Towards an idiographic education

Savi, O.A.

Publication date
2019
Document Version
Final published version
License
Other

TOWARDS AN IDIOGRAPHIC EDUCATION

ALEXANDER SAVI

Towards an Idiographic Education

Alexander Savi

alexandersavi.nl

The research in this dissertation was supported by a grant from the Netherlands Organisation for Scientific Research, project number 314-99-107.

Typeset:	LaTeX (using the dissertate class)
Artwork:	Xavier Delansa
Print:	Off Page
Digital:	dare.uva.nl & thesiscommons.org
DOI:	10.31237/osf.io/7gqhz

Towards an Idiographic Education

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 17 mei 2019, te 14.00 uur

door

Olof Alexander Savi
geboren te Baarn

Promotiecommissie

Promotores:	prof. dr. G. K. J. Maris	Universiteit van Amsterdam
	prof. dr. H. L. J. van der Maas	Universiteit van Amsterdam
Overige leden:	prof. dr. M. Meeter	Vrije Universiteit Amsterdam
	prof. dr. D. Borsboom	Universiteit van Amsterdam
	prof. dr. E. M. Wagenmakers	Universiteit van Amsterdam
	dr. C. P. B. J. van Klaveren	Vrije Universiteit Amsterdam
	dr. B. R. J. Jansen	Universiteit van Amsterdam
Faculteit:	Faculteit der Maatschappij- en Gedragswetenschappen	

Voor pa en ma.

Contents

1

Introduction

1.1 The educational sequence

Education can be seen as a long chain of interventions in a self-organizing developmental system. Sure, more poetic definitions exist, such as the popularized "lighting of a fire". Yet, for the purpose of this book, the initial definition provides us with a much more workable point of departure. An example from gardening will show you why.

If you have ever tried to grow your own vegetables, you know that it can be a delicate activity. You also know, that whereas some of the care is a one-off, such as finding a spot with the right temperature and desired amount of sunlight, most of it is structural, such as the provision of nutrition and water. Depending on your skill, the demands of the plant, and the fitness of its environment, it will prosper or wither.

Now, you may ask, why bother me with plant care when the actual topic is education? The reason is simple. Plant care can simply be viewed as a long chain of interventions—water, fertilize, water again, remove aphid, and so on. This is not any different in education. Sure, the interventions that constitute education are far more complex than the clear interventions in the plant analogy. But education can too be viewed as a long chain of interventions—motivate,

provide effortful practice, instruct, and so on—this time aimed at the cognitive (or social, emotional, or affective) growth of the student, rather than the physical growth of a plant.

On top of that, the goals of gardening and education are comparable. Typically, a gardener aims to create the optimal conditions for her or his plants to flourish with minimal structural effort, such that ideally the garden becomes increasingly self supporting. Of course, with many different types of plants, this is not a trivial task. Indeed, teachers too aim to provide the individual students with the optimal conditions for self supporting learning. Quite a daunting task, to say the least.

The plant analogy and the abstract, sequential, intervention conception of education (hereafter called the *educational sequence*) help us discover the main themes in this book. In the following I introduce those main themes, clarify the educational sequence, and finally use it as a thread to introduce the chapters that follow.

1.2 Idiography

The primary theme of this book starts where the plant analogy stops. Centuries of breeding and selection have created cultivars with all kinds of desirable characteristics. As a result, in a field full of say tulips, each individual plant benefits similarly from the same treatment. Now, rather than a field with tulips, picture a classroom full of students. These students may benefit very differently from the educational sequence. Factors like family situation, health and wealth, and extracurricular activities, may all contribute to an enormous variability in which educational intervention suits a student best and at what moment those interventions should be applied.

This heterogeneity calls for an idiographic approach to the science of education. Idiography is defined as the study of the individual, and idiographic science is often contrasted with nomothetic science, the formulation of universal laws. In scientific psychology, Molenaar (2004) explains that idiographic science "brings back [...] the dedicated study of the individual, prior to pooling across other individuals. Each person is initially conceived of as a possibly unique system of interacting dynamic processes, the unfolding of which gives rise to an individual life trajectory in a high-dimensional psychological space."

In education, the idiographic approach is for instance justified by the fact that individual tutoring gives superior learning outcomes over traditional classroom instruction. A straightforward consequence is the idea that the educational sequence must be adapted to the individual. A caricatural description of traditional education on the other hand, may consist of frontal instruction and linear methods, creating the exact opposite—sequences of interventions that

are highly similar for each and every individual. The back cover of this book illustrates such undesirable sequences, where the letters represent interventions, the colors represent different sorts of interventions, and the rows represent individual sequences of interventions.

One may argue that such a caricature hardly exists, but a quite novel educational approach, that of Massive Open Online Courses (MOOCs), comes considerably close. Although there is much to say for learning at scale, the difficulty of tailoring MOOCs to the individual student's needs is a serious issue, and one that we—Savi, van der Maas, and Maris (2015)—addressed in Science, in response to a timely and constructive discussion of MOOC research by Reich (2015):

> We agree wholeheartedly with J. Reich that research on the effectiveness of Massive Open Online Courses (MOOCs) must focus on learning rather than mere clicking ("Rebooting MOOC research," Education Forum, 2 January, p. 34). Our biggest challenge will be figuring out what is most appropriate for an individual student at a given moment.
>
> Ideally, a MOOC would work like the GPS navigation device in your car. You tell it where you want to go, it figures out where you are, and it guides you along the most optimal route. Keeping with the analogy, current MOOCs are like having all GPS navigation devices instruct every car driver to turn right at 9:15 on Monday morning.
>
> If we can't adapt teaching and practice to the individual learner, MOOCs will never be more than a digital version of classroom teaching. To personalize the learning experience, we first need a detailed description of what a student already can and cannot do. Such information can be determined by traditional tests or by more powerful methods such as the practice-based trackers that already exist in other domains of online education (Klinkenberg, Straatemeier, & van der Maas, 2011). The A/B testing discussed in the Education Forum provides us with ideal methodology to start putting roads on the educational map. Once we gather information about various conditions, we can map each student's optimal route.

On the dimension of educational sequences, the one extreme is populated with sequences that are identical for each and every student (illustrated on the back cover), whereas on the other extreme all sequences are perfectly tailored to the individual (illustrated on the front cover). All educational programs can be seen to lie somewhere in between those extremes, and in this book I explore methods that may help increase the tailoring of education.

1.3 Real-world laboratory

The second theme of this book is substantiated by a very much related dissimilarity between plants and humans. Whereas plants can be studied not only under extremely controlled conditions, such as in greenhouses, they can too be bred and even genetically manipulated. Now consider this approach with humans. Most of us would agree that we would soon face serious ethical constraints. It is one of the reasons that sophisticated plant models exist (e.g., Vos et al., 2009) that can explain phenomena like growth speed and branching, whereas psychology has tremendous difficulty with mapping the enormous variability in humans.

Putting it mildly, humans are a profoundly difficult subject to study, and thus so are students. Cognitive psychologists have been successful in adopting the experimental method, which has facilitated the discovery of many important effects in the domain of learning (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Karpicke & Roediger, 2008). Furthermore, educational psychologists adopted randomized controlled trials—considered a 'gold standard' in clinical research—to study the effect of large educational interventions. In education, such trials face various criticisms, among which a rather painful one: double blinding is nearly impossible.

One specific form of field experiments provides an elegant solution to many of the problems faced in educational research. In this book, we gratefully exploit the rise of large-scale online learning environments. These environments are exciting not only because randomization and double-blinding are generally easily achieved, but because they provide a window into natural and authentic learning contexts. Although evidently not the same as the actual classroom, online learning environments can be seen as a model of human learning, shaped by both the online experiences and the experiences in the actual classroom.

Moreover, large-scale online learning environments unlock data that are difficult or even impossible to attain in traditional education. In Chapter 2 we show that these environments may not only unlock student notebooks by gathering the type, sequence, and amount of problems tried, but may too gather types of errors, response times, problem difficulties, and student abilities. Impressively, it does so live, with little effort, and on a massive scale, and is consequently increasingly able to capture the microgenetics of in vivo learning. Thus, analogous to the greenhouses of biologists, online learning environments provide educational psychologists with the means for a systematic inquiry into human's natural learning.

1.4 Learning and ability

Finally, the third theme of this book concerns an apparent dichotomy in scientific psychology. Most famously, Cronbach (1957)—then president of the American Psychological Association—addressed "the separation of the disciplines": the observation that experimental and correlational psychology work in relative isolation. Here, experimental psychology refers to the efforts to explain variation within persons, whereas correlational psychology refers to the efforts to explain variation between persons. Notice the similarity between this issue, and the nomothetic and idiographic approaches to science.

Just a couple of years before Cronbach's presidential address, Ferguson (1954) discussed the very issue in relation to intelligence. In an article that reads like a novel, he proposes a single conceptual framework that must bridge human learning and human ability. In his words, "[t]hose concerned with the description and classification of man's abilities have usually adopted an individual difference approach. They have paid scant attention to problems of learning. The experimentalists, engrossed in the study of learning, have for various theoretical and practical reasons shown little interest in individual differences. They seem unaware that they too are students of man's abilities."

Although Klinkenberg et al. (2011) might not have been aware, their computer-adaptive practice environment for arithmetic—which is studied in Chapter 2, 4, and 6—comes close to unifying the two disciplines. On the one hand, their adaptive algorithm that matches students to problems deals with important principles in learning, such as the zone of proximal development and instructional scaffolding. On the other hand, the exact same algorithm provides ability estimates that track the development of each student.

In Chapter 7, we redeem the promise of unifying the two disciplines from a strong theoretical point of view. We propose a formal framework that explains fundamental phenomena in human ability, and that offers the crucial bridge to human learning. Our idiographic theory allows not only for the evaluation of theories of learning, but also for studying the effect of educational interventions, and thus gives the educational sequence its appropriate place in the study of human ability.

1.5 Overview

The discussed themes give great context for understanding the chapters to come. Additionally, in the following I briefly go over each of the chapters, and explain how they connect. To begin

with, in Chapter 2 we discuss Math Garden, the previously mentioned computer-adaptive practice environment. Math Garden adapts the educational sequence by matching students to problems on the basis of ability and difficulty estimates. Specifically, we discuss the various challenges that come even with such a rudimentary approach.

Then, in Chapter 3, we discuss the state of online learning and propose a methodology aimed at improving the return on investment of such learning environments. We argue that online learning environments should indeed be viewed as real-world laboratories, ultimately benefiting the student. In Chapter 4 we use the suggested approach—online randomized experiments—to optimize Math Garden. Additionally, we introduce a novel approach to guaranteeing the reliability of exploratory research.

Naturally, tailored educational sequences and evidence-based improvements based on experimental comparisons require versatile learning environments. Unfortunately, not all learning environments offer such versatility. Therefore, in Chapter 5 I introduce a software protocol that can connect learning environments like MOOCs to the survey software Qualtrics. This way, the means for adaptivity and experimental comparisons that is offered by Qualtrics—albeit rudimentary—can now be used by teachers and researchers alike.

