is a smoothing parameter that determines the variance in the ratings in the system. See Klinkenberg et al. (2011), Pelánek (2016) and Brinkhuis and Maris (2009) for more details about Elo in the context of adaptive testing.

Based on these estimates, relevant items are selected for each player at each time-point, such that children were expected to provide 60, 75 or 90% correct responses when playing at the hard, medium or easy difficulty level (for more details see Jansen et al., 2016). For the current study we only selected children who played at the medium and hard difficulty levels and for which the average proportion of correct responses were around the aimed proportions, respectively .60 and .75. This ensured that the analyzed ability estimates were reliable indicators of the true skills at different periods.

Simulation Study

We performed different simulation studies to investigate two issues. First, we tested the recovery of the estimated coupling (γ) and self-feedback (β) parameters and of the correlational structure on the latent change factors ($\Delta X_t, \Delta X_{t+1}, \Delta Y_t, \Delta Y_{t+1}$) under different scenario's. Second, we investigated the model comparison of the bidirectional coupling model (the full model), the no-coupling model and the g-factor model with varying strengths of the coupling parameters.

Parameter recovery



Real Dynamic Structure
A
Estimated Dynamic Structure
Real Correlational Structure
Estimated Correlational Structure
Real Dynamic Structure
B
Estimated Dynamic Structure
Real Correlational Structure
Estimated Correlational Structure
Real Dynamic Structure
C
Estimated Dynamic Structure
Real Correlational Structure
Estimated Correlational Structure
X$_t$
X$_{t+1}$
Y$_t$
Y$_{t+1}$

Figure C.1: The results of the simulation study under three different scenario's: A = true model; B = selection of time-points; C = selection of variables. Here you see a simplified depiction, where only the relevant part of the full bivariate LCSM is shown. The circles represent change scores of variables X and Y at times T and T+1. The first column shows the coupling (green; indicating a positive effect) and self-feedback (red; indicating a negative effect) effects of the true model and the second column shows the average estimated effects over the 500 runs. The third and fourth columns show the true correlational structure (zero in all true models) and estimated correlation structure respectively between the change factors. We set the maximum value of the edges to .7 to clearly depict the difference between the estimated model parameters.

We tested the parameter recovery of the correlations between the change scores in the bivariate latent change score model under three different scenarios: (1) parameter recovering in the true model, (2) selection of time-points, and (3) subset of variables. 500 data sets were simulated for each scenario with a bivariate dual latent change score model of three time-points and 5,000 participants. These settings were chosen to resemble the configurations of the data sets and the estimated model parameters of the current study. In the dynamic structure, the coupling parameters (γ) and the feedback parameters (β) were respectively set to .2 and $-.2$, the average of change score factors (Δ) to .5 with a variance of .3. To test the effects of the different scenarios on the correlational structure we set the correlations between the change factors (correlational structure) to zero in the true model. The variance of the scores at t1 were .5 and the error variances of all time-points were .1. The model specifications were equal to the models presented in section6, but to simplify the model we omitted the in this case unnecessary covariates (age and counts) at T1 because.

Scenario A: parameter recovery in the true model First, we investigated the recovery of the parameters under the true model. The results showed that under the specifications both the parameters of the dynamic and correlational structure we used were perfectly recovered, see upper-panel of Figure C.1.[1]

Scenario B: selection of time-points Second, we investigated the effect of collecting a subset of the time-points on which the real developmental process unfold. To this end we generated data sets with seven time-points while only three time-points were observed and analyzed (the first, fourth and seventh time-point). Hence, in the true model multiple iterations were present between the dynamical structure of both coupled variables between the observed time-points. A comparison of the real and the estimated parameters in the dynamical structure (see middle-panel of Figure C.1) shows that, as expected, these parameters were overestimated (on average the estimated parameters are around twice as large as the true parameters). More importantly, also the correlational structure is largely overestimated (with an average correlation of .32; varying between .03 and .5). This indicated that this correlational structure can merely result from a data selection problem, and an inflation of these correlations are expected if only a subset of the time-points of the true dynamical system are analyzed.

Scenario C: subset of variables Third, we investigated the effect of collecting just a subset of all variables that thrive the full dynamical system. Therefore, we again simulated 500 data sets based on a LSCM with four variables with the same configurations as discussed before. As expected, in the dynamical system the self-feedback parameters were underestimated (-.1 instead of -.2) and the coupling parameters were overestimated (.25 instead of .2). This can be explained by the positive relation with other variables that were not selected, although these variables did drive the changes scores upwards in the data simulation. Also, the correlational structure between the change scores was overestimated (correlations ranging from .06 to .24 with an average of .16). This simulation indicated that a significant correlational structure between the change scores in the LSCM can be caused

[1] The plot was made with the qgraph package (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) in *R*

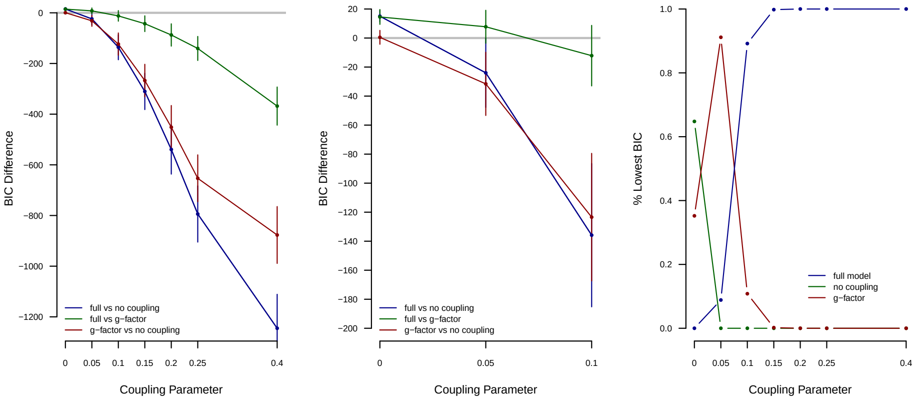


Figure C.2: Model comparison using BIC differences based on simulated data sets of the bidirectional coupling model with varying coupling parameters.

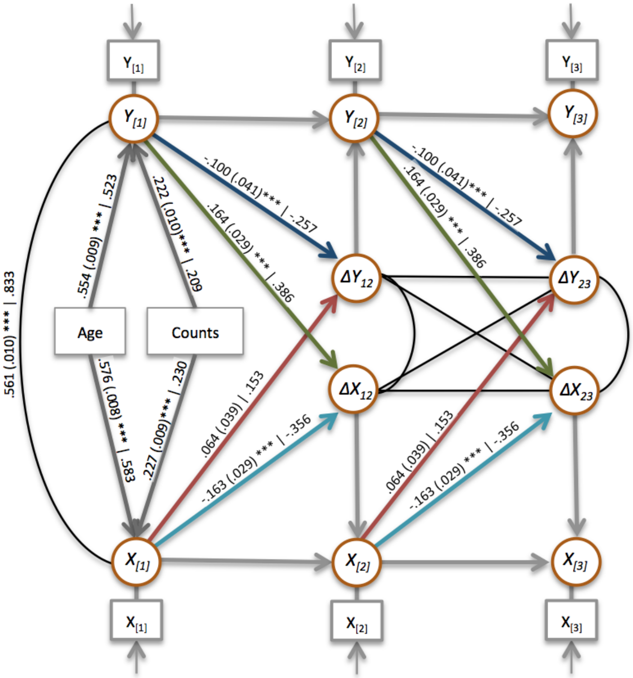
by the analyses of just a small subset of the possibly larger set of relevant variables that thrive the changes in the dynamical system.

To conclude, the model shows very good parameter recovery under the true model, but both a selection of time-points and including only a subset of all variables of the real developmental system will result in an inflated correlational structure on the change scores variables. These results should be taken into account in the interpretation of such a correlational structure in a real data example. Crucially, the results of Scenario A and B show that under a true model with (positive) mutualistic effects and no correlations between change scores, a selection of time-points or than analyses of a subset of all variables results in an overestimation of these correlations. This shows that in addition to a *g*-factor explanation of the correlational structure, these correlation can also be caused by a mutualistic account on development.

Model comparison

In a second simulation study we investigated the model comparison between the bidirectional coupling model (the full model), the no-coupling model and the *g*-factor model with varying strengths of the coupling parameters. We used the same set-up as in the previous simulation study, but with varying coupling parameters. These were sequentially set to 0, .05, .1, .15, .2, .25 and .4, ranging from absent to moderately high. This resulted in seven sets of 500 data sets. For each set we calculated the differences between AIC and BIC values for the comparison of the full model with coupling parameters, constrained model without coupling parameters and the *g*-factor model. The results are presented in Figure C.2. The left-panel plot shows that, as expected, the differences between the BIC values were higher for higher coupling parameters (negative differences indicates that the first model has a lower BIC values than the second model and thus results in a better fit). The vertical lines indicate the standard deviations (2 * SD) of the differences. Both the full and the *g*-factor model clearly outperform the no coupling model when the coupling parameter is higher than or equal to .1. This indicated that under the utilized data generation we have enough power to reliably classify coupling parameters of .1 as significantly different from zero. Also, as expected, the full model outperforms the *g*-factor model for high values of the coupling parameters. A close look at the results of the smaller coupling parameters (middle-panel) indicates that if the true coupling parameters are zero the BIC differences indicate that the full model is, as expected, outperformed by both the no-coupling and the *g*-factor model. Additionally, the BIC differences between the no-coupling model and the *g*-factor model are distributed around zero, which shows that in this case both models result in a comparable fit. This is also shown in the right-panel plot of Figure C.2: if the coupling parameters were set to zero, in about 60% of the runs the no-coupling model outperforms the other models, whereas the *g*-factor model is the best model in the other 40% of the runs. An unexpected result was the high performance of the *g*-factor model, outperforming the other models in 90% of the runs, when the coupling parameters were set to .05. Thus under these settings the *g*-factor model outperforms both the no-coupling model and the full model. To conclude, both the no-coupling model and the *g*-factor model provide a good baseline model to test the presence of mutualistic coupling parameters and that the full model outperforms the other models under the current settings of the data simulation if such coupling is introduced (coupling values above .1).

Model parameters multiplication and division



Means:
$X_{[1]}$ = -.115
ΔX_{12} = .559 (.010)***
ΔX_{23} = .449 (.012)***
$Y_{[1]}$ = -.427
ΔY_{12} = .438 (.013)***
ΔY_{23} = .388 (.014) ***

Correlations Δ:
$COR(\Delta X_{12}, \Delta Y_{12})$ = .120 (.006) *** |.657
$COR(\Delta X_{23}, \Delta Y_{23})$ = .110 (.006) *** |.638
$COR(\Delta X_{12}, \Delta X_{23})$ = .024 (.006) *** |.120
$COR(\Delta Y_{12}, \Delta Y_{23})$ = .031 (.008) *** |.199
$COR(\Delta X_{12}, \Delta Y_{23})$ = .027 (.006) *** |.163
$COR(\Delta Y_{12}, \Delta X_{23})$ = .019 (.006) ** |.104

Variances:
$X_{[1]}$ = .592 (.012) *** | .607
ΔX_{12} = .197 (.013) *** | .969
ΔX_{23} = .205 (.016) *** | .974
$Y_{[1]}$ = .766 (.016) *** | .683
ΔY_{12} = .168 (.023) *** | .981
ΔY_{23} = .144 (.025) *** | .996

$X_{[1-3]}$ = .033 (.008) ***
$Y_{[1-3]}$ = .026 (.013) *

Figure C.3: The estimated model parameters for the best fitting mutualism model in the multiplication and division data set. The first value denotes the unstandardized coefficient, with the standard error between brackets, and the value after the vertical bar denotes the standardized coefficient. The stars indicate the significant levels (* = p <.05, ** = p < 0.01 and *** = p < .001). The observed scores (X and Y) are the latent trait estimates provided by Math Garden.

REFERENCES

Adolph, K. E., Robinson, S. R., Young, J. W., & Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychological review*, *115*(3), 527. 2, 111, 165

Akin, O., & Chase, W. (1978). Quantification of three-dimensional structures. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 397. 12

Ambrose, R., Baek, J., & Carpenter, T. P. (2003). Children's invention of multidigit multiplication and division algorithms. *The development of arithmetic concepts and skills: Constructive adaptive expertise*, 305–36. 123

Anderson, M. (2017). Binet's error: Developmental change and individual differences in intelligence are related to different mechanisms. *Journal of Intelligence*, *5*(2), 24. 89

Ashby, F. G., & Alfonso-Reese, L. A. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442. doi: 10.1037/0033-295X.105.3.442 52

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*, 149–178. doi: 10.1146/annurev.psych.56.091103.070217 51

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161. doi: 10.1080/17405620344000031 30

Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. *Mathematical cognition*, *1*, 3–34. 57

Ashcraft, M. H., & Guillaume, M. M. (2009). Mathematical cognition and the problem size effect. *Psychology of learning and motivation*, *51*, 121–151. 54, 57, 123

Ashkenazi, S., Mark-Zigdon, N., & Henik, A. (2013). Do subitizing deficits in developmental dyscalculia involve pattern recognition weakness? *Developmental Science*, *16*(1), 35–46. 25, 26

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *96*(4), 699. 54, 55, 90

Balakrishnanl, J., & Ashby, F. G. (1992). Subitizing: Magical numbers or mere superstition? *Psychological research*, *54*(2), 80–90. 13, 25

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, *128*(4), 612. 103

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, thomson and bartlett. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 569–582. doi: 10.1348/000711008X365676 89

Batchelder, W. H., & Bershad, N. J. (1979). The statistical analysis of a thurstonian model for rating chess players. *Journal of Mathematical Psychology*, *19*(1), 39–60. 5, 168

Benoit, L., Lehalle, H., & Jouen, F. (2004). Do young children acquire number words through subitizing or counting? *Cognitive Development*, *19*(3), 291–307. 13, 26

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance: Classes, strategies, or rules for the balance scale task? *Cognitive Development*, *16*(2), 717–735. doi: 10.1016/S0885-2014(01)00056-9 35

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review*, *110*(2), 203. 89

Bouwmeester, S., Sijtsma, K., & Vermunt, J. K. (2004). Latent class regression analysis for describing cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology*, *1*(1), 67–86. doi: 10.1080/17405620344000031 36

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339–353. 75

Brinkhuis, M. J. S., Bakker, M., & Maris, G. (2015). Filtering data for detecting differential development. *Journal of Educational Measurement*, *52*(3), 319–338. 95

Brinkhuis, M. J. S., & Maris, G. (2009). Dynamic parameter estimation in student monitoring systems. *Measurement and Research Department Reports (Rep. No. 2009-1). Arnhem: Cito*. 146

Buitrago, M. M., Schulz, J. B., Dichgans, J., & Luft, A. R. (2004). Short and long-term motor skill learning in an accelerated rotarod training paradigm. *Neurobiology of learning and memory*, *81*(3), 211–216. 73

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D. Berliner. (Ed.), *Review of research in education.* Washington, DC: American Educational Research Association. 73

Campbell, J. I. D., & Austin, S. (2002). Effects of response time deadlines on adults' strategy choices for simple addition. *Memory & Cognition*, *30*(6), 988–994. 127

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59–68. 26

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for research in Mathematics Education*, 179–202. 103

Carr, M., & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy use: Social and metacognitive influences. *Journal of Educational Psychology*, *89*, 318. 57, 68

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Evolution of teaching and learning paradigms in intelligent environment*, 183–221. 106

Cattell, R. B. (1971). *Abilities: Their structure, growth and action*. Houghton-Mifflin, Boston. 90

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *NJ: Lawrence Earlbaum Associates*, *2*. 83

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons. 39

Coomans, F., Hofman, A. D., Brinkhuis, M. J. S., van der Maas, H. L. J., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PloS one*, *11*(5), e0155149. 69, 107, 128

Crump, M. J. C., & Logan, G. D. (2010a). Episodic contributions to sequential control: Learning from a typist's touch. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 662. 72

Crump, M. J. C., & Logan, G. D. (2010b). Hierarchical control and skilled typing: evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1369. 72

Curran, P. J., Howard, A. L., Bainter, S. A., Lane, S. T., & McGinley, J. S. (2014). The separation of between-person and within-person components of individual change over time: a latent curve model with structured residuals. *Journal of consulting and clinical psychology*, *82*(5), 879. 93

Dandurand, F., & Shultz, T. R. (2009). Modeling acquisition of a torque rule on the balance-scale task. In *Proceedings of the 31st annual conference of the cognitive science society. austin, tx: Cognitive science society* (pp. 1541–6). 52

Dandurand, F., & Shultz, T. R. (2013). A comprehensive model of development on the balance-scale task. *Cognitive Systems Research*. doi: doi.org/10.1016/j.cogsys.2013.10.001 31, 35, 52

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559. 56

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28. 54, 56

De Brauwer, J., Verguts, T., & Fias, W. (2006). The representation of multiplication facts: Developmental changes in the problem size, five, and tie effects. *Journal of Experimental Child Psychology*, *94*, 43–56. 57

De Groot, A. D. (1978). *Thought and choice in chess* (Vol. 4). The Hague, The Netherlands: Mouton (Original work published in 1946). 27

Dehaene, S. (1987). *The number sense*. Oxford, UK: University Press. 26

Dehaene, S., & Cohen, L. (1994). Dissociable mechanisms of subitizing and counting: neuropsychological evidence from simultanagnosic patients. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 958. 12

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624. 55

Demeyere, N., Lestou, V., & Humphreys, G. W. (2010). Neuropsychological evidence for a dissociation in counting and subitizing. *Neurocase*, *16*(3), 219–237. 12

De Smedt, B., Holloway, I. D., & Ansari, D. (2011). Effects of problem size and arithmetic operation on brain activation during calculation in children with varying levels of arithmetical fluency. *Neuroimage*, *57*, 771–781. 55

Desoete, A., Ceulemans, A., Roeyers, H., & Huylebroeck, A. (2009). Subitizing or counting as possible screening variables for learning disabilities in mathematics education or learning? *Educational Research Review*, *4*(1), 55–66. 26

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*, 2380. 55

DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence*, *56*, 82–92. 55, 68

Dowker, A. (2005). *Individual differences in arithmetic: Implications for psychology, neuroscience and education*. East Sussex: Psychology Press. 3, 54, 123, 166

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica: Revista de metodología y psicología experimental*, *32*, 107–132. 58

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub. 5, 20, 94, 107, 146, 168

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. 149

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107. doi: 10.1037/0096-3445.127.2.107 52

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307–314. 26

Fendrick, P. (1937). Hierarchical skills in typewriting. *Journal of Educational Psychology*, *28*(8), 609. 73

Ferrer, E., & McArdle, J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling*, *10*(4), 493–524. 94

Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental psychology*, *40*(6), 935–951. 94

Ferrer, E., & McArdle, J. J. (2010). Longitudinal modeling of developmental changes in psychological research. *Current Directions in Psychological Science*, *19*(3), 149–154. 91

Ferrer, E., Shaywitz, B. A., Holahan, J. M., Marchione, K., & Shaywitz, S. E. (2010). Uncoupling of reading and iq over time. *Psychological Science*, *21*(1), 93-101. 103, 131

Ferretti, R. P., & Butterfield, E. C. (1986). Are children's rule-assessment classifications invariant across instances of problem types? *Child Development*, *57*(6), pp. 1419-1428. doi: 10.2307/1130420 34, 51