Where experiments are crucial in determining which intervention works best in the educational sequence, measurement is crucial in determining when to use a particular intervention. Where the aforementioned Math Garden uses an individual's ability estimate to determine the timing of a problem, in Chapter 6 we explore a diagnostic approach. Cognitive diagnosis aims at understanding an individual's deficits in learning and understanding, and here we introduce an intuitive model to identify the misconceptions that cause a student's errors. We apply the model to single digit multiplication and discuss how it may serve as diagnostic approach to personalization.

Finally, in Chapter 7 we introduce a novel idiographic approach to understanding intelligence and its development. Importantly, this chapter provides a theoretical framework for understanding adaptations to the educational sequence, and its influence on cognitive development.

No masterpiece was ever created by a lazy artist.

Salvador Dalí

2

Active analytics

Summary

With the advent of computers in education, and the ample availability of online learning and practice environments, enormous amounts of data on learning become available. The purpose of this chapter is to present a decade of experience with analyzing and improving an online practice environment for math, which has thus far recorded over a billion responses. We present the methods we use to both steer and analyze this system in real-time, using scoring rules on accuracy and response times, a tailored rating system to provide both learners and items with current ability and difficulty ratings, and an adaptive engine that matches learners to items. Moreover, we explore the quality of fit by means of prediction accuracy and parallel item reliability. Limitations and pitfalls are discussed by diagnosing sources of misfit, like violations of unidimensionality and unforeseen dynamics. Finally, directions for development are discussed, including embedded learning analytics and a focus on online experimentation to evaluate both the system itself and the users' learning gains. Though many challenges remain open, we believe that large steps have been made in providing methods to efficiently manage and research educational big data from a massive online learning system.

2.1 Introduction

The societal expectations of educational data and learning analytics are high. As more and more educational institutes routinely use computerized tools for training and testing, enormous amounts of data on learning are collected. These data support the idea that, "[...] in the near future it will be possible to continuously assess and store the unfolding life history (trajectory in behavior space) of each individual" (Molenaar, 2004, p. 216), and thereupon allow for a detailed study and targeted improvement of education. It should, for instance, be possible for teachers to create completely individualized educational programs based on the progress and learning difficulties of each student.

The role of learning analytics in shaping online learning systems is still emerging. In this paper, we contribute to this conversation by providing a case study of Math Garden (Klinkenberg et al., 2011), an online practice system for arithmetic. Math Garden aims to live up to the aforementioned promise by providing individualized computer-adaptive practice to over 400 000 primary school children in the Netherlands—by means of real-time ability estimates—and by giving teachers the tools to track the children's progress. In this case study, we first share design considerations, such as embedding learning analytics in the educational model, and then critically inspect the level to which those analytics are reliable. To this end, we determine the fit of the computer-adaptive model and explore sources of misfit. We finally share important considerations for the future of learning analytics.

We believe such a case study is particularly valuable, as designing learning analytics for such large-scale systems is not an easy task. Learning environments can have a sizable impact on education in general and individual students in particular. Moreover, interventions or design considerations in such systems, for instance based on learning analytics, may too have a significant impact on students' learning experiences, and must arguably be addressed with the same scrutiny as is demanded in traditional education.

To illustrate the reach, Math Garden involves almost 853 300 000 responses from over 452 thousand K-12 children, distributed across 5300 schools and many more household subscriptions, playing in 26 arithmetic domains totaling more than 37 000 different items. The rate at which items are answered is currently about 900 000 per school day. Then, to illustrate learning in Math Garden, Figure 2.1 shows the development of the domain *addition* over time (the Methods section explains the ability estimation procedure in detail). For each birth cohort (based on birth year), the development of average monthly performance is plotted, and nicely shows the

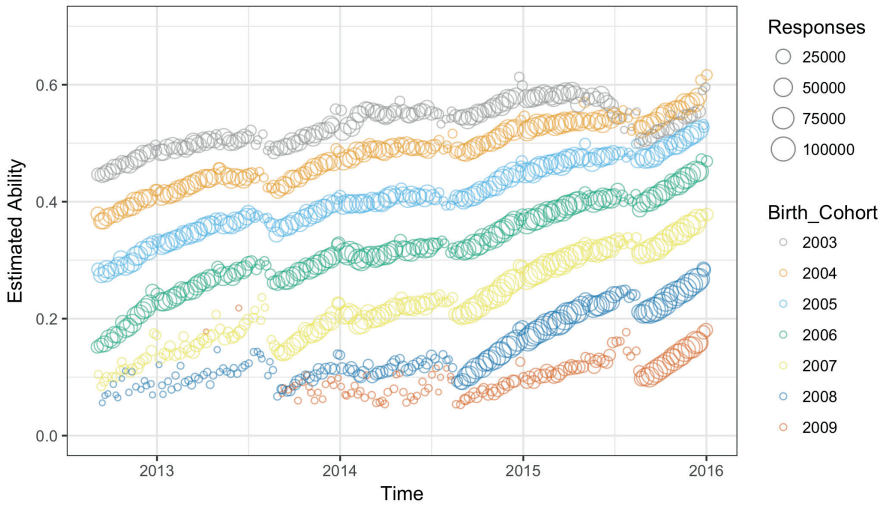


Figure 2.1. Growth in monthly average addition ability per grade, over a period of 6 years. Estimated ability represents the proportion of correct responses, if one responded to all addition items in the item bank. The number of responses in this graph totals 39, 391, 617, and the number of monthly responses can be seen to increase over time. For every new school year, the development of each grade is well visible. Also the continuation of progress over school years is clearly shown.

development throughout a school year. After the school year, the development continues in the next class in the next year. Figure 2.1 illustrates how learning progresses through the years, and how classes compare to one another and over time. The graph includes almost 40 million responses from over six years of data[1].

Traditional psychometric methods, like classical test theory and item response models (e.g., Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968; Rasch, 1960), fall short in systems like these due to the scale and adaptive nature of these systems. Therefore, Math Garden utilizes a different approach, and analyzes student responses on the fly, while continuously updating the estimates of student abilities and item difficulties. In this case study, we draw upon the lessons we learned during a decade of analyzing Math Garden data.

Design considerations and the implemented computer-adaptive model are discussed in the Methods section. We scrutinize the fit of this model in the Results section by inspecting whether the predictions of the model are in accordance with the actual observed responses. Additionally, we determine the reliability of the model's estimates of item difficulties by comparing the estimates of parallel items (e.g., $n \times m$ and $m \times n$). We then take a deep dive into the many

[1] All analyses are performed using R (R Core Team, 2016)

different possible causes of the small but significant amount of misfit observed in the system. We consider user-dependent responses processes, on both the global and local level: that is, the user-specific and item-specific strategies that children use for solving the items. The discussed sources of misfit primarily pertain to violations of the strict unidimensionality assumption that underlies the model.

This multi-method approach exposes the fact that many facets play a role in educational systems that embed learning analytics into their educational model. In the Discussion section, we suggest that this embedded approach, combined with other active forms of learning analytics such as online experiments, might prove worthwhile as a first step towards a more coherent field. The central challenge in this approach is ensuring that the defined educational model works as desired—for which the following sections provide a case study.

2.2 Methods

2.2.1 Math Garden

Dating from 2007, Math Garden is a computer adaptive practice system for arithmetic items, mainly focused on K-12 (Klinkenberg et al., 2011; Straatemeier, 2014). Originally, it was designed to freely capture long and dense time series data for the microgenetic study of cognitive development in general, and mathematical development in particular. Due to popular demand, it was commercialized in 2009[2] and different domains were developed, such as the adaptive practice of languages (English and Dutch), statistics (Groeneveld, 2014; Klinkenberg, 2014), and typing (van den Bergh, Schmittmann, Hofman, & van der Maas, 2015). Each system hosts eight to 26 games that each train a distinct ability relevant to the domain.

Children who log in to Math Garden land on a personalized page with a garden and various plants (see Figure 2.2). Each plant represents a mathematical domain, such as addition, multiplication, or fractions. By clicking the plant, children start practicing that domain (see Figure 2.3 for an example). Plants grow and flourish when the corresponding domain is frequently practiced, while plants wither when a domain is neglected. In each practice session, a set of 15 items is sequentially presented for 20 seconds each. Depending on the domain, children either pick the correct response from a set of alternatives or respond in an open format. Children may, within certain limits, hit a question mark button to skip items that seem too difficult.

After each game, children return to the landing page where they can again choose a domain

[2] www.oefenweb.com

Figure 2.2. Landing page with a garden and various plants that represent mathematical domains. The smileys can be used to select the difficulty level. The buttons in the top right can be used to navigate to other parts of the environment, such as a bonus garden with more domains or a prize cabinet. Camera symbols communicate the availability of instruction videos.

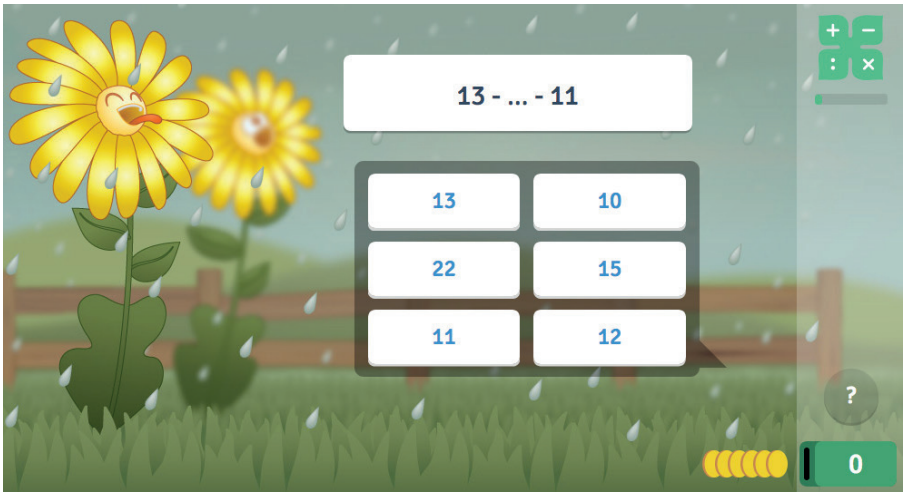


Figure 2.3. An item in the domain 'series'. Children must fill in the number that completes the incomplete series. The virtual coins indicate the remaining time. Children earn the remaining amount of coins if the answer is correct, and lose the remaining amount of coins if the answer is incorrect. The question mark can be used to skip the item.

to practice. Math Garden adaptively matches children to items with an appropriate level of difficulty: an individual child who fails to solve an item or does so too slowly will receive easier items, whereas a child who succeeds within the expected time will receive more difficult items. Additionally, children can set the difficulty level (i.e., expected probability correct) themselves individually, to either easy (about 90% correct), medium (about 75% correct), or hard (about 60% correct). Adaptive item selection in the context of a practice system such as Math Garden is therefore quite different from item selection in computer adaptive tests (CATs). CATs optimize measurement efficiency by selecting maximally informative items for measuring ability (i.e., items with a probability correct of about 50% might be selected) to obtain maximum measurement precision within a limited set of items (e.g., van der Linden & Glas, 2000; Wainer, 2000). Adaptive practice systems on the other hand, choose items to facilitate learning and motivation, as discussed by Veldkamp, Matteucci, and Eggen (2011) and shown by Jansen et al. (2013). In Math Garden, no *optimal* item selection is currently attempted, but items are sampled, taking into account the preferred difficulty level of the learner and recent history of answered items to avoid recent items.