Ferretti, R. P., & Butterfield, E. C. (1992). Intelligence-related differences in the learning, maintenance, and transfer of problem-solving strategies. *Intelligence*, *16*(2), 207–223. doi: 10.1006/drev.1997.0437 34

Ferretti, R. P., Butterfield, E. C., Cahn, A., & Kerkman, D. (1985). The classification of children's knowledge: Development on the balance-scale and inclined-plane tasks. *Journal of Experimental Child Psychology*, *39*(1), 131–160. doi: 10.1016/0022-0965(85)90033-5 33, 34

Finnie, R., & Meng, R. (2006). *The importance of functional literacy: Reading and math skills and labour market outcomes of high school drop-outs* (Tech. Rep.). Statistics Canada, Analytical Studies Branch. 106

Fischer, B., Köngeter, A., & Hartnegg, K. (2008). Effects of daily practice on subitizing, visual counting, and basic arithmetic skills. *Optometry & Vision Development*, *39*(1). 27

Fischer, K. W., & Silvern, L. (1985). Stages and individual differences in cognitive development. *Annual Review of Psychology*, *36*(1), 613–648. doi: 10.1146/annurev.psych.36.1.613 51

Freudenthal, H. (1991). *Revisiting mathematics education*. Dordrecht, the Netherlands: Kluwer. 54

Geary, D. C., Brown, S. C., & Samaranayake, V. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. *Developmental psychology*, *27*(5), 787. 88

Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology*, *88*(2), 121 - 151. doi: http://dx.doi.org/10.1016/j.jecp.2004.03.002 88

Geary, D. C., Widaman, K. F., & Little, T. D. (1986). Cognitive addition and multiplication: Evidence for a single memory network. *Memory & Cognition*, *14*, 478–487. 54

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383. 111

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., … Dorie, V. (2009). *arm: Data analysis using regression and multilevel/hierarchical models (r package, version 9.01).* 111

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number.* Harvard University Press. 26

Gentner, D. R., Larochelle, S., & Grudin, J. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. *Cognitive Psychology*, *20*(4), 524–548. 73

Ghisletta, P., & De Ribaupierre, A. (2005). A dynamic investigation of cognitive dedifferentiation with control for retest: evidence from the swiss interdisciplinary longitudinal study on the oldest old. *Psychology and aging*, *20*(4), 671. 92

Gierasimczuk, N., van der Maas, H. L. J., & Raijmakers, M. E. J. (2013). An analytic tableaux model for deductive mastermind empirically tested with a massively used online learning system. *Journal of Logic, Language and Information*, *22*(3), 297–314. 6, 169

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence*, *42*, 89–97. 90, 91, 97

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(3), 377–394. 5, 168

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). West Sussex, UK: Wiley. 75, 78, 86

Groeneveld, C. M. (2014). Implementation of an adaptive training and tracking game in statistics teaching. In *International computer assisted assessment conference* (pp. 53–58). 2

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. 88

Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013). Topographic representation of numerosity in the human parietal cortex. *Science*, *341*(6150), 1123–1126. 12

Hill, L. B., Rejall, A. E., & Thorndike, E. L. (1913). Practice in the case of typewriting. *The Journal of Genetic Psychology*, *20*, 516–529. 73

Hofman, A. D., Visser, I., Jansen, B. R. J., Marsman, M., & van der Maas, H. L. J. (Submitted). Fast and slow strategies in multiplication.
doi: osf.io/aw3qq 94, 107, 128

Hofman, A. D., Visser, I., Jansen, B. R. J., & van der Maas, H. L. J. (2015). The balance-scale task revisited: a comparison of statistical models for rule-based and information-integration theories of proportional reasoning. *PloS one*, *10*, e0136449. 68

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. Routledge. 75

Hoyles, C., Wolf, A., Molyneux-Hodgson, S., & Kent, P. (2002). Mathematical skills in the workplace: final report to the science technology and mathematics council.
106

Huang, G.-H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, *69*(1), 5-32. doi: 10.1007/BF02295837 37

Imbo, I., vandierendonck, A., & Rosseel, Y. (2007). The influence of problem features and individual differences on strategic performance in simple arithmetic. *Memory & Cognition*, *35*, 454–463. 54

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. *London, Kegan Paul*. 30, 31, 125

Jansen, B. R. J., Hofman, A. D., Savi, A., Visser, I., & van der Maas, H. L. J. (2016). Self-adapting the success rate when practicing math. *Learning and Individual Differences*, *51*, 1–10. 107, 110, 146

Jansen, B. R. J., Hofman, A. D., Straatemeier, M., Bers, B. M. C. W., Raijmakers, M. E. J., & van der Maas, H. L. J. (2014). The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology*, *32*(2), 178–194. 94

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*(3), 321–357. doi: 10.1006/drev.1997.0437 33, 34, 35, 50, 51

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology*, *81*(4), 383–416. doi: 10.1006/jecp.2002.2664 30, 31, 33, 34, 35, 38, 47, 50, 51, 125

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport CT USA: Praeger Publishers / Greenwood Publishing Group Inc. 8, 90, 170

John, B. E. (1996). Typist: A theory of performance in skilled typing. *Human-Computer Interaction*, *11*(4), 321–355. 72

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, *22*(1), 36–46. 12, 27

Jost, K., Beinhoff, U., Hennighausen, E., & Rösler, F. (2004). Facts, rules, and strategies in single-digit multiplication: evidence from event-related brain potentials. *Cognitive Brain Research*, *20*, 183–193. 55

Julious, S. A. (2001). Inference and estimation in a changepoint regression problem. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(1), 51–61. 13

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan. 30

Kan, K.-J. (2012). *The nature of nurture: the role of gene-environment interplay in the development of intelligence* (Doctoral dissertation, University of Amsterdam). Retrieved from `http://hdl.handle.net/11245/1.392628` 103

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498–525. 12

Kievit, R. A., Brandmaier, A., Ziegler, G., van Harmelen, A.-L., de Mooij, S., Moutoussis, M., . . . Dolan, R. (2017). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *bioRxiv*. doi: 10.1101/110429 91

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., & Dolan, R. J. (2017). Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood. *Psychological Science*, 10.1177/0956797617710785. doi: 10.1177/0956797617710785 89, 101, 103, 104

Kirk, E. P., & Ashcraft, M. H. (2001). Telling stories: the perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 157. 55

Klahr, D., & Siegler, R. S. (1978). The representation of children's knowledge. *Advances in Child Development and Behavior*, *12*, 61–116. 12, 30, 31

Klinkenberg, S. (2014). High speed high stakes scoring rule. In *International computer assisted assessment conference* (pp. 114–126). 2

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. doi: :10.1016/j.compedu.2011.02.003 2, 4, 5, 6, 19, 20, 21, 25, 39, 55, 74, 89, 94, 106, 107, 115, 146, 165, 167, 168, 169

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 333–353. 106

Kroesbergen, E. H., van Luit, J. E. H., van Lieshout, E. C. D. M., van Loosbroek, E., & van de Rijt, B. A. M. (2009). Individual differences in early numeracy: The role of executive functions and subitizing. *Journal of Psychoeducational Assessment*, *27*(3), 226–236. 12, 27

Kruis, J., & Maris, G. (2016). Three representations of the ising model. *Scientific Reports*, *6*, 1–11. 89

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., & Carver, S. M. (1995). Strategies of knowledge acquisition. *Monographs of the society for research in child development*, i–157. 123

Lashley, K. S. (1951). The problem of serial order in behavior. In J. L. A. (Ed.), *Cerebral mechanisms in behavior* (p. 112-136). New York: Wiley. 72

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*, 284. 54, 55, 57

Lemaire, P. (2010). Executive functions and strategic aspects of arithmetic performance: The case of adults' and children's arithmetic. *Psychologica Belgica*, *50*(3-4). 123

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, *124*, 83. 54, 55, 57, 123

Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional rasch analysis of gender differences in pisa mathematics. *Journal of applied measurement*, *9*, 18. 54, 55

Logan, G. D. (2003). Simon-type effects: chronometric evidence for keypress schemata in typewriting. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(4), 741. 72

Logan, G. D., & Crump, M. J. C. (2009). The left hand doesn't know what the right hand is doing: The disruptive effects of attention to the hands in skilled typewriting. *Psychological Science*, *20*(10), 1296–1300. 72

Logan, G. D., & Crump, M. J. C. (2011). Hierarchical control of cognitive processes: The case for skilled typewriting. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 1. 72, 86

Logan, G. D., & Zbrodoff, N. J. (1998). Stroop-type interference: Congruity effects in color naming with typewritten responses. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 978. 72

Lövdén, M., Ghisletta, P., & Lindenberger, U. (2004). Cognition in the berlin aging study (base): the first 10 years. *Aging Neuropsychology and Cognition*, *11*(2-3), 104–133. 104

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, *46*(1), 1–27. 31

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of experimental psychology: learning, memory, and cognition*, *29*(4), 650. doi: 10.1037/0278-7393.29.4.650 51

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*(5), 945–952. doi: 10.3758/BF03196726 51

Maddox, W. T., & David, A. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of experimental psychology: learning, memory, and cognition*, *31*(1), 100. doi: 10.1037/0278-7393.31.1.100 51

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2017). cluster: Cluster analysis basics and extensions [Computer software manual]. (R package version 2.0.6) 119

Mahmoud, H. (2008). *Pólya urn models*. CRC press. 129

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*(1), 1. 12, 13, 25, 26

Maniar, H. S., Council, M. L., Prasad, S. M., Prasad, S. M., Chu, C., & Damiano, R. J. (2005). Comparison of skill training with robotic systems and traditional endoscopy: implications on training and adoption. *Journal of Surgical Research*, *125*(1), 23–29. 73