In order to anticipate the multidimensional structure of math practice, Math Garden is designed such that each of the games consists of a separate ability, hence is assumed to be unidimensional. This can be regarded as quite a conservative approach given the involvement of 26 games and thus as many dimensions. Many psychometric models assume unidimensionality, as does the Elo rating system (ERS). Hence, for each single game a separate rating scale is implemented. However, due to the large amount of data, it is still possible to distinguish different dimensions within a single game. This is demonstrated in the Results section, where misfit is discussed. Yet, the amount of bias introduced by this multidimensionality, together with all other sources of misfit, is believed to be limited, as is discussed in the section Prediction accuracy. For each game within the Math Garden, two psychometric innovations are implemented: scoring rules and adaptive item selection, both discussed hereafter.

2.2.2 Scoring rule

Scoring rules play an important role in such diverse domains as sports, games, educational testing, and recruitment. In all these domains, they are introduced to elicit specific behavior that one somehow wants to quantify (e.g., answering correctly within a certain amount of time), thereby discouraging unwanted behavior that can reduce the validity and reliability (also called the accuracy or dependability, cf. Cronbach, 1951) of the measuring procedure (e.g., guessing).

See for example Lazer, Kennedy, King, and Vespignani (2014) for a general discussion on measurement in big data analysis, and Klinkenberg (2014) for an evaluation of the reliability and validity of the scoring rule. In large-scale computerized educational frameworks such as Math Garden, scoring rules additionally serve as a means to control the progression of the global system and to steer it towards a desired goal. To this end, it is important that the scoring rule is explicitly known and understood by the students, and that students act accordingly.

The scoring rule used in Math Garden, introduced by Maris and van der Maas (2012) and displayed in Figure 2.4, can easily be made explicit to the individual student. This scoring rule has the following form:

$$S_{pi} = (2x_{pi} - 1)(d - t_{pi}), \tag{2.1}$$

where S_{pi} denotes the score earned by user p after responding to item i, d denotes the time limit and $x_{pi} \in \{0, 1\}$ and $t_{pi} \in [0, d]$ denote, respectively, the accuracy and the response time of user p on item i. In Math Garden the time limit d is generally fixed at 20 seconds. The absolute value of the score S_{pi} is determined by the remaining time until the time limit, $d - t_{pi}$, whereas the sign of S_{pi} is determined by the accuracy x_{pi}. In this way, the scoring rule discourages fast guessing and imposes an explicit speed-accuracy trade-off (Wickelgren, 1977). The form of the scoring rule makes it easy to visualize the score to the individual user. At the start of an item the user sees a number of coins equal to the time limit in seconds, as visible in Figure 2.3. Each second one coin disappears. When a correct response is given the remaining number of coins is added to the total. In case of an incorrect response, it is subtracted. In Math Garden, children can collect these coins to buy virtual prizes. To allow users to omit an item that they do not know the correct response to, without having to wait until the time limit has passed, a question mark button has been built in. By using this question mark button, a user can go to the next item directly, and earns a score of zero on the skipped question, though its use is now limited to constrain strategic behavior in which students only try very easy items to maximize their points. Unless otherwise stated, all analyses presented in this paper are based on data from which these question mark responses are removed.

From the Signed Residual Time (SRT) scoring rule, Maris and van der Maas (2012) derived a response model. To estimate the response model's parameters on the incoming data streams from Math Garden, a rating system is implemented for each of the 18 games, facilitating real-time parameter updates, and driving the adaptive item selection discussed hereafter.

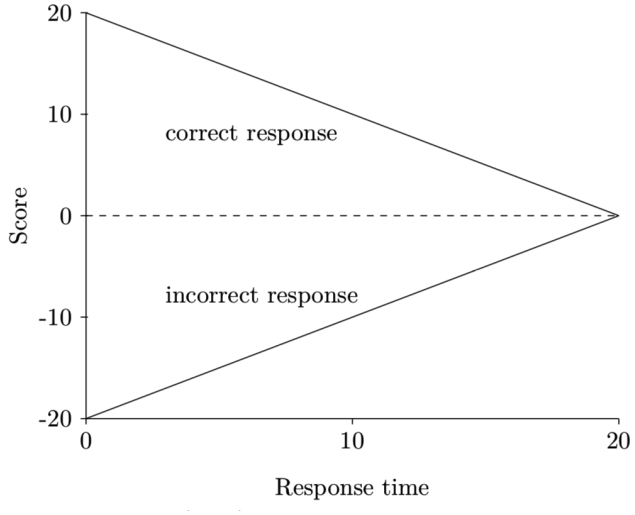


Figure 2.4. The Signed Residual Time scoring rule. If a user's response is correct, the score equals the remaining time until the time limit (shown by the top slope). If a user's response is incorrect, the score equals minus the remaining time until the time limit (shown by the bottom slope).

2.2.3 Adaptive item selection

To provide adaptive item selection, there is a need to determine what items are suitable to present to a specific student at a specific time. An algorithm based on the Elo Rating System (ERS) that both continually estimates the difficulty of the items and the ability of the students is used for this purpose.

The ERS has a history in the chess community, where dynamically changing abilities of chess players are expressed in Elo ratings (Batchelder & Bershad, 1979; Batchelder, Bershad, & Simpson, 1992; Elo, 1978). This provides a means to estimate dynamic ratings in setups that involve possibly massive paired comparisons. Hence, it is suitable for application in an educational context where item responses can be regarded as person–item paired comparisons, and we expect abilities and item difficulties to change over time (e.g. Brinkhuis, Bakker, & Maris, 2015; Klinkenberg et al., 2011; Pelánek, 2014; Wauters, Desmet, & den Noortgate, 2010).

To use the ERS in a computerized adaptive system like Math Garden, several modifications are required. First of all, the opposing player is replaced by an item i such that a user p responding to an item i is considered a match between the user and the item. This match is won by the user if the response is correct, and won by the item if the response is incorrect. The ratings

correspond to the user ability ϑ_p and the item difficulty δ_i, the score corresponds to the SRT score (Equation 2.1), which takes values in the interval $[-d, d]$, and the response model from which the expected score is computed is provided by the SRT model. After user p responds to item i and achieves SRT score S_{pi}, the user and item ratings are updated as follows (Klinkenberg et al., 2011):

$$\begin{aligned} \vartheta_p &\rightarrow \vartheta_p + K(S_{pi} - \mathcal{E}(S_{pi}))\,, \\ \delta_i &\rightarrow \delta_i - K(S_{pi} - \mathcal{E}(S_{pi}))\,, \end{aligned} \tag{2.2}$$

where K is a scaling factor and the expected score $\mathcal{E}(S_{pi})$ is based on the current ability estimate ϑ_p and item difficulty estimate δ_i:

$$\mathcal{E}(S_{pi}|\vartheta_p, \delta_i) = d\,\frac{\exp\left(2d\big(\vartheta_p - \delta_i\big)\right) + 1}{\exp\left(2d(\vartheta_p - \delta_i)\right) - 1} - \frac{1}{\vartheta_p - \delta_i} \tag{2.3}$$

where d is the time limit. Brinkhuis and Maris (2009, p. 11) provide an intuitive visualization of how such updates work.

The ERS has two specific advantages that are beneficial in the context of adaptive practice. First, the method in which ratings are updated makes them self-correcting. In Equation 2.2, the part $S_{pi} - \mathcal{E}(S_{pi})$ is simply the observed minus expected score. These differences facilitate the ERS to be self-correcting in its ratings—updates always steer in the right direction (e.g., a score that is higher than expected always gains points)—and the update size is related to the difference between observed and expected scores (e.g., for an unexpected correct response, the difference between observed and expected is quite high, and therefore the rating update is quite large, while for an expected correct response, the rating update is relatively small or can even be negative if the response given is too slow). This self-correcting feature makes the rating system quite robust: after every new response, ratings are updated in a sensible direction and hence adapt to changes in the underlying parameters. The K factor in Equation 2.2 functions is a scaling factor, and determines the size of the influence of the current response on the update of the ratings. A high K factor allows for ratings to quickly adapt to changes in the underlying parameters, yet introduces noise, whereas a lower K obtains smoother rating developments at the risk of adapting too slowly. This can be regarded as a classical bias-variance trade-off. Discussions on how to optimize K can be found in Elo (1978), Glickman (1999, 2001), Klinkenberg et al. (2011), and Sonas (2015).

A second beneficial feature of this rating system is that it is iterative and computationally

light. When the ERS was first introduced in the 1960's, updates could be calculated by hand, with the assistance of simple tables (Elo, 1978). The expected win probabilities depend on straightforward functions of estimated parameters, not on past data, and can be easily obtained. Clearly, with the advent of big data, this allows for real-time calculations on possibly large streams of data, with little computational load. In the implementation of Math Garden, parameters are therefore updated in real-time as responses become available. Note that real-time updates of parameters with IRT models is challenging—see for example Veldkamp et al. (2011) for an approach on updating ability parameters, or Brinkhuis (2014, pp. 83–114) for a (time-intensive) approach to updating parameters on a daily basis.

The ERS allows us to obtain up-to-date estimates of both person ability estimates ϑ_p and item difficulty estimates δ_i, which are continually adapted to possible changes. These ratings are used to facilitate many functions, such as adaptively selecting items at different difficulty levels, and providing teachers with child ratings and reference groups.

2.3 Results

Having discussed the design considerations and embedded learning analytics in Math Garden's computer-adaptive system, in this section we evaluate its model fit. Relevant for both the field of learning analytics in general, and computer-adaptive practice environments in particular, we scrutinize model fit by exploring various causes of misfit. To this end, we use a variety of methods on very diverse sets of data from the Math Garden ecosystem.

2.3.1 Evaluation of model fit

We start by a general evaluation of the computer-adaptive Elo model that underlies Math Garden, specifically by evaluating the quality of fit of the model. Model fit is evaluated in two specific traditions. First, we use prediction accuracy from the field of machine learning. Second, we use reliability measures from the field of psychometrics.

Prediction accuracy

Figure 2.5 gives an indication of the quality of fit of Math Garden's computer adaptive practice model. It shows the amount of practice, and the extent to which it is able to predict a child's responses. To be more precise, the figure shows for one particular child over the course of three years the difference between the observed and expected SRT scores, normalized over d, to every

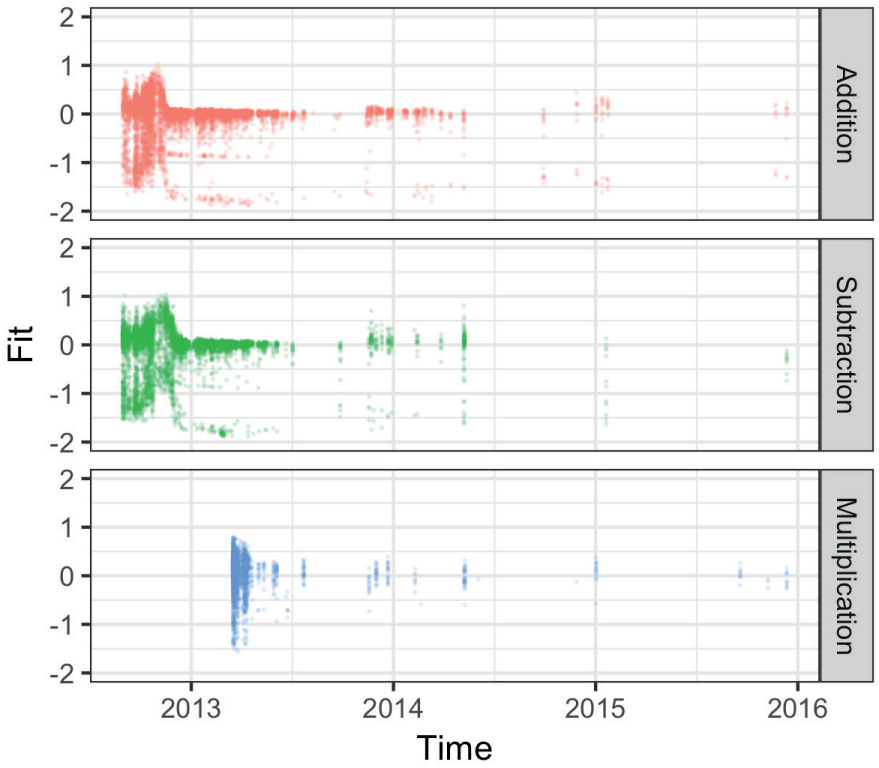


Figure 2.5. The development of model fit for one particular individual over time, for the domains addition, subtraction, and multiplication. For every single response over the course of three years ($n = 20,392$), the differences between the observed and expected SRT scores are shown, normalized over the time limit. The smaller this difference, the more accurate the expected score. For this individual, fit can be seen to improve over time on all three domains. The onset of practice differs between domains, and the amount of practice drops for these domains after May 2014. Some bimodality can be observed (partly due to guessing).

single item (s)he attempted in the fields of addition, subtraction, and multiplication. These differences can be interpreted as a proxy for model fit for this person. Hence, the closeness of these differences to the zero-lines correspond to good prediction, and therefore good model fit. The RMSE decreases from .45 ($n = 5,804$) to .29 ($n = 4,312$) for addition, and from .54 ($n = 4,369$) to .31 ($n = 3,305$) for subtraction, comparing 2012 and the rest. The total RMSE is .41 ($n = 20,392$).

In addition, one can see intensive practice starting in September 2012, and lasting through June 2013. Observations span a couple of years, since in Math Garden users are encouraged to revisit domains occasionally. Practice levels of this child and for these domains decline sharply after May 2014. First, narrowing down to the quality of fit, we see that after some initial phase, the difference between observed and expected responses tends to get centered closely around zero. At the onset of a new domain there is quite some noise, which reduces after some time. Since the estimated item difficulty parameters are readily available, the improvement in fit is not only due to better parameter estimation. This user increasingly conforms to the scoring rule and the response model that goes with it, which can also be observed by the relatively fast increase in fit in the multiplication domain. Hence, the estimated model parameters facilitate a good prediction. That Elo ratings can provide good prediction accuracy is not unique to Math Garden, and for instance also shown by Nižnan, Pelánek, and Řihák (2015).

Figure 2.6 provides another representation of the difference between observed and expected responses, this time for a large group of students. For one particular day, May 26, 2015, we have selected all responses for all games in Math Garden. As this day is situated at the end of the school year, we expect few new students and hence expect fit to have converged for this group, e.g., see Figure 2.5 for an improved fit near the end of the school year. On this day, 13,608 students provided 463,729 responses to 10,983 items on 17 different games. The differences between the observed and expected scores, normalized over time limit d, for all these responses are provided in Figure 2.6. The mean difference is close to zero (~ 0.02), which means that the ERS appears to do a good job at adjusting the expected scores toward the observed scores.

As the expected responses are best guesses at the time the actual responses are observed, the achieved accuracy indicates a significant amount of control on the dynamics in the environment. Nonetheless, Figure 2.5 also shows a constant stream of responses for which the observed score is not close to the expected one, and in Figure 2.6 one can clearly see that the histogram is bimodal.

We estimated a mixture of two normal distributions on this data,[3] and obtained a smaller distribution with 21% of the observations at $\mathcal{N}(-0.90, 0.31)$, and a larger proportion of 79%

[3]Using the mixtools R package (Benaglia, Chauveau, Hunter, & Young, 2009)

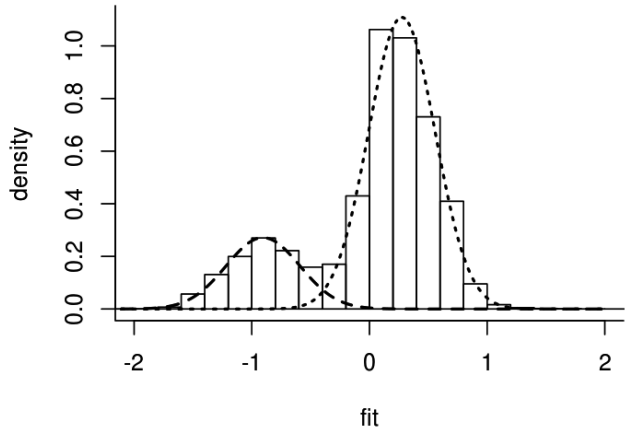


Figure 2.6. Histogram of the difference between the observed and expected SRT scores, normalized over the time limit, for all 463, 729 Math Garden responses on May 26, 2015, excluding skips. Overlaid is a fitted mixture of two normal distributions. The smaller distribution on the left contains 21% of the observations at $\mathcal{N}(-0.90, 0.31)$ (dashed), and the larger distribution on the right contains 79% of the observations at $\mathcal{N}(0.28, 0.28)$ (dotted). The mean difference between observed and expected scores is close to zero (0.02). The smaller distribution on the left corresponds to person-item interactions, where the expected result was correct, but the observed score was fast and incorrect—typical for typing errors and guessing.

at $\mathcal{N}(0.28, 0.28)$. The smaller distribution seems to collects all sorts of unexpected errors, such as typing errors, or fast guessing (Wang & Xu, 2015). When rather easy items are selected for the students, which have positive expected value, a quick error results in a quite large negative observed score (Figure 2.4), resulting in a (large) negative difference between observed and expected scores. These errors can also be observed in Figure 2.5, where in all three domains points can be seen hovering at the lower end of the panels. Since these errors are asymmetric, they introduce some bias in the estimated expected scores. Such bias might be removed, for example by disregarding quick incorrect responses in updates of the ratings. The larger component $\mathcal{N}(0.28, 0.28)$ includes responses that conform to the SRT score model (i.e., excluding guessing, typing errors, etc.), and allows us to estimate the RMSE of prediction to be 0.28, across all Math Garden games, excluding quick incorrect answers. Further considerations of fast and slow processes are discussed in section 2.3.2.

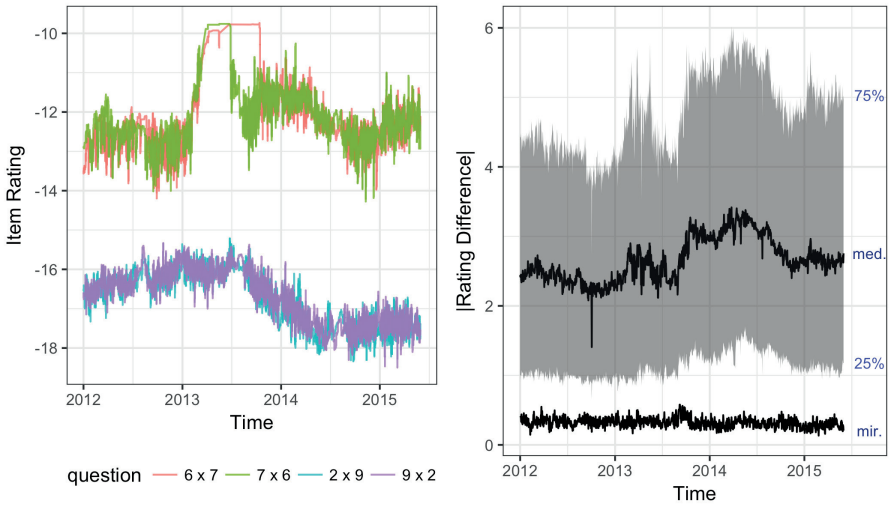


Figure 2.7. (left) Temporal development of the daily average rating of 2 pairs of parallel items from the multiplication domain. The lower pair of ratings constitutes the items 2×9 and 9×2 and the upper pair constitutes the items 6×7 and 7×6. Item ratings are non-transformed Elo ratings. (right) Temporal development (grey area) of the distribution of the absolute item pair-rating difference of all 4005 item pairs by the 90 non-symmetric items in the multiplication table. The absolute item pair-rating difference of the 45 mirror item pairs can be found at the bottom of the figure. Med. refers to the median of the distribution of non-mirror items (with 25% - 75% boundaries) and mir. refers to the median of the mirror items.

Reliability of parallel item development

In addition to prediction accuracy, the quality of fit of the adaptive system can be investigated by comparing the temporal development of the ratings of parallel items. Such temporal developments can be interpreted as a measure of reliability of the measurement, since similar items should have a similar (development of) difficulty. Parallel items look different superficially but share a number of features, which make them more or less equivalent. See Brinkhuis et al. (2015) for an approach to detect differential development of item pairs. An example of such parallel items is provided by mirror items like 2×9 and 9×2. In the left panel of Figure 2.7, the temporal development over a period of 3.5 years of the daily average ratings of two pairs of parallel items from the multiplication domain in Math Garden are displayed. It is clear that for both pairs the ratings of the parallel items remain very similar over time: their temporal rating evolutions overlap to a large extent, though the ERS allows for estimating their item difficulties independently.

The right panel of Figure 2.7 generalizes these findings to all 45 parallel item pairs that can be formed by the 90 non-symmetric items in the multiplication table. For every day in the 3.5-year period, the median, 25%-, and 75%-quantiles are determined of the absolute value of the item pair differences in daily average rating for all 45 item pairs. To put this in the right perspective, the figure also displays the temporal development of the median of the absolute item pair-rating differences computed over all 4005 item pairs that can be formed by the 90 non-symmetric items in the multiplication table. This figure makes it quite clear that the ratings of parallel items remain much closer over time than the ratings between two generic items. Even though other single-digit multiplication items can be equally difficult, the consistent small differences between mirror items is an indicator of the reliability of these ratings. See van der Ven, Straatemeier, Jansen, Klinkenberg, and van der Maas (2015) for considerations on the difficulties of single-digit multiplication, and when items cannot be considered mirror items.

Taking the above results together, both the prediction accuracy and the item ratings of parallel items suggest that the computer-adaptive architecture is able to create a considerable amount of stability within such a complex dynamical system. Given this achieved stability, the observed data collected with the system allows for a detailed look at the cognitive processes used in learning arithmetic. However, we also observed a certain amount of misfit, which we investigate next.

2.3.2 Diagnosis of model misfit

The methods in the previous section give different perspectives on the quality of fit, and give rise to further—more specific—explorations aimed at diagnosing misfit. Figures 2.5 and 2.6 display a number of promising results that indicate a good working of the mechanics underlying Math Garden. However, the dashed mixture component in Figure 2.6 clearly indicates a set of responses for which there are substantial differences between the observed and expected scores. These types of responses indicate alternative behavior (e.g., typing errors or fast guessing), which may lead to an incorrect assessment of the ratings, and to misfit of the response model.