Maris, G., & van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633. 4, 5, 20, 21, 38, 74, 94, 107, 115, 127, 146, 167, 168

Marsman, M. (2014). *Plausible Values in statistical inference* (Doctoral dissertation, University of Twente, Enschede, the Netherlands). doi: 10.3990/1.9789036537445 140

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank ising networks. *Scientific reports*, *5*. 140

Maylor, E. A., Watson, D. G., & Hartley, E. L. (2011). Effects of distraction on visual enumeration in children and adults. *Developmental psychology*, *47*(5), 1440. 13, 26

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In C. R., duToit S., & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (p. 342-380). Lincolnwood: Scientific Software International. 91, 92

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual review of psychology*, *60*, 577–605. 91

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental psychology*, *38*(1), 115. 89, 90

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In M. G. M. Morris (Ed.), *Parallel distributed processing, implications for psychology and neurobiology* (pp. 8–45). Oxford: Clarendon Press. 31

McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In T. J. Simon & G. S. Halford

(Eds.), *Developing cognitive competence: New approaches to process modeling*. Lawrence Erlbaum Associates. 31, 51

McCutcheon, A. L. (1987). *Latent class analysis* (No. 64). Sage. 35

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 46. 57

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence*, *38*(6), 611–624. 90, 97

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218. 2, 88, 89, 106, 165

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in medicine*, *22*(19), 3055–3071. 13

Murnane, R. J., John, B. W., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics*, *77*(2), 251–266. 88

Musch, J., & Grondin, S. (2001). Unequal competition as an impediment to personal development: A review of the relative age effect in sport. *Developmental review*, *21*(2), 147–167. 103

Muthén, B. O., Kao, C.-F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new irt-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1–22. doi: 10.1111/j.1745-3984.1991.tb00340.x 90

Nan, Y., Knösche, T. R., & Luo, Y.-J. (2006). Counting in everyday life: Discrimination and enumeration. *Neuropsychologia*, *44*(7), 1103–1113. 12

Normandeau, S., Larivée, S., Roulin, J.-L., & Longeot, F. (1989). The balance-scale dilemma: Either the subject or the experimenter muddles through. *The Journal of genetic psychology*, *150*(3), 237–250. doi: 10.1080/00221325.1989.9914594 33

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*, 23–32. 55, 128

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. *Journal of educational and behavioral statistics*, *24*, 342–366. 140

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of educational and behavioral Statistics*, *24*, 146–178. 140

Pelánek, R. (2015). Metrics for evaluation of student models. *JEDM-Journal of Educational Data Mining*, *7*, 1–19. 59

Pelánek, R. (2016). Applications of the elo rating system in adaptive educational systems. *Computers & Education*, *98*, 169–179. 5, 146, 168

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2016). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, *27*(1), 89–118. 5, 110, 116, 124, 127, 168

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, *121*(1), 147–153. 26

Plaisier, M. A., Tiest, W. M. B., & Kappers, A. M. (2009). One, two, three, many–subitizing in active touch. *Acta psychologica*, *131*(2), 163–170. 13

Plaisier, M. A., Tiest, W. M. B., & Kappers, A. M. (2010). Range dependent processing of visual numerosity: similarities across vision and haptics. *Experimental brain research*, *204*(4), 525–537. 13

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, *28*(01), 1–14. doi: 10.1017/S0140525X05000014 30

Price, G. R., Mazzocco, M. M. M., & Ansari, D. (2013). Why mental arithmetic counts: brain activation during single digit arithmetic predicts high school math scores. *The Journal of Neuroscience*, *33*, 156–163. 55

Quinlan, P. T., van der Maas, H. L. J., Jansen, B. R. J., Booij, O., & Rendell, M. (2007). Re-thinking stages of cognitive development: An appraisal of connectionist models of the balance scale task. *Cognition*, *103*(3), 413–459. doi: 10.1016/j.cognition.2006.02.004 30, 36

Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study. *Child development*, *86*(1), 159–175. 90, 93

R Core Team. (2013). R: A language and environment for statistical computing.
40, 111, 134

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, *25*, 111–164. doi: 10.2307/271063 40

Raijmakers, M. E., Jansen, B. R. J., & van der Maas, H. L. J. (2004). Rules and development in triad classification task performance. *Developmental Review*, *24*(3), 289-321. doi: doi.org/10.1016/j.dr.2004.06.002 36

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*. 56

Reed, H. C., Stevenson, C., Broens-Paffen, M., Kirschner, P. A., & Jolles, J. (2015). Third graders' verbal reports of multiplication strategy use: How valid are they? *Learning and Individual Differences*, *37*, 107–117. 55

Reeve, R., Reynolds, F., Humberstone, J., & Butterworth, B. (2012). Stability and change in markers of core numerical competencies. *Journal of Experimental Psychology: General*, *141*(4), 649. 12, 26

Reyna, V. F., & Brainerd, C. J. (2007). The importance of mathematics in health and human judgment: Numeracy, risk communication, and medical decision making. *Learning and Individual Differences*, *17*(2), 147–159. 106

Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 317–345. 107, 128

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. 106

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/` 96

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. 58

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive science*, *6*(1), 1–36. 72

Salthouse, T. A. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological bulletin*, *99*(3), 303. 72

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, *22*(6), 800. 104

Savi, A. O., Ruijs, N., Maris, G., & van der Maas, H. L. J. (2017). Online learning solves sturdy problems in education experiments.
128

Savi, A. O., van der Maas, H. L. J., & Maris, G. (2015). Navigating massive open online courses. *Science*, *347*(6225), 958–958. 128

Savi, A. O., Williams, J. J., Maris, G., & van der Maas, H. L. J. (2017). The role of a/b tests in the study of large-scale online learning.
doi: 10.17605/OSF.IO/83JSG 128

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, *110*(3), 395–411. doi: 10.1016/j.cognition.2008.11.017 30, 31, 52

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, *8*(2), 23–74. 96

Schleifer, P., & Landerl, K. (2011). Subitizing and counting in typical and atypical development. *Developmental science*, *14*(2), 280–291. 12, 25

Schmidt, W. C., & Ling, C. X. (1996). A decision-tree model of balance scale development. *Machine Learning*, *24*(3), 203–230. 31

Schwartz, J. E., & Stone, A. A. (1998). Strategies for analyzing ecological momentary assessment data. *Health Psychology*, *17*(1), 6. 73

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464. doi: http://www.jstor.org/stable/2958889 40, 62

Shaffer, L. H. (1976). Intention and performance. *Psychological Review*, *83*(5), 375. 72

Shaffer, L. H., & Hardwick, J. (1968). Typing performance as a function of text. *The Quarterly Journal of Experimental Psychology*, *20*(4), 360–369. 73

Shanks, D. R. (2010). Learning: From association to cognition. *Annual review of psychology*, *61*, 273–301. doi: 10.1146/annurev.psych.093008.100519 30

Shapiro, S., & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591-611. doi: doi.org/10.2307/2284728 48

Shultz, T. R. (2003). *Computational developmental psychology*. MIT Press. 31

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning*, *16*(1-2), 57–86. 31

Shultz, T. R., & Takane, Y. (2007). Rule following and rule use in the balance-scale task. *Cognition*, *103*(3), 460–472. doi: 10.1016/j.cognition.2006.12.004 30, 36

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive psychology*, *8*(4), 481–520. 30, 31, 33, 34, 35, 50, 52

Siegler, R. S. (1978). *The origins of scientific reasoning.* Lawrence Erlbaum Associates, Inc. 35

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*(2), 1-74. 33, 47, 50

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*, 258. 55, 67, 68, 69

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford University Press. 30, 123

Siegler, R. S. (2006). Microgenetic analyses of learning. *Handbook of child psychology*. 115, 123

Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, *10*, 104–109. 68

Siegler, R. S., & Alibali, M. W. (2005). *Children's thinking (4th ed.)*. Upper Saddle River, NJ: Prentice Hall. 90

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, *46*(6), 606–620. 123

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., . . . Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological science*, *23*(7), 691–697. 88

Siegler, R. S., & Lortie-Forgues, H. (2014). An integrative theory of numerical development. *Child Development Perspectives*, *8*(3), 144–150. 88

Snijders, T., & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage. 75

Snitz, B. E., Small, B. J., Wang, T., Chang, C.-C. H., Hughes, T. F., & Ganguli, M. (2015). Do subjective memory complaints lead or follow objective cognitive change? a five-year population study of temporal influence. *Journal of the International Neuropsychological Society*, *21*(9), 732–742. 90

Spearman, C. (1927). *The abilities of man.* Macmillan. 89

Starkey, P., & Cooper, R. G. (1995). The development of subitizing in young children. *British Journal of Developmental Psychology*, *13*(4), 399–420. 13, 26

Stevenson, C. E., Hickendorff, M., Resing, W. C., Heiser, W. J., & de Boeck, P. A. (2013). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence*, *41*(3), 157–168. 101

Straatemeier, M. (2014). *Math Garden: A new educational and scientific instrument* (Doctoral dissertation, University of Amsterdam). Retrieved from `http://hdl.handle.net/11245/1.417091` 2, 3, 6, 55, 89, 94, 106, 131, 146, 165, 166, 168

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological science*, *12*(6), 462–466. 75

Svenson, O., & Sjöberg, K. (1983). Speeds of subitizing and counting processes in different age groups. *The Journal of genetic psychology*, *142*(2), 203–211. 13

Swift, E. J. (1904). The acquisition of skill in type-writing; a contribution to the psychology of learning. *Psychological Bulletin*, *1*(9), 295. 72

Tenison, C., Fincham, J. M., & Anderson, J. R. (2014). Detecting math problem solving strategies: An investigation into the use of retrospective self-reports, latency and fmri data. *Neuropsychologia*, *54*, 41–52. 55

SLO. (2009). SLO leerlijnen rekenen/wiskunde [learning program arithmetic/mathematics]. Retrieved from `www.tule.slo.nl` (Accessed: 2017-08-27) 69, 124, 127, 131

Thomson, G. (1951). *The factorial analysis of human ability*. London, England: University of London Press. 103

Towse, J. N., & Hitch, G. J. (1996). Performance demands in the selection of objects for counting. *Journal of experimental child psychology*, *61*(1), 67–79. 14

Trick, L. M. (2008). More than superstition: Differential effects of featural heterogeneity and change on subitizing and counting. *Attention, Perception, & Psychophysics*, *70*(5), 743–760. 13, 19, 25

Trick, L. M., Enns, J. T., & Brodeur, D. A. (1996). Life span changes in visual enumeration: The number discrimination task. *Developmental Psychology*, *32*(5), 925. 26

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 331. 12, 13

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological review*, *101*(1), 80. 26

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental psychology*, *45*(4), 1097. 88

van Maanen, L., Been, P., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In *Mathematical psychology in progress* (pp. 267–287). Springer. 51

van den Bergh, M., Hofman, A. D., Schmittmann, V. D., & van der Maas, H. L. J. (2015). Tracing the development of typewriting skills in an adaptive e-learning environment. *Perceptual and motor skills*, *121*(3), 727–745. 2

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, *85*(4), 475–511. 107

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. 127

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006a). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological review*, *113*(4), 842. 8, 131, 170

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006b). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological review*, *113*(4), 842–861. 89, 90, 92, 103

van der Maas, H. L. J., Kan, K.-J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. seriously. *Journal of Intelligence*, *2*(1), 12–15. 90

van der Maas, H. L. J., Kan, K. J., Hofman, A. D., & Raijmakers, M. E. J. (2014). Dynamics of development: a complex systems approach. In P. C. M. Molenaar, R. M. Lerner, & K. M. Newell (Eds.), *Handbook of developmental systems theory and methodology* (p. 270-286). New York: Guilford Press. 89, 131

van der Maas, H. L. J., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Preprints*. 89

van der Maas, H. L. J., & Nyamsuren, E. (2017). Cognitive analysis of educational games: The number game. *Topics in cognitive science*, *9*(2), 395–412. 6, 169

van der Maas, H. L. J., Quinlan, P. T., & Jansen, B. R. J. (2007). Towards better computational models of the balance scale task: A reply to shultz and takane. *Cognition*, *103*(3), 473–479. doi: :10.1016/j.cognition.2007.01.009 36

van der Maas, H. L. J., & Straatemeier, M. (2008). How to detect cognitive strategies: commentary on differentiation and integration: guiding principles for analyzing cognitive change. *Developmental science*, *11*(4), 449–453. doi: 10.1111/j.1467-7687.2008.00690.x 35

van der Ven, S. H. G., Boom, J., Kroesbergen, E. H., & Leseman, P. P. M. (2012). Microgenetic patterns of children's multiplication learning: Confirming the overlapping waves model by latent growth modeling. *Journal of experimental child psychology*, *113*(1), 1–19. 123

van der Ven, S. H. G., Klaiber, J. D., & van der Maas, H. L. J. (2017). Four and twenty blackbirds: how transcoding ability mediates the relationship between visuospatial working memory and math in a language with inversion. *Educational Psychology*, *37*(4), 487–505. 6, 169

van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012). The development of executive functions and early mathematics: A dynamic relationship. *British Journal of Educational Psychology*, *82*(1), 100–119. 68, 88, 90, 106 van der Ven, S. H. G., Straatemeier, M., Jansen, B. R. J., Klinkenberg, S., & van der Maas, H. L. J. (2015). Learning multiplication: An integrated analysis of the multiplication ability of primary school children and the difficulty of single digit and multidigit multiplication problems. *Learning and Individual Differences*, *43*, 48–62. 6, 54, 57, 64, 94, 114, 169

van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory & Cognition*, *10*(4), 396–404. 12

van Rijn, H., van Someren, M., & van der Maas, H. L. J. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257. doi: 10.1207/s15516709cog2702_4 31, 34, 50

van Veen, R., Evers-Vermeul, J., Sanders, T., & van den Bergh, H. (2013). The influence of input on connective acquisition: a growth curve analysis of english because and german weil. *Journal of child language*, *40*(5), 1003–1031. 78

Verachtert, P., De Fraine, B., Onghena, P., & Ghesquière, P. (2010). Season of birth and school success in the early years of primary education. *Oxford Review of Education*, *36*(3), 285–306. 103

Verguts, T., & Fias, W. (2005). Interacting neighbors: A connectionist model of retrieval in single-digit multiplication. *Memory & cognition*, *33*, 1–16. 54

Vetter, P., Butterworth, B., & Bahrami, B. (2008). Modulating attentional load affects numerosity estimation: evidence against a pre-attentive subitizing mechanism. *PLoS One*, *3*(9), e3269. 26

Visser, I., & Speekenbrink, M. (2010). depmixs4: An r-package for hidden markov models. *Journal of Statistical Software*, *36*(7), 1–21. 40

Wagenmakers, E.-J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196. Retrieved from `http://dx.doi.org/10.3758/BF03206482` doi: 10.3758/BF03206482 40, 96, 100

Wainer, H. (2000). *Computerized adaptive testing*. Wiley Online Library. 3, 167

Watson, D. G., Maylor, E. A., & Bruce, L. A. (2007). The role of eye movements in subitizing and counting. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1389. 12

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85. 4, 168

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer. 7

Wilkening, F., & Anderson, N. (1991). Representation and diagnosis of knowledge structures in developmental psychology. *Contributions to information integration theory*, *3*, 45–80. 30

Wilkening, F., & Anderson, N. H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin*, *92*(1), 215. 30, 34, 35, 37, 40, 48, 134

Wohlwill, J. F. (1973). *The study of behavioral development*. Academic Press. 88

Wu, C., & Liu, Y. (2008). Queuing network modeling of transcription typing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *15*(1), 6. 72

Yamaguchi, M., & Logan, G. D. (2014). Pushing typists back on the learning curve: Revealing chunking in skilled typewriting. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 592. 86

NEDERLANDSE SAMENVATTING

Grote longitudinale datasets zijn nodig om fundamentele vragen over cognitieve ontwikkeling en leren te kunnen beantwoorden (Adolph et al., 2008; P. C. Molenaar, 2004). Om de ontwikkelingspatronen vast te leggen, moet deze data worden verzameld terwijl kinderen zich ontwikkelen.

Rekentuin, een online adaptief oefen- en meetsysteem voor rekenen, is ontwikkeld om dit te doen: de ontwikkeling in wiskundige vaardigheden vast te leggen met als opbrengst intensieve tijdreeksdata voor grote aantallen kinderen (Klinkenberg et al., 2011; Straatemeier, 2014). Rekentuin bevat een breed scala aan spellen die studenten kunnen gebruiken om wiskundige vaardigheden (bijv. tellen en vermenigvuldigen) en cognitieve vaardigheden (bijv. proportioneel redeneren en werkgeheugen) te oefenen. Elk spel bestaat uit een set van vragen van verschillende moeilijkheid, zodat alle kinderen vragen kunnen maken die overeenkomen met hun vaardigheid.

Rekentuin begon in 2007 als een onderzoeksproject aan de Universiteit van Amsterdam en werd in 2009 gecommercialiseerd als reactie op de toenemende populariteit van het systeem. Dit was het begin van Oefenweb, een spin-off bedrijf dat Rekentuin en andere adaptieve oefensystemen voor het leren van Nederlands, Engels, statistiek en typen ontwikkelt en host. Op dit moment (augustus, 2017), tien jaar later, heeft Rekentuin 22 spellen, met in totaal 32.720 items en 831.280.316 responsen van 713.985 gebruikers verzameld, met een snelheid van ongeveer één miljoen reacties per dag[2]. Dit proefschrift onderzoekt deze unieke maar gecompliceerde dataset, met als doel cognitieve ontwikkeling te bestuderen.

Sleutelkenmerken van Rekentuin

Rekentuin is ontworpen voor kinderen om op school of thuis wiskunde te oefenen. Aangezien kinderen vrijwillig spelen, moet het systeem aantrekkelijk en motiverend zijn. Mede hierom is Rekentuin speels, adaptief en is er directe feedback na elk antwoord. Zo zijn verschillende taken opgezet als spellen met beloning voor oefenen. Figuur 1 toont een schermafbeelding van de landingspagina van Rekentuin. Elke plant in de tuin vertegenwoordigt een spel dat kan worden geselecteerd om een oefensessie te starten. Kinderen

[2] De meest frequente speler van het schooljaar 2016-2017 heeft 48.231 vragen gemaakt.