A good working of the system requires the detection of the sources of this misfit after which appropriate steps can be taken to properly deal with these. First, we diagnose misfit by investigating different response processes (section Global response processes). Second, we analyze local item strategies (section Local response strategies). Finally, we provide a short overview of other sources of misfit, illustrating the complexity of identifying and correcting sources of misfit (section Other sources).

Global response processes

An attempt to assess the misfit of the SRT model and to situate the model in a more generalized framework of speed-accuracy response models can be found in Coomans, Hofman, Brinkhuis, van der Maas, and Maris (2016). In that paper the quality of fit of several of these models, including the SRT model, is investigated in the simplest possible non-trivial setup: persons try to solve two problems only; it is registered whether or not their response is correct and whether their response time is faster than half the time limit or slower than half the time limit. Hence, in this setup there are four different ways a person can answer a single item (fast and correct, slow and correct, slow and incorrect, and fast and incorrect) and 16 different ways a person can answer a pair of items. This simplistic setup is advantageous because:

1. It gives access to data from a large number of item pairs, spanning such diverse subject areas as basic arithmetic, language learning, and intelligence-related problems, with large numbers of independent observations per item pair.
2. Different speed–accuracy response models predict qualitatively different probability distributions of the 16 possible response patterns in a population of test takers. By inferring these distributions empirically by using, for example, Math Garden data, we can easily get a handle on the allowed speed–accuracy trade-off mechanisms.

To give an example of the analysis done in Coomans et al. (2016), reconsider the item pair 9×2 and 2×9, previously discussed in Figure 2.7. We obtained the response patterns of 13, 152 persons who responded to this pair of items within one day, and collapsed the response times in two categories: response times smaller than half the time limit are classified as slow, others as fast. The resulting data is summarized in the contingency table displayed as Table 2.1. In Coomans et al. (2016) it is demonstrated that the SRT model constrains the expected frequencies of the response patterns on anti-diagonals $(9, 6, 3)$, $(13, 10, 7, 4)$, and $(14, 11, 8)$ to be monotonically increasing or decreasing along these anti-diagonals. However, Table 2.1 is clearly incompatible with these predictions: the frequencies of the events on anti-diagonals $(9, 6, 3)$, $(13, 10, 7, 4)$, and $(14, 11, 8)$ are not monotonically increasing or decreasing along these anti-diagonals, but instead exhibit a dip (along $(9, 6, 3)$), a dip (along $(13, 10, 7, 4)$), and a peak (along $(14, 11, 8)$). The same features are found for numerous other item pairs in different domains, for all of which there are a great many observations that can be easily extracted from the Math Garden database.

Coomans et al. (2016) concludes that these features cannot be accounted for by simple 'one-process' models, such as the SRT model, and that a more complex model is needed. Therefore, they consider a 'two-process' model developed in Partchev and Boeck (2012) and which explicitly

		2 × 9			
		incorr./fast	incorr./slow	corr./slow	corr./fast
9 × 2	incorr./fast	313^{1}	107^{2}	137^{3}	434^{4}
	incorr./slow	124^{5}	98^{6}	153^{7}	230^{8}
	corr./slow	132^{9}	130^{10}	684^{11}	1221^{12}
	corr./fast	440^{13}	190^{14}	1211^{15}	7550^{16}

Table 2.1. Item pair contingency table for the items 9 × 2 and 2 × 9, constructed from 13,152 persons who responded to the item pair within a single day (over the period 2011-03-01 to 2015-06-29). The cells are numbered using superscript. The cells 1, 4, 13, and 16 constitute the events for which both responses on the item pair are fast. The cells 6, 7, 10, and 11 constitute the events for which both responses on the item pair are slow. All remaining cells constitute the events for which the speed of both responses on the item pair differs.

distinguishes between fast and slow responses, showing that this model results in a better fit than the more parsimonious SRT model, which does not make such a distinction. A similar conclusion was reached in Hofman, Visser, Jansen, Marsman, and van der Maas (2017).

Local response strategies

We will now turn to an example where the adaptive nature of Math Garden steers towards undesired behavior, ultimately resulting in misfit of the model. This example was encountered in the balance-scale task (Inhelder & Piaget, 1958), implemented in Math Garden in 2010. In this task, children predict the movement of a balance-scale (see Figure 2.8), with a varying number of blocks on each peg and varying distance between the blocks and the fulcrum. The task is famous for the interesting (erroneous) strategies used by children (and adults). To discriminate between these strategies, Siegler (1976) classified items to different item types. Simple items are included to discriminate between children who use a simple strategy based on only one dimension; counting only the number of weights or only looking at the distance between the location of the blocks and the center of the scale. Next to the simple items, complex items are added where children need to integrate both the weight and the distance information to correctly solve the item.

In the first implementation of the task, the adaptive item selection was based on the differences between estimated item difficulties and user abilities, as in all other Math Garden domains. Interestingly, when items are selected based on the Elo parameters, the collected estimated ability ratings for individual users show jumps between qualitatively different strategies. However, when items are selected in a fixed order, no such development is present (see left-panel

Figure 2.8. Example of the balance-scale task, as implemented in Math Garden. The coins reflect the used scoring-rule: if a correct/incorrect response is provided, the coins are added/subtracted to/from the child's virtual savings. Children respond by clicking the left, middle, or right picture of the balance scale—depicting the three possible states of the balance scale—or by clicking the question mark button.

of Figure 2.9 for both patterns). Also, a closer look at the responses to items (see right-panel of Figure 2.9) reveals that item responses are clustered when items are selected based on Elo ratings (groups of correct and incorrect responses are visible).

This clustering has an intriguing cause. The users seem to develop a local strategy that only works on the cluster of items presented at that specific moment. For example, a user might recognize that the response 'balance' is not correct for the first few items and learns that the balance response is always incorrect. Between item position 25 and 50 some of these items are presented but made incorrect, hence the ability estimate does not increase in the left panel of Figure 2.9. Since the system adapts the item difficulty estimates based on these responses, the difficulty estimates of these items increase and of the remaining items decrease. This results in an automatic clustering of items for which this local strategy fails versus items on which the strategy succeeds. After some incorrect responses, and receiving feedback, this child learns that he/she should provide only balance responses, as can be seen around item 80. This results in an increase in the ability estimate, what eventually results in the selection of items of yet another type (around 130), and a new local strategy seems to be learned.

In this example, the dynamic estimation of the item and user ratings, in combination with local strategies, result in dynamics that reinforce the reward of developing erroneous local strategies. Importantly, in this situation an ability estimate is based on the local cluster of items that

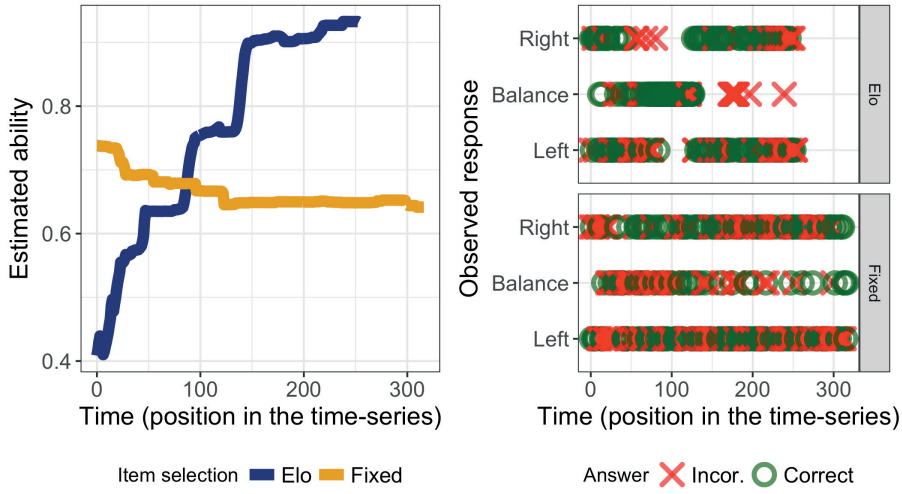


Figure 2.9. The development of responses to the balance-scale task, for a single child. The left-panel shows the rating development of a single user. When items are selected with the adaptive item selection based on the Elo estimates, large jumps in rating are observed, but when items are selected with a fixed sequence, a decrease in rating is observed for all users. The right-panel shows the responses (left, balance, or right), and whether the response was correct or incorrect, for both the data collected with the adaptive item selection procedure (clustering is visible) and fixed item selection procedure (no clustering visible).

the user has practiced, and does not generalize to the other clusters. This violates the assumption of unidimensionality and results in misfit.

To solve this undesired state of the system we intervened on the item selection by presenting a fixed order of items to all children, thus making the system less adaptive. The collected data and estimated ratings in the new implementation of the task showed large deviations compared to the first implementation. For example, the observed responses and ability ratings do not show a clear developmental pattern when items of different types are mixed (see the lower-panel of Figure 2.9). Although changing the item selection resolved the development of local strategies, still large discrepancies are found in the strategies used by children in Math Garden compared to strategies invoked by more traditional paper-and-pencil tests (Hofman, Visser, Jansen, & van der Maas, 2015). Clearly, such interactions between the content of the domain and the adaptive algorithm are not easily foreseen and require careful investigation into sources of misfit.

Other sources

The previous example illustrates that complex learning systems can have undesirable side effects, and that one should be on guard for unexpected behavior in different forms. Having diagnosed sources of misfit in alternative response processes and specific item strategies, in this section we briefly identify four more sources of misfit that are currently active areas of research in Math Garden. This collection of sources further demonstrates the rich variation in sources of misfit, the diverse set of methods required for their diagnosis, and hence the complexity in reducing misfit.

One source of misfit can be found in the (lack of) adherence to the scoring rule. A good working of the system implies that its users respond in accordance with the scoring rule, i.e., that their ability is reflected in the score that they earn. This is ensured in part by the form of the SRT scoring rule which strongly discourages guessing and thus prevents low ability users to earn scores that are too high and do not correspond to their actual ability. However, despite this explicit penalizing of fast incorrect responses, a substantial amount of guessing remains. Moreover, the particular form of the scoring rule can have a negative effect on less confident, yet high ability users. They might be scared by the high stakes associated to fast responding and produce a slow response resulting in a score that is too low for their actual ability. For these reasons it is important to develop methodology that enables an evaluation of scoring rules to find out if the (majority of) users conform to the SRT scoring rule, as discussed by Klinkenberg (2014).

A second source of misfit is very much related to the issue of users not adhering to the scoring rule. As mentioned in the Methods section, children earn virtual coins by giving correct responses, and faster responses yield more coins. Some children who aim to maximize the number of collected coins are observed to quickly skip problems that they deem too difficult to answer within a short time. They quickly use the question mark button to proceed to the next problem, as they're not penalized for doing so, and wait for an item that they can quickly answer correctly. This way, they somewhat circumvent the adaptive item selection by only choosing items that yield the most coins. Ultimately, this strategic behavior results in subtle misfit, as these children's abilities cannot be assessed correctly. For assessing such misfit, standard errors of estimates in the ERS would be beneficial, as explored by Brinkhuis and Maris (2010).

Interestingly, a solution to this issue was implemented in the Math Garden ecosystem. Savi, Ruijs, Maris, and van der Maas (2018) explain how a large online randomized experiment revealed that a simple delay in making the question mark button available, decreased the number of question marks used, and increased the amount of effort put into the children's responses. The development team of Math Garden has subsequently implemented such a delay throughout their ecosystem. The degree to which this intervention helped decrease misfit in the adaptive system is a subject of study.

For a third source of misfit, we investigate single-person-by-item time series. The size of the Math Garden data allows investigation of development in a new level of detail; that is, the individual development of accuracy, including the error types and response times, on a single item (Klinkenberg et al., 2011). These time series show interesting patterns from a developmental perspective and allow testing, for example, of whether learning of one set of items is related to learning another set of items. To illustrate the different patterns, we selected three of these series. Figure 2.10 shows the development of three different users on three different items ($5 + 1$, $3 + 4$, and 4×3) over a long period, with a maximum of 136 weeks.

The upper panel shows the responses of a child who learns to add five and one (and the parallel item), in three different stages. In the first stage, until position 25, he or she provides incorrect responses, mostly answering five. Thereafter, in the second stage, the correct solution is learned. In the third stage, from position 38 onward, the observed response time decreases indicating a more efficient strategy or faster sampling from memory (Ashcraft, 1982) compared to the previous stages.

On the other hand, the middle panel shows a time series of a child who does not learn $3 + 4$, while practicing this item for 61 times. Some correct responses are stated, but these are alternated with errors. These errors provide insight into the highly variable cognitive process of this child

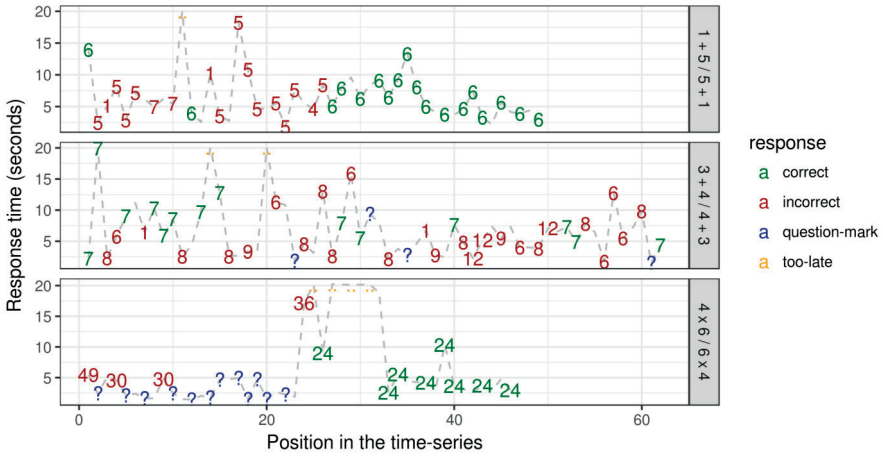


Figure 2.10. The development of both accuracy and response time, for responses to three different items (and their parallel items). For each item, the responses by one particular individual over time are shown. The ordered position in the time series are on the horizontal axis, which can span a maximum 62 responses. The upper and lower panel show a three-stage pattern, moving from mainly incorrect responses, to slow correct responses, to fast correct responses.

over time. The child alternates responses that can be labeled as close misses (6, 8 and 9) and responses labeled by wrong multiplication operand strategy (12). This highlights possibilities for tailoring instruction and feedback to misconceptions of a child as detailed as to a single item.

The development depicted in the lower panel shows a different pattern. This child starts with fast question-mark responses. Around position 22, he provides (too) slow responses, and seems to learn during this period the correct response. In the last part of the series he gives correct responses to this item. The dynamics of this child also highlight the differentiation between fast and slow processes, discussed previously.

The quantification of these developmental patterns, and the connection between multiple single-item time series, is ongoing research. Especially the connection between different time series provides insight into an important type of misfit. That is, the possible presence of item clusters within a certain domain, see for example Pelánek, Papoušek, Řihák, Stanislav, and Nižnan (2016) or the previous discussion of the balance-scale task. These possible item clusters show that learning a subset of items is strongly related to some items, whereas it is unrelated to other items within the same domain. These clusters provide insights in qualitative differences between the solutions strategies and learning patterns of children. Furthermore, the inclusion of these possible clusters in the measurement model can reduce the amount of misfit.

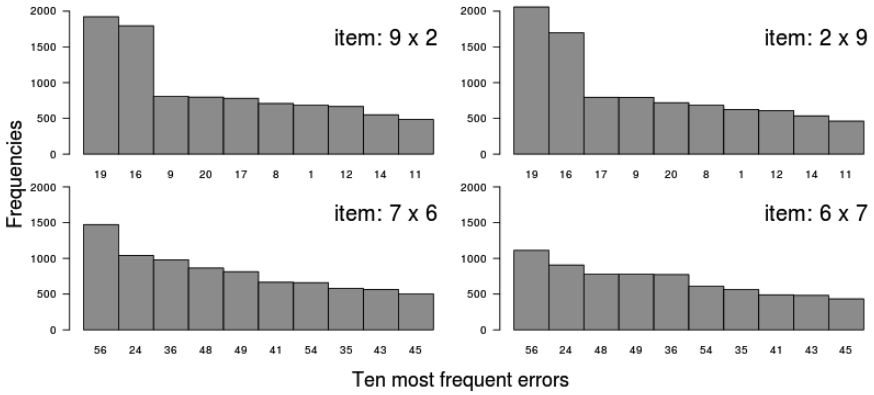


Figure 2.11. Frequencies of the ten most frequent errors on the items 2×9 and 6×7, and their parallel items.

The fourth and final source of misfit is captured in the just mentioned error responses. Error analyses provide an interesting direction for investigating misfit, because information about individual cognitive processes is contained in the types of errors that students make. More specifically, since different students may have different misconceptions, and different items are susceptible to different misconceptions, error analyses can help detect violations of unidimensionality.

An example of error analyses is shown in Figure 2.11. It shows the frequencies of the most made errors of the same set of items as depicted in Figure 2.7. Different aspects are highlighted by this plot. First, the observed correspondence in the error frequencies between the two sets of parallel items supports the reliability of the system. Second, the most frequent error for each parallel item pair seems to indicate different processes. The response 19 to the item 2×9 implies a mistake in counting, since children missed the correct response by one. Whereas the response 56 to the item 6×7 implies an operand relevant mistake, since the answer is correct for another multiplication problem (Straatemeier, 2014, pp. 99–128).

However, the classification of observed errors to error types is often ambiguous (J. S. Brown & VanLehn, 1980), and is therefore an active area of research in Math Garden. Take for example the incorrect response 18 to the item 9×9. This error can indicate that this child (1) adds instead of multiplies (wrong operand), (2) incorrectly reverses 81 to 18, or (3) states the response to an item from within the same table (operand related error). To solve this issue in error-classification, Straatemeier (2014, pp. 99–128) introduced and compared multiple classification methods (two literature-based approaches and four data-based approaches) that can be used

with the availability of big data to uncover what types of errors a child makes on a particular item. Using the weighted frequency rule, more than 80% of the 1, 104, 865 errors can be classified as coming from a certain strategy, and distinctions in age can be made. Importantly, since these errors provide information about which erroneous strategies children use to provide a response, the classification of these errors can provide a valuable tool in educational computer programs, as it allows for providing personalized feedback.

2.4 Discussion

The Math Garden ecosystem, like other large-scale learning environments, contributes to what Molenaar (2004) described as an opportunity to "continuously assess and store the unfolding life history (trajectory in behavior space) of each individual". However, although the phrase nicely catches the opportunities of today's educational data, it fails to draft the desired way forward. In this final section, we reflect on the research discussed in the current paper, and give important considerations for the future of learning analytics.

2.4.1 Active analytics

A primary characteristic of learning systems is their educational objective, and this objective should have a central place in learning analytics. Savi et al. (2015) argue that reaching a desired educational objective requires one to first accurately track a student's development, and subsequently map each student's learning route. That is, the spectrum of each student's ability should be assessed and tracked over time, such that an accurate learner model is created. This learner model may for instance encompass the discussed ability measures for various scholastic domains, or possibly some diagnosed misconceptions, and should give rise to the creation of an optimal learning path for this particular student.

To this end, we believe that learning analytics should be an *active* exercise. Rather than passively collecting analytics about a learning environment, learning analytics must actively help direct a student towards his or her educational objective—such as effortful practice on the level of the individual child in the case of Math Garden. Math Garden applies active forms of learning analytics on multiple levels. First and foremost, as laid out in the Methods section, it utilizes *embedded learning analytics*: the Elo rating system at the core of Math Garden estimates item difficulties and user abilities on the fly, and dynamically steers each student's learning experience in the desired direction.

Moreover, as we show in the Results section, making sure a learning system optimally directs the student towards the intended goal additionally requires active development. In this paper, we took model fit as the primary approach, and showed how different data selections and different methods shed diverse lights on the problem of misfit. We showed that although in general the adaptive system both accurately predicts student responses and reliably estimates item difficulties, these analytics may be biased by systematically misfitting responses. Multiple sources of this misfit were discussed, such as the possibly distinct processes underlying observed responses and local response strategies for subsets of items. Finally, the rich diversity of possible sources became evident when we discussed four more explorations of misfit, including a diversity in possible error patterns, and unexpected and undesired consequences of the used scoring rule.

The nevertheless good fit of the system illustrates that embedded learning analytics can help track and direct the development of an individual student. We hope to have conveyed that model fit can be seen as a central endeavor in learning analytics, with implications for very diverse parts of a learning system. Moreover, active analytics, such as the embedded learning analytics employed in Math Garden, need to assure that the system and its users reach their educational objectives.

Besides the embedded learning analytics, we believe a second form of active analytics deserves careful consideration: *experimentation*. The different sources of misfit in the Results section illustrate that without careful supervision, the ecosystem may move towards an unintended or even undesirable goal. Moreover, the necessary continuous maintenance of a large-scale online learning system like Math Garden unmistakably changes the system in both intentional and unintentional ways. In such a goal-directed system, these changes can alter the degree to which the goal is reached. Experiments serve to detect how an intervention alters the complex system, and to make sure it does not behave in unintentional and possibly detrimental ways.

An experimental method particularly suited to large-scale online learning systems is the online randomized controlled experiment (Savi, Williams, Maris, & van der Maas, 2017), commonly known as the A/B test. In the Other sources section, we briefly discussed one such experiment, aimed at preventing undesirable strategic responses that increase the misfit of the adaptive system. Additionally, besides using experiments to evaluate the mechanics of a learning system, experimental comparisons of pedagogical interventions can provide additional leverage. The learning sciences provide a wealth of possible interventions targeted at achieving learning gains, and often well suited for testing. Similarly, large online educational systems provide an exceptional testing ground for such interventions.