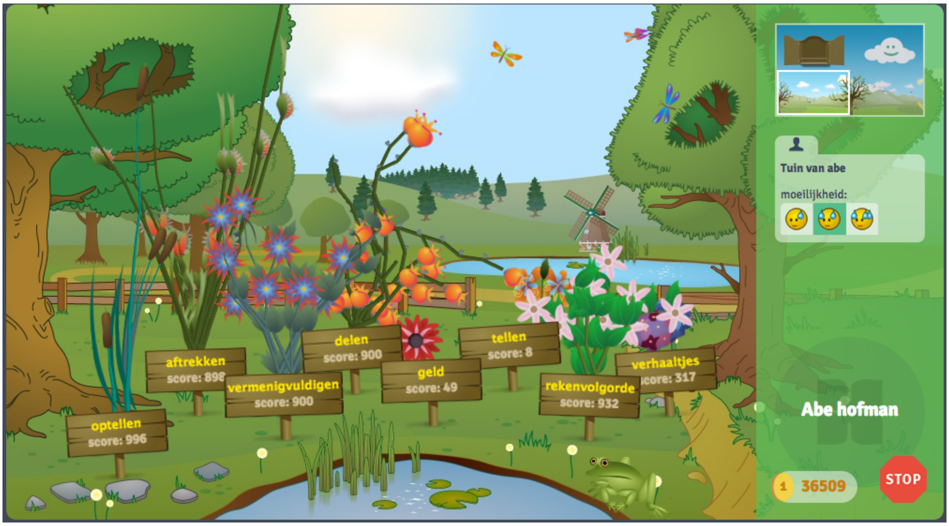


Figure 1: De landingspagina van Rekentuin. Elke plant staat voor een spel, en de score op het bord geeft een indicatie van de vaardighied van de speler (op een schaal van 1 tot 1000). Spelers kunnen de moeilijkheid van het spel bepalen door te klikken op het figuur met een, twee of drie zweetdruppels. Dit bepaaldt de kans op het correct maken van een opgave (90, 75, of 60 procent kans op een correct antwoord)

moeten de spellen regelmatig bezoeken om te voorkomen dat de planten verwelken. Ten tweede komen de oefeningen overeen met de vaardigheden van de kinderen, die erg divers zijn, zie bijvoorbeeld Straatemeier (2014, p. 13) en Dowker (2005). Ten derde wordt feedback gegeven na elk antwoord, om gepersonaliseerd leren te bevorderen en inspanningen te belonen.

Psychometrie van Rekentuin

De basis van Rekentuin is een uitbreiding op klassieke computer-adaptieve testmethoden (CAT). CAT is een testmethode die is gebaseerd op item respons theorie (IRT), die bestaat uit een grote familie van IRT modellen. De methode die door Rekentuin wordt gebruikt, is gebaseerd op het eenvoudigste IRT model, het 1-PL of Rasch-model:

$$P(x=1|\theta_p,\beta_i)=\frac{exp(\theta_p-\beta_i)}{1+exp(\theta_p-\beta_i)}$$

waar de kans op een goed antwoord wordt bepaald door het verschil tussen de persoonsvaardigheid (θ_p) en de moeilijkheid van de vraag (β_i). Het Rasch-model gaat uit van één-dimensionaliteit (één enkele latente vaardigheid wordt gemeten door alle vragen in het spel) en conditionele onafhankelijkheid (de responskansen zijn onafhankelijk van elkaar, gegeven de latente vaardigheid).

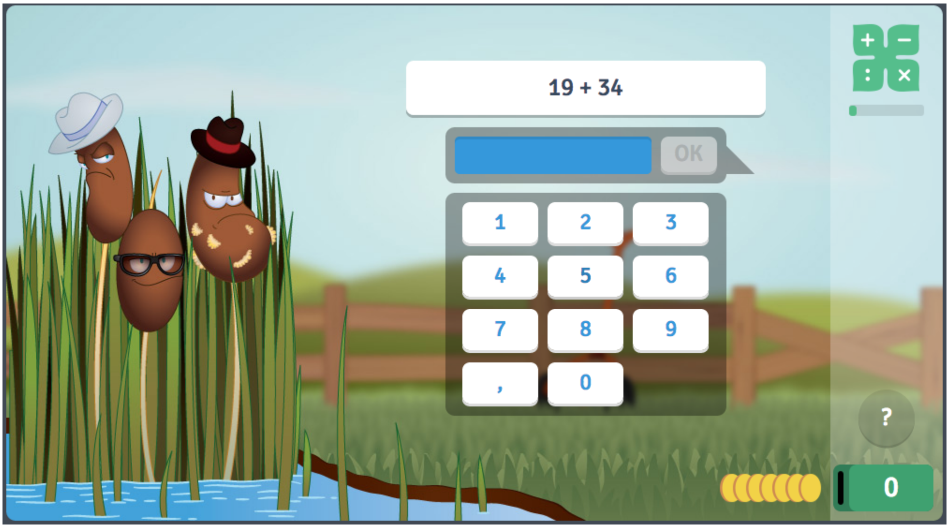


Figure 2: Een voorbeeldvraag van het optellen spel. Spelers kunnen hun eigen toetsenbord of het numeriek toetsenbord op het scherm gebruiken om antwoord te geven, of op het vraagteken drukken wanneer ze het antwoord niet weten. De muntjes aan de onderkant van het scherm visualiseren de geimplementeerde scoreregel. Elke seconde verdwijnt er een muntje. Na een goed antwoord wint de speler de resterende munten en na een fout antwoord verliest de speler de resterende munten.

In CAT hangt de volgorde van presentatie van de vragen af van hoe iemand reageerde op eerdere vragen (Wainer, 2000): als het onmiddellijk voorgaande antwoord correct was, wordt een moeilijker item weergegeven en vice versa. Het voordeel van het gebruik van CAT is dat vaardigheden, zoals rekenvaardigheid, kunnen worden geschat met minder vragen dan bij standaardtests. Momenteel wordt CAT voornamelijk gebruikt voor meten, maar in Rekentuin wordt het tegelijkertijd gebruikt voor meten en trainen van vaardigheid. Daarom heeft Klinkenberg et al. (2011) Rekentuin geïntroduceerd als een *computer adaptive practice*-systeem (CAP).

Rekentuin maakt gebruik van een uitgebreide CAT-techniek op basis van twee cruciale innovaties: (1) zowel snelheid als accuratesse bepalen de score en (2) het bijwerken van de persoonsvaardigheid en de moeilijkheid van de vraag gebeuren real-time met behulp van het Elo-algoritme.

De eerste innovatie, het gebruik van zowel accuratesse als reactietijd bij het bijwerken van de persoonsvaardigheid en de moeilijkheid van de vraag, vindt plaats door middel van een nieuwe scoreregel (Maris & van der Maas, 2012; Klinkenberg et al., 2011). Responstijden zijn opgenomen omdat ze belangrijke aanvullende informatie geven over de vaardigheid van een kind en een extra spel-element toevoegen. In Rekentuin hebben items meestal een tijdslimiet van twintig seconden. Volgens de geïmplementeerde *signed-residual time*-scoreregel (SRT) is de score gelijk aan de resterende tijd voor de juiste antwoorden (twintig seconden minus de responstijd), maar minus de resterende tijd voor foute antwoorden.

Raden is dus riskant, en als een kind niet weet wat het correcte antwoord is kan hij of zij het beste niet reageren of het vraagteken indrukken, wat een score van nul oplevert. De SRT-scoreregel wordt visueel gepresenteerd via munten onder aan het scherm die verwijzen naar de resterende tijd (zie Figuur 2). Door deze visualisatie, met een munt die elke seconde verdwijnt, kunnen zelfs jonge kinderen de scoreregel begrijpen. Verder heeft deze nieuwe scoreregel twee belangrijke voordelen. Ten eerste lost het het beruchte probleem van de *speed-accuracy trade-off* op (Wickelgren, 1977) omdat kinderen nu weten hoe snelheid en accuratesse worden gewogen in de score van hun antwoorden. Ten tweede heeft Maris and van der Maas (2012) aangetoond dat onder bepaalde milde statistische aannames deze scoreregel een standaard twee-parameter IRT-model impliceert. Daarom is er veel bekend over de eigenschappen van het meetmodel, bijvoorbeeld over de marginale en conditionele verdelingen van de parameters in het model.

De score kan formeel worden uitgedrukt als:

$$S = (2X_{pi} - 1)(d - T_{pi}),$$

waarbij X_{pi} 1 is voor een goed en 0 voor een fout antwoord op vraag i van speler p. T_{pi} is de responstijd en d de deadline. De verwachte score ($E(S)$) die volgt uit het meetmodel op basis van de SRT-scoreregel is:

$$E(S|\theta_p, \beta_i) = \frac{exp(2d(\theta_p - \beta_i)) + 1}{exp(2d(\theta_p - \beta_i)) - 1} - \frac{1}{(\theta_p - \beta_i)}.$$

De tweede innovatie, is het gebruik van een 'on-the-fly' Elo-algoritme (Klinkenberg et al., 2011), dat zijn oorsprong vindt in schaken (Elo, 1978). Het schatten van de parameters door middel van het Elo-algoritme resulteert in een zelf-organiserend systeem waarin zowel de schattingen van de vaardigheden van kinderen als de moeilijkheid van de vragen continu worden bijgewerkt, onmiddellijk na de reactie van een kind. De betrouwbaarheid van het Elo-algoritme is analytisch en in simulaties goed bestudeerd (Batchelder & Bershad, 1979; Glickman, 1999; Pelánek, 2016; Pelánek et al., 2016). Het meest opvallende voordeel van dit systeem is dat het niet de tijdrovende en dure procedure vereist om de vragen vooraf te testen zoals in normale CAT.

In het Elo-algoritme zijn de bijgewerkte schattingen gebaseerd op een gewogen som van eerdere schattingen en het verschil tussen de waargenomen en de verwachte score:

$$\begin{aligned} theta_{nieuw|p} &= theta_{oud|p} + K_p * (S - E(S)) \\ beta_{nieuw|i} &= beta_{oud|i} - K_i * (S - E(S)). \end{aligned}$$

De K-factor bepaalt het gewicht van de huidige respons bij het bijwerken van de parameters, en implementeert een *bias-variance trade-off*. De K-factor in Rekentuin neemt toe wanneer kinderen herhaaldelijk onder of boven de verwachte score scoren, of wanneer ze nieuw zijn in het systeem (zie Klinkenberg et al. (2011), Straatemeier (2014)).

Onderzoek met Rekentuin

De groeiende populariteit van Rekentuin biedt onderzoekers een rijke dataset. Het onderzoek in dit proefschrift kan grofweg worden gecategoriseerd door drie verschillende benaderingen. De eerste benadering is gebaseerd op directe analyses van de parameters die uit het systeem volgen. Eerder onderzoek met deze aanpak werd uitgevoerd door Klinkenberg et al. (2011), van der Ven et al. (2015, 2017) en Gierasimczuk et al. (2013). Klinkenberg et al. (2011) tonen aan dat de persoonsparameters van verschillende rekenkundige spellen sterk correleren met meer traditionele tests. Bovendien laat het werk van van der Ven et al. (2015, 2017) zien dat itemparameters overeenkomen met de effecten zoals voorspeld door verschillende theoretische modellen over wiskunde. Bovendien laten Gierasimczuk et al. (2013) en van der Maas and Nyamsuren (2017) zien dat de parameters van personen en items in zowel het spel *Deductive Mastermind* als een nummerreeks-spel, kunnen worden verklaard aan de hand van inhoudelijke (cognitieve) modellen die voor deze taken zijn ontwikkeld. Deze resultaten bieden ondersteuning voor de validiteit van sommige spellen in Rekentuin. In hoofdstuk 2 gebruiken we deze benadering en analyseren we de itemparameters van het tel-spel om verschillende strategieën te onderzoeken. We ondersteunen onze bevindingen verder met een experiment op twee basisscholen.

Een tweede onderzoeksbenadering is gericht op het begrijpen van de cognitieve strategieën die worden gebruikt door spelers in Rekentuin. Hiertoe wordt een cross-sectionele dataset samengesteld op basis van antwoorden (accuratesses en reactietijden) op vragen van een subset van kinderen die regelmatig een spel speelden. In deze benadering worden 'ruwe' data geanalyseerd met een uitgebreid latent variabel model dat gedetailleerdere processen kan vastleggen dan het huidige meetmodel van Rekentuin. In hoofdstuk 3 bestuderen we de regels die kinderen gebruiken om vragen uit de bekende balanstaak op te lossen. We bieden een vergelijking tussen een op expliciete regels gebaseerd model en een informatie-integratiemodel, over twee verschillende datasets: een meer traditionele pen en papier dataset en een dataset verzameld met Rekentuin. In hoofdstuk 4 onderzoeken we de strategieën bij vermenigvuldiging met behulp van uitgebreide IRT-modellen. Meer specifiek testen we of de vaardigheid om snelle accurate antwoorden te geven gelijk is aan de vaardigheid bij langzame responsen. Deze onderzoeken werpen licht op de strategieën die kinderen gebruiken bij het oplossen van vragen, en bieden daardoor ook aanwijzingen voor het geven van feedback aan kinderen en mogelijke verbeteringen van het Rekentuin-systeem.

Beide onderzoeksbenaderingen, en het grootste deel van het gepubliceerde werk tot nu toe, zijn gebaseerd op deze cross-sectionele gegevens. Het volgsysteem van Rekentuin stelt ons echter ook in staat ontwikkelingsprocessen te onderzoeken met behulp van een longitudinale subset van de data. De derde onderzoeksbenadering betreft dergelijke longitudinaal onderzoek. Ter illustratie van de longitudinale data toont Figuur 3 de gemiddelde vaardigheidsschattingen van een enkel domein gedurende vier opeenvolgende schooljaren, van zeven cohorten kinderen (elk cohort bestaat uit kinderen die in hetzelfde

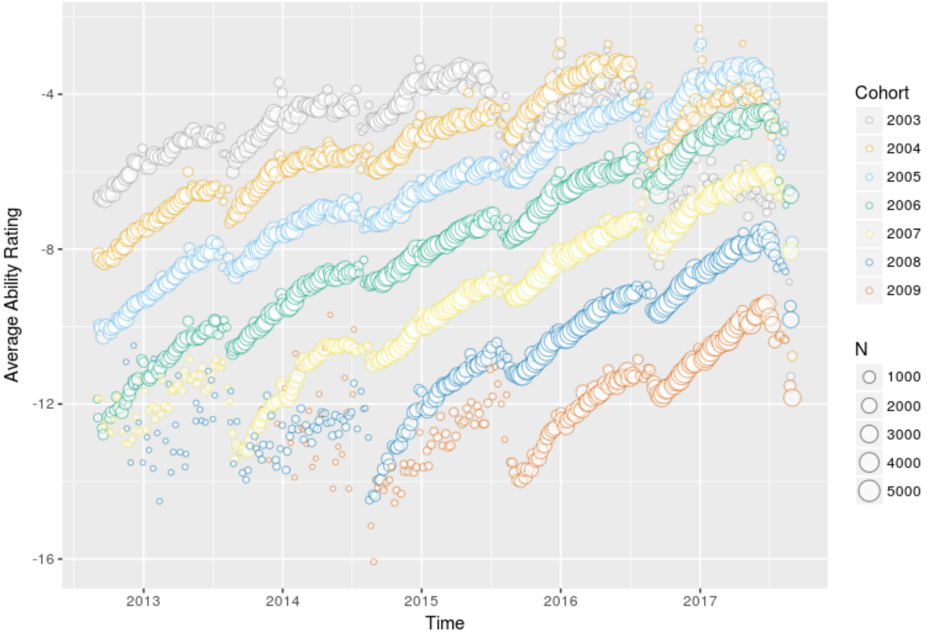


Figure 3: De ontwikkeling van de gemiddelde vaardigheid in het optellen spel van kinderen geboren in verschillende jaren. Elke lijn (reeks van stippen) geeft de ontwikkeling van kinderen weer uit een bepaald geboortejaar. De grootte van de stippen laat het aantal kinderen zien die in de desbetreffende week gespeeld hebben.

jaar zijn geboren). Deze figuur geeft data weer van 274.383 kinderen die het optelspel hebben gespeeld, en laat zien dat de gemiddelde optelvaardigheid van kinderen in de loop van de tijd toeneemt. Er worden echter enkele interessante extra patronen waargenomen: (1) tijdens de vakantie worden dalingen gevonden in de gemiddelde vaardigheid en (2) de grootste vooruitgang wordt waargenomen op jonge leeftijd, en de voortgang neemt langzaam af terwijl kinderen ouder worden.

In het tweede deel van dit proefschrift presenteren we verschillende analyses van longitudinale subsets van de data. In hoofdstuk 5 onderzoeken we de verbanden tussen de ontwikkeling van tellen en de ontwikkeling van de optellen, en tussen de ontwikkeling van vermenigvuldigen en de ontwikkeling van delen. Daarvoor analyseren we de vaardigheden van kinderen met behulp van tijdreeks-modellen, om zowel een mutualisme-benadering (van der Maas et al., 2006a) als een *g*-factor-benadering (Jensen, 1998) van ontwikkeling te vergelijken. In hoofdstuk 6 onderzoeken we de ontwikkelingsprocessen van leren met behulp van de reactietijden van toetsaanslagen van een groep kinderen die een typecursus volgden in de Typetuin (`www.typetuin.nl`).

In hoofdstuk 7 presenteren we verschillende *learning analytics* gericht op het beschrijven van tijdreeksreeksen van antwoorden van kinderen op specifieke vragen. Deze data komen van een subgroep van kinderen die bijna dagelijks, en gedurende langere perioden, hebben

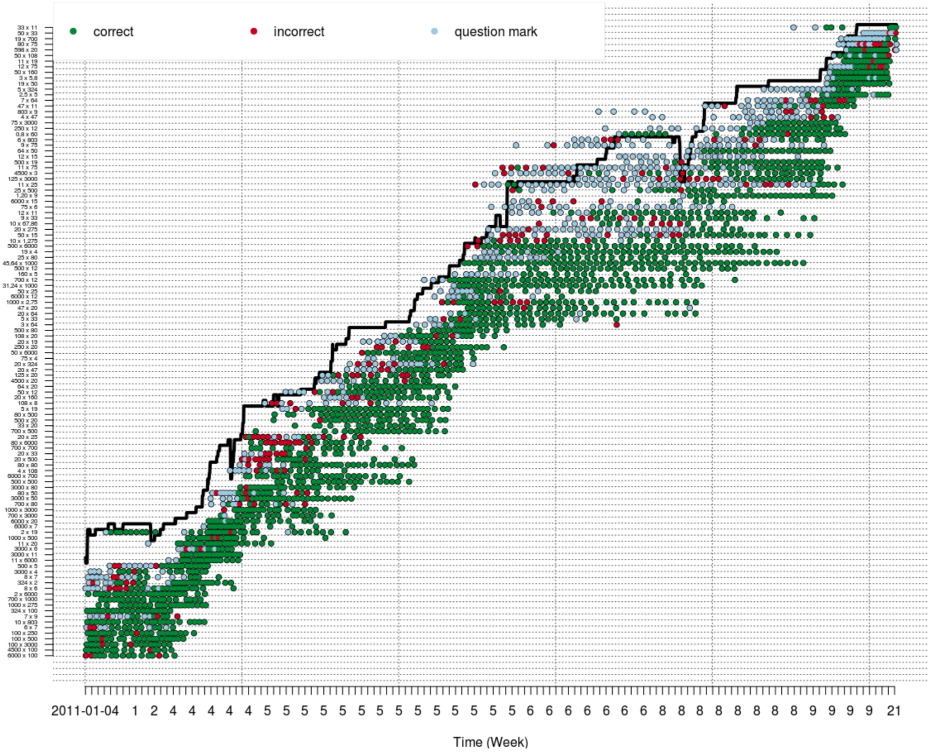


Figure 4: De ontwikkeling van antwoorden van een kind op een verschillen vermenigvuldigingsvragen. De vragen op de y-as zijn gesorteerd op de schatte moeilijkheid (laag is makkelijk, hoog is moeilijk). De lijn geeft de geschatte vaardigheid weer van dit kind.

gespeeld in de Rekentuin. De gegevens van deze kinderen zijn rijk aan kwantiteit en dynamiek, zoals Figuur 4 illustreert. De antwoorden van dit kind laten zien dat het kind voor sommige vragen niet in staat is om correcte antwoorden te geven, maar na een set vraagtekens of onjuiste antwoorden lijkt hij het juiste antwoord te leren en als gevolg daarvan neemt de schatting van zijn vaardigheid toe. In hoofdstuk 7 beschrijven we deze en andere ontwikkelingspatronen, en verzamelen we verschillende *learning analytics* om de stabiliteit van de antwoorden te onderzoeken.