2.4.2 Conclusions

Although a vast share of research on learning is conducted within the safe boundaries of confined experiments, that is not where the actual everyday learning happens. Everyday learning happens *in vivo* — in a complex, dynamic, ecological system. Such a system is inherently difficult to track, let alone deliberately navigate towards a desired goal. Fortunately, an ever-increasing worldwide accessibility to the internet and serious efforts to scale learning technologies, increasingly succeed to unlock the big data of learning. These data, with an unprecedented granularity, combined with advanced methods, are now starting to provide a window into the complexity and dynamics of learning in vivo. In this paper we reported on a decade of experience from one such system, Math Garden. We described what we have learned and how we are still learning from a system that develops while we observe learning as it happens.

Nature has been experimenting since the beginning of time, with a boldness and complexity far beyond the resources of science.

Lee J. Cronbach

3

An experimental agenda

Summary

Although large-scale online learning increasingly succeeds in attracting learners worldwide, to date it fails to deliver on its promise. We first show the immense popularity of online learning and discuss its (unsatisfactory) effectiveness. We then discuss large-scale online randomized controlled experiments (A/B tests) as a powerful complimentary means to enable the desired leap forward. Although these experiments are widely and intensively used for web page optimization, and are slowly being adopted by the online learning community, their use, benefits, and challenges have only limitedly seeped through to the larger learning community. We summarize existing efforts in employing A/B tests in online learning, argue that such tests should take into account the typical nature of (online) learning, and encourage the use of knowledge from the various learning sciences to identify interventions that promise improved learning. We finally discuss both the limitations and promises of A/B tests, and show how such tests can ultimately contribute to learning that is tailored to each individual learner. The insights and priorities that arise from this overview and synthesis of A/B tests in online learning may help advance and direct the field.

3.1 Introduction

Online learning is sometimes attributed, rather boldly, the power to solve the 2 sigma problem: the finding that learners that are tutored one-to-one perform two full standard deviations better than learners that receive conventional instruction (Bloom, 1984). Although this is not its only promise, and 2 sigma likely is an overestimation (VanLehn, 2011), it does set a challenge. Indeed, online learning's large scale, easy adaptability, and inference from its generously generating data, are argued to enable unprecedented optimization and personalization and significantly increase learning gains. To date, however, the majority of online learning seems to fail to convincingly deliver on its promise, and both researchers and (online) educators seek to discover which components actually do make it truly effective. In this paper, we motivate the urgent need to prioritize the effectiveness of online learning and discuss a large-scale online experimental approach that is slowly being adopted by some providers of large-scale online learning, and that may potentially provide a leap forward. We ultimately argue that this approach is not just a powerful way to greatly increase the effectiveness of online learning, but also an opportunity to expand our knowledge of effective components in learning and education as a whole.

In the first section, we define large-scale online learning (hereinafter simply referred to as online learning) and show its rapidly increasing popularity, followed by a discussion of its disputed effectiveness. This should not only introduce less familiar readers to the field of online learning, but also convey the urgency to prioritize its return on investment. Return on investment may be defined as the increased achievement with respect to devoted time, or as the increased aspiration or perseverance in learning. We are however reluctant to pinpoint its definition, as it is greatly determined by one's personal educational philosophy and associated goals.

In the second section, we discuss the use of A/B tests (large-scale online randomized controlled experiments) to identify and iteratively optimize learning interventions for the online environment. We give examples of existing efforts to employ such experiments, followed by an extensive discussion of its most pertinent requirements and accompanying challenges. Finally, we discuss its limitations. Again, this should not only provide less familiar readers with a brief overview of the field of A/B testing in large-scale online learning, but should also locate pressing issues that need to be dealt with in order to advance the field.

3.2 A primer on online learning

Online learning comes in a great variety. It ranges from learning activities that resemble traditional education such as massive open online courses (i.e., xMOOCs, such as Coursera, edX, FutureLearn, and Udacity), to a plethora of activities characterized by a practice-approach and the use of various gamification elements (e.g., Codecademy, Duolingo, Khan Academy, KnowRe, and Math Garden), and ultimately full-fledged games that promote implicit learning through interaction with the game mechanics (e.g., DragonBox Algebra, Geniverse, Mars Generation One: Argubot Academy, Mathbreakers, Wuzzit Trouble, and Zoombinis). These activities may take place entirely online or blended with traditional education (e.g., Molnar, 2013), range from learning limited concepts (e.g., Slice Fractions, and Vax!) to earning complete degrees (e.g., Minerva Project, and Udacity's Online Master of Science in Computer Science in collaboration with Georgia Tech and AT&T), cover many subjects relevant from kindergarten to higher education, and are too used across virtually all ages.

We reflect this variety by defining large-scale online learning as any learning activity that is provided online and that scales. Large-scale online learning may thus be seen as an umbrella term for similar and related concepts such as blended or hybrid learning, distance learning, e-learning, intelligent tutoring systems, massive open online courses (cMOOCs and xMOOCs), and serious games, provided that these obey the terms in our definition. Naturally, lumping these together seems to do no justice to the rich history and diversity of each of the individual concepts, and in many contexts prudence is called for when comparisons between them are being made. However, in the context of the current paper it nicely illustrates the broad applicability of the experimental approach that is proposed and discussed in the second section.

3.2.1 The rise and rise of online learning

The number of users of online learning has rapidly increased the past few years. Although accurate figures are sparse, many of those figures pertain solely to the U.S., and moreover do not equally represent the whole spectrum of online learning, the few figures that are available do give a compelling picture of the popularity of online learning (a global group of organizations called 'GlobalOHER initiative' aims to provide survey-based figures on online higher education worldwide (ICDE, 2013), however no report was released yet). In the following we summarize these figures and purposely echo the terms used for online learning in their respective sources, since each may pertain to some class of online learning and not necessarily to the whole of online learning.

To begin with primary and secondary education (roughly aged 4 to 18), a survey by the private Evergreen Education Group (Watson, Murin, Vashaw, Butch, & Rapp, 2013) estimates that 310 000 U.S. students were enrolled in fully online schools in 2012/2013. Another survey by the U.S. National Center for Education Statistics (NCES) shows that enrollment of U.S. high school students (roughly aged 16 to 18) in distance education courses has rapidly increased over the past years (Aud et al., 2012). In 2002/2003 roughly 222 000 students enrolled, which increased to 310 000 students in 2004/2005, and 1.3 million students in 2009/2010. Twelve percent of the school districts serving these students in 2009 to 2010 reported that these students could fulfill all requirements for graduation with distant courses. Finally, a survey on teaching with digital games by the Joan Ganz Cooney Center estimates 74% of K-8 teachers to use digital games for instruction (Takeuchi & Vaala, 2014).

In fall 2012, the NCES also started collecting enrollment figures of U.S. post-secondary students (roughly aged 18 and above) in distance education courses (NCES, 2012) and issued the first report in 2014 (Ginder & Stearns, 2014) (summarized in Kena et al., 2014). The figures show that, in fall 2012, 2.6 million students were enrolled exclusively in these courses, whereas another 2.8 million students were enrolled in some but not all of these courses. The private Online Learning Consortium (former Sloan Consortium) also provides yearly estimates of the students of online learning in higher education in the U.S. In their 2013 report they estimate a total number of 7.1 million students (Allen & Seaman, 2014), although some argue that this might be an overestimation (Kolowich, 2014).

Also, the self-reported worldwide coverage of some large providers of online learning adds to this picture of popularity. For instance, the xMOOC provider Coursera reports to serve 7.5 million students worldwide and offer courses from 100 institutions (Ng, 2014), including leading universities such as Stanford and Yale. Codecademy, a platform for learning to code, reports to serve over 24 million students worldwide (Sims, 2014). Finally, Khan Academy, a video tutorial and practice platform for a multitude of subjects, reaches over 10 million unique students per month (Murphy, Gallagher, Krumm, Mislevy, & Hafter, 2014).

Naturally, learning is not limited to the academy. Consultants from Roland Berger, a large strategy consultancy firm, estimate that in Europe 3000 companies are involved in e-learning (Vernau & Hauptmann, 2014), and they expect an average increase of 13% per year. Also, Udacity and its spin-off Nanodegree partner with companies to build and provide courses specifically tailored to future employees (Chafkin, 2013). Other examples are edX Professional Education, HBX of Harvard Business School, the French First Business MOOC, and the IOC Athlete MOOC by the International Olympic Committee. Finally, also governments have

started adopting online learning, with the OpenupEd initiative in the European Union and Coursera's several governmental partnerships (Coursera, 2014a, 2014b) being notable examples.

3.2.2 All that glitters is not gold

Unfortunately, all that glitters is not gold. Here we briefly give a few examples from right across the spectrum of online learning, showing some of the struggles different online educators face in optimizing the return on investment of learning. As a first illustration: Sebastian Thrun, founder of Udacity, already warned in 2012 that his MOOCs were experimental and that he had not "seen a single study showing that online learning is as good as other learning" (Lewin, 2012). He substantiated this claim in 2013, when he revealed that despite Udacity's efforts to teach and engage students with courses that use quizzes and gamification techniques, only 7% managed to finish them (Chafkin, 2013).

Another example comes from Khan Academy: Greenberg, Medlock, and Stephens (2011) found that Khan Academy's lecture videos, a core component of the platform, were mostly avoided by its users. Instead of watching the videos, many users sought help from peers and teachers, and used hints from the platform. Both Coursera and Udacity, xMOOC providers that likewise heavily rely on videos, later confirmed this finding on their respective platforms (Simonite, 2013).

On the other end of the spectrum of online learning struggles are apparent as well. Although incomparable to MOOCs and on a decidedly different scale, in a synthesis of some of the research of the Community College Research Center the issue of low retention is also noticed in online courses in community colleges (Jaggars, Edgecombe, & Stacey, 2013). Moreover, they show that performance outcomes were lower for online than traditional courses, that students who took online courses were less likely to return the following semester, and that the more online courses they took the less likely they were to obtain a degree or transfer to a four-year institution. Additionally, achievement gaps that existed in traditional courses increased in online courses.

A slightly more recent study by the Public Policy Institute of California (Johnson & Mejia, 2014), also in community colleges, challenged the finding that students taking online courses are less likely to obtain a degree. Although in the short term similar disappointing results were found as by CCRC, results show that in the long-term students that take an online course are more likely to obtain a degree or transfer to a four-year institution than students that only take traditional courses. A multitude of explanations may account for increased long-term outcomes.

Students may for instance enrich their curriculum with online courses such that they are better prepared for further education, or it might simply be the most motivated students that take online courses. As the authors put it: "[f]or some students, online courses offer a useful tool that helps them reach their goals". Naturally, this does not necessarily mean that the online courses are effective on their own and the short-term results may signal that online courses, with their lower completion rates and larger achievement gaps, under-perform to traditional courses.

Possibly disappointing findings such as the few we summarize here naturally influence the perceived promise of online learning. This is for instance reflected in the aforementioned survey by the Online Learning Consortium (Allen & Seaman, 2014), which showed that, compared to 2012, in 2013 proportionally less academic leaders indicate that "online learning is critical to their long-term strategy" and proportionally more academic leaders think that "the learning outcomes for online education are inferior to those of face-to-face instruction". Moreover, only a small minority believes that "there will no longer be concerns about the relative quality of online courses".

Summarizing, we have seen that learners and educators worldwide rapidly adopt online learning, and that companies and countries jump on the bandwagon. To date online learning has enabled affordable learning activities for a broad public and many around the world already benefit greatly. We have however also seen that its effectiveness leaves a lot to be desired. Taking together its inevitable growth, its according impact and role within education as a whole, and an unarguable need to increase its effectiveness, necessitates the use of a robust and reliable method to identify interventions that promise improved learning.