Kortom, het huidige proefschrift bouwt voort op eerder onderzoek met Rekentuin, en breidt de analyse van item- en persoonsparameters uit naar andere spellen en naar koppelingen tussen spellen. Een belangrijke stap is dat dit proefschrift analyses van longitudinale data toont, wat een beter beeld geeft van de cognitieve ontwikkelingsprocessen. De voorbeelden in dit proefschrift gaan verder dan *snapshots* van wat zich ontwikkelt en tonen de dynamiek van ontwikkelingsprocessen. Dit proefschrift laat zien dat we aan de hand gegevens van Rekentuin, hoewel niet gemakkelijk te analyseren, een nieuwe blik kunnen werpen op cognitieve ontwikkeling.

DANKWOORD

De eerste stap van dit proefschrift is gezet tijdens mijn masterthese in de onderzoeksmaster psychologie. Ik zocht een project met een combinatie van psychometrie en ontwikkelingspsychologie, en kwam logischerwijs uit bij jou Han, en bij de Rekentuin. Ons gezamenlijke project over de balans-taak—waarbij we eerst even de adaptieve item-selectie uit moesten zetten, zodat we betrouwbare data konden verzamelen—vond ik een mooi project en zorgde ervoor dat ik op dit thema wilde promoveren. Mede door de manier waarop je bijna dagelijks even op mijn scherm keek, om vervolgens te concluderen dat het tijd was voor een potje tafeltennis. Deze ontspanning, open en altijd geïnteresseerde houding (ook voor alle zaken die niet met wetenschap te maken hebben) heb ik zeer gewaardeerd. Je talent voor het altijd weer een nieuwe invalshoek verzinnen bij een probleem, het soms streng zeggen dat dat paper nu echt de deur uit moet en je positieve houding hebben ervoor gezorgd dat dit proefschrift nu af is. Dat geldt zeker ook voor mijn beide copromotoren Brenda en Ingmar. Brenda, dank voor je gedetailleerde feedback en je belangrijke focus op de planning. Door je enthousiaste begeleiding van ons onderzoekspracticumgroepje in het tweede studiejaar wist ik dat ik methodenleer wilde doen. Ingmar, dank voor je waardevolle bijdrage bij meerdere hoofstukken in de dit proefschrift. Zowel aan de wiskundige modellen als aan het schrijfproces. Door het binnenhalen van een grote onderzoeksbeurs door Gunter werd de labmeeting een stuk groter. Bedankt Gunter en de andere leden van het *creative industries team* (Frederik, Matthieu, Marjan en Alexander). Ik heb genoten van het meedraaien. De *monday-morning-meetings*, de hei-dag en de fundamentele discussies waren leerzaam en inspirerend. En natuurlijk alle andere collega's bij methodenleer, jullie zijn een top team en ik ben dankbaar dat ik daar een onderdeel van ben.

Halverwege mijn promotieonderzoek ben ik deeltijd bij Oefenweb gaan werken. Zonder de bijdrage van alle mensen die hier aan de Rekentuin en de andere programma's sleutelen was er helemaal geen data geweest om te onderzoeken. Jullie maken samen echt een fantastisch programma (al ben ik natuurlijk wat bevooroordeeld). Ik geniet van de balans tussen wetenschap en het verbeteren van de verschillende leerprogramma's in samenwerking met de programmeurs en later ook de CoPs (content-psychometrie team). Dank voor de ontspannen cultuur met veel samenwerking en natuurlijk het *hacky-sacken*.

De kerels van *de (nieuwe) leven*, bedankt dat naast mijn onderzoek en later ook nog naast de twee kinderen toch tijd was om de kroeg in te duiken of te ontspannen in de Ardennen, op Texel en in Catalonië. Die houden we erin! Peter, fascinerend om te zien hoe parallel onze levens blijven lopen, laten we dat ook samen vasthouden!

En natuurlijk mijn twee paranimfen. Lucas, ik heb het idee dat we vaak op dezelfde golflengte zitten. Misschien niet altijd als we het hebben over statistische modellen, maar juist wel als over zaken praten die er echt toe doen. Dank dat je naast me wil staan op deze belangrijke dag. Dat geldt natuurlijk ook voor jou Alexander (ondertussen meer vriend dan collega). Mooi dat je alweer een plek had bemachtigd in de ITGWO groep, zonder dat ik daar enig idee van had. Je ontspannen houding, de potjes tafeltennis, onze gezamenlijke onderwijsklussen, het avontuur in Finland; op veel vlakken heb ik van je af kunnen kijken.

Tot slot, bedank ik graag onze *extended* familie. Pa, bedankt voor je steun en vertrouwen. Karin, je bent een geweldige oma. We zijn samen toch geen adviesbureau begonnen in onderwijskundig onderzoek, maar misschien komt dat nog. Mijn zusjes, Dieuwertje en Keye, jullie zijn toppers! Willien en Matthieu, Joke en Agaath; meemoeders (vader), peetouders of oppas, hoe je het ook wil noemen, jullie zijn een zeer belangrijke steun voor mij en ons gezin. De manier waarop jullie betrokken zijn is zeker niet vanzelfsprekend en daar zijn we dankbaar voor. Mijn moeder, Pieternel, je hebt alleen de eerste zes jaar van mijn leven mee kunnen maken, maar toch geloof ik dat de liefde die je daarin hebt uitgedeeld nog steeds een verschil maakt.

Finne en Ciske, wat zijn jullie een prachtige aanwinst in ons leven (en staan terecht op de voorkant). Jullie aanwezigheid relativeren alle uitdagingen op mijn werk. Wat zorgen jullie voor veel vreugde. En als allerlaatste, natuurlijk mijn vrouw Jessie. Zonder jou had ik dit nooit gekund. Je bent een fantastische vrouw. De manier waarop je in het leven staat inspireert me enorm. Het was niet altijd makkelijk, maar het is volbracht! Laten we dit samen gaan vieren.

LIST OF PUBLICATIONS

This section lists all manuscripts that are (close to being) submitted or accepted for publication at the time of submission of my dissertation.

Part of this dissertation

- Jansen, B. R. J., **Hofman, A. D.**, Straatemeier, M. Bers, B. M. C. W., Raijmakers, M. E. J., & van der Maas, H. L. J. (2014). The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology, 32*(2), 178–194
- **Hofman, A. D.**, Visser, I., Jansen B. R. J. & Van der Maas, H. L. J., (2014). The Balance-Scale Task Revisited: A Comparison of Statistical Models for Rule-Based and Information-Integration Theories of Proportional Reasoning. *PloS one 10* (10), e0136449
- **Hofman, A. D.**, Visser, I., Jansen, B. R. J., Marsman, M., & van der Maas, H. L. J. (Submitted). Fast and slow strategies in multiplication. Retrieved from *https://psyarxiv.com/aw3qq/*
- **Hofman, A. D.**, Kievit, R. A., Stevenson, C. E. , Molernaar, D., Visser, I., & van der Maas, H. L. J. (Submitted). The Dynamics of the Development of Mathematical Ability: A Comparison of *g*-Factor and Mutualistic Network Theories.
- van den Bergh, M. **Hofman, A. D.**, Schmittmann, V., van der Maas, H. L. J. (2015) Tracing the Development of Typewriting Skills in an Adaptive E-Learning Environment. *Perceptual & Motor Skills: Learning & Memory*, 121, 3, 1-19.
- **Hofman, A. D.**, Jansen, B. R. J., de Mooij, S. M. M., Stevenson, C. E. & van der Maas, H. L. J (2018) A Solution to the Measurement Problem in the Idiographic Approach Using Computer Adaptive Practicing. *Journal of Intelligence*, 6 , (1), 14

Other publications

- van der Maas H. L. J., Kan K.-J., **Hofman, A. D.** & Raijmakers M. E. J., (2012). Dynamics of Development: A Complex Systems Approach. *Handbook of Developmental Systems Theory & Methodology.*

- Brinkhuis, M. J. S., Savi, O. A., **Hofman, A. D.**, Coomans, F. van der Maas, H. L. J., & Maris, G (2018). Learning As It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System. Retrieved from *https://psyarxiv.com/g4z85/*

- Coomans, F., **Hofman, A. D.**, Brinkhuis, M. J. S., van der Maas, H. L. J., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PloS one, 11* (5), e0155149.

- Böing-Messing, F., van Assen, M. A., **Hofman, A. D.**, Hoijtink, H., & Mulder, J. (2017). Bayesian evaluation of constrained hypotheses on variances of multiple independent groups. *Psychological Methods, 22*(2), 262.

- Dekkers, L. M. S., Bexkens, A., **Hofman, A. D.**, De Boeck, P., Collot d'Escury, A. L., & Huizenga H. M. (2017). Formal Modeling of the Resistance to Peer Influence Questionnaire: A Comparison of Adolescent Boys and Girls With and Without Mild-to-Borderline Intellectual Disability. *Assessment*, 1-14

- Jansen B. R. J., **Hofman, A. D.**, Savi, O. A., Visser, I. & van der Maas, H. L. J. (2016) Self-adapting the success rate when practicing math. *Learning and Individual Differences*, 51, 1-10

- Burgoyne, A., **Hofman, A. D.**, van der Maas H. L. J. & Honing H.-J. (2015). Adaptive music recognition games for dementia therapy. *In Proceedings of the European Society for the Cognitive Sciences of Music*. Manchester, England.

- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., **Hofman, A. D.**, Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1-28.