3.3 An experimental approach to improve online learning

In this section, we focus on one such method. We discuss the use of large-scale online randomized controlled experiments (i.e., A/B tests): a within-platform approach for iterative and incremental improvement of large-scale online learning. The popularity (and accordingly large scale) of particular online learning environments calls, at least partly, for such an online-only and within-platform approach. Moreover, we focus on A/B testing as we think it is an opportunity that receives too little attention, in spite of its broad applicability. The use of such experiments within the online learning community is quite limited, especially in comparison to web page optimization outside of the online learning context, and also not well-known in the larger educational community. And although it will not turn all glitters into gold, it certainly is a powerful complimentary method that may potentially provide a significant contribution.

In the following we introduce the use of A/B tests, give examples of existing efforts to employ such experiments in online learning, discuss important challenges that are specific to the study of online learning, and meet the inevitable limitations of the method.

3.3.1 A/B testing

Both educational practice and research benefit greatly from randomized experiments (Slavin, 2002). Traditionally, different methods are used for varying levels of (causal) inference, such as small-scale laboratory experiments on precise learning interventions, and large-scale randomized controlled trials (RCTs) for comparing distinct educational methods. In online learning, the A/B test (also known as 0/1 test, bucket test, content experiment, parallel flight, and split test) provides us with an online RCT that is intensively used in website optimization and that is both convenient and robust. In a typical A/B test different variations of the same web page are shown to different randomly chosen groups of visitors. After a specified amount of time, the groups are compared with respect to an evaluation criterion (Kohavi, Longbotham, Sommerfield, & Henne, 2008). In other words, in an A/B test a large number of visitors is randomly assigned to a control or treatment variant in order to discern the effects of the treatment variant. Such an experimental test is particularly strong since it enables the establishment of causal rather than correlational relationships. Moreover, A/B tests can be extended to involve more than two conditions (i.e., A/B/n tests) or additional variables (i.e., multivariate tests).

Conducting A/B tests reliably requires a tremendous number of users (Kohavi, Deng, Longbotham, & Xu, 2014), however in the first section we saw that online learning can actually provide that scale. The opportunity this opens up for learning research is unprecedented. In traditional learning mass experimentation (in the form of traditional RCTs) is troublesome and costly, whereas the advent of online learning offers plenty of opportunity for cheap mass experimentation. Moreover, in an online environment it is not only much easier to allocate learners randomly, but also to readily and homogeneously implement the experimental interventions.

Finally, it is important to stress that A/B tests differ with respect to one more essential aspect from their offline randomized controlled counterpart. Whereas traditional RCTs for instance enable comparisons between situations that may differ quite significantly (with all accompanying challenges, e.g., G. Norman, 2003), A/B tests on the other hand are best suited for the comparison of more precise interventions, as its online nature enables very homogeneous and precise adjustments. A/B tests are characterized as minimally invasive (Heffernan & Heffernan, 2014), and enabling iterative improvement (Williams et al., 2014). Summarizing, on the one

hand A/B tests share their precision and ease of randomized allocation and double-blinding with traditional laboratory experiments, whereas they share the scale and ecological validity with traditional randomized controlled trials.

3.3.2 Existing efforts

Although the (reported) use is very limited, some online educators are already performing A/B tests, or are currently preparing their frameworks to enable the use of these tests. Examples are Coursera (Novet, 2013), Duolingo (Ungerleider, 2014), edX ("Content Experiments," n.d.), Khan Academy (Rowan, 2013), and Udacity (Simonite, 2013). At Coursera they discovered, using large-scale A/B tests, that students' engagement dropped when those students were reminded of their homework, whereas a reminder of their previous activities increased engagement, and at Udacity they discovered that black and white rather than colorized lessons gave better test results (more details were not provided; Simonite, 2013). At Khan Academy, they discovered that showing a sneak peek of more advanced content demotivated learners to continue studying (Fox, 2014), that a mindset intervention in the header of math problems increased mastery of those math concepts (Williams, Paunesku, Haley, & Sohl-Dickstein, 2013), and they have tested different models for assessing students' proficiencies (Hu, 2011). At Duolingo they discovered that introducing the English pronoun 'it' too early in the curriculum confused Spanish students as the word is sometimes used differently in both languages (Stevenson, 2014). Finally, quite some A/B tests were conducted on both the edX framework (some of which were summarized by Reich, 2015) and the ASSISTments tutoring system (Heffernan & Heffernan, 2014). Unfortunately, results from A/B tests such as summarized here are rarely found, especially in the scientific literature. Given the recent efforts among online educators to incorporate A/B tests, we do however predict an increase in, hopefully publicly available, studies.

3.3.3 Tailoring A/B tests to online learning

A/B tests have great potential for determining and increasing the effectiveness of online learning, however blindly adopting A/B tests from the domain of website optimization without taking into account the specificities of online learning will fail in fully exploiting that potential. A/B tests are often used in website optimization for relatively simple tweaks, such as textual or graphical changes to for instance the landing page of a retail website. The overall evaluation criterion (i.e., dependent or outcome variable) is in its turn usually expressed as a conversion rate, which is the rate to which visitors or users act in some desired way; this may be the rate to

which visitors for instance buy a product from the retail website, or in case of online learning the rate to which users study the course material, watch an instructional video, or continue to practice math problems. However, although those relatively simple interventions may indeed increase the effectiveness of a learning experience, presumably the real progression comes from interventions that are both not that easily implemented and varied, and not that easily captured in a conversion metric (such as measures of transfer, Barnett & Ceci, 2002). Rather, maximizing the return on investment of online learning has two important requirements. First it requires determining and tracking each learner's position on an educational map, and second it requires offering individualized navigation towards an educational goal (Savi et al., 2015).

For the first requirement, determining a learner's position, one accessible and accordingly popular method is to use proxy measures for learning, such as frequency of play, time on task, or proportion correct. However, such proxy measures carry important challenges, with the most pressing one being which measures best capture learning (Reich, 2015). On top of that, we deal with a variable that is latent (e.g., ability, learning) rather than manifest (e.g., profit, conversion rate), and one that is inherently subject to change, requiring a microgenetic method (Siegler & Crowley, 1991).

Powerful yet involved methods that can fulfill the above requirements include Intelligent Tutoring Systems (ITS) and Item Response Theory (IRT). One excellent example of the successful implementation of the latter, in a large-scale educational context, is Math Garden's computer adaptive math practice platform (Klinkenberg et al., 2011). Math Garden uses an IRT approach to continuously estimate and update the abilities of learners and difficulties of problems, and an Elo Rating System (ERS) to adaptively match learners with problems that have the desired difficulty.

The proper implementation of such methods is far from trivial and may not be applicable in each and every online educational method. However, tracking or monitoring a learner's development through these and similarly sophisticated methods can open up new opportunities for measurement (e.g., Brinkhuis et al., 2018), as these enable measuring the development of each individual learner and estimating the idiosyncratic effects of an intervention.

For the second requirement, individualized navigation towards an educational goal, accurate tracking is again very beneficial. Tracking each learner's development, and identifying for instance their misconceptions or level of expertise, can aid in adapting interventions to individuals, as we have to deal with both individual and contextual differences. For instance, cognitive variability is notoriously large (Siegler, 1994), and may contribute to considerable individual differences. Such differences, for instance in level of expertise, can require quite distinct in-

terventions (Kalyuga, Ayres, Chandler, & Sweller, 2003). Meaningful contextual differences on the other hand, for instance across domains (such as reading and math, or addition and multiplication), may likewise require distinct interventions. Solely assessing the main effect of an intervention, which is usually done in simple A/B tests, will not suffice in an educational context. Rather, it must be assessed how each of the different conditions serve the different individuals and contexts (Savi et al., 2015; Williams et al., 2014).

Once these individual and contextual differences are identified, A/B tests can help determine which learner benefits from which intervention at which moment. This way, either the effect of a global intervention (i.e., targeting all learners in the environment) can be determined with respect to the different individual and contextual characteristics, or the effect of a local intervention (i.e., targeting a relevant subgroup, such as learners with a specific misconception or level of expertise). The former can be considered a grapeshot method, revealing which individual differences are meaningful with respect to the intervention, yet potentially hurting certain subgroups that do not benefit from the intervention or are even harmed by it. On the other hand, the latter is more specific and especially beneficial as it provides the first step towards fully adaptive interventions.

A final consideration in tailoring A/B tests to online learning is the intervention to choose. As mentioned previously, whereas in traditional website optimization a simple intervention such as coloring the buy button might suffice; optimizing the learning rate of individual learners requires more effort. Fortunately, the learning sciences provide a wealth of knowledge on pedagogical interventions that likely improve achievement and that can be scaled, tested and implemented online. Examples of interventions that have already proven to scale, as diverse as increasing the motivation to learn, optimizing the exerted effort in learning, and providing proper guidance in the learning process, include mindset interventions (Paunesku et al., 2015), spaced practice (Xiong & Beck, 2014), and personalized feedback (Piech et al., 2015).

3.3.4 Limitations of A/B tests

Measuring up to the requirements discussed above will enable the full potential of A/B tests in online learning, however its opportunities naturally have limitations. The first is its online nature. For instance, most online learning settings allow control over delivery of content and interactions with technology, but do not allow for tight control over the amount and time of study; students learn at their own pace and time. Allied to that, students may receive additional education that enhances similar abilities that are trained in the online environment. Also, online

learning likely has a selection effect; students using learning online might not be representative of the total population of students (e.g., due to self-selection). These problems arise due to the in vivo nature of A/B tests. Although randomization and large subject pools may eliminate possible bias coming from the treatment assignment, heterogeneity likely is large and in order to secure high internal validity one must anticipate serious challenges (similar and other challenges of A/B tests are for instance discussed by Kohavi et al., 2014; Kohavi et al., 2008; Lu & Liu, 2014).

Also, as mentioned previously, whereas comparisons in traditional (offline) RCTs can go across institutions and cover distinctly different educational methods that may or may not include online components, A/B tests are limited to a single online platform. Typical A/B tests are therefore limited in breadth and support mainly incremental steps, ultimately facing the risk of ending up in a local maximum (i.e., optimal learning given the constraints of the platform that is used). Although offline RCTs and A/B tests thus utilize the same principals, they do serve quite distinct goals. A/B tests are most appropriate to identify the interventions that contribute to the development of an effective platform for learning, are less suited to the comparison of entirely different online educational frameworks, and naturally fail if comparisons are made with traditional (offline) learning. Thus, although performing offline RCTs in online learning remains a serious challenge (Lack, 2013), they do complement A/B tests and fill a gap.

Another gap that A/B tests fail to fill is the vast amount of non-experimental data online learning accumulates, which is complementary to experiments. Learning data, from demographics to activity measures and error patterns, provide a rich source of (non-experimental) information and may also guide the development of effective platforms for learning (Long & Siemens, 2011). For example, it may reveal the level of expertise or typical misconceptions of a learner. As discussed previously, accurate identification of such expertise or misconceptions may in turn enable the use of increasingly personalized interventions. Machine learning provides powerful data mining techniques that are most commonly used for this purpose. The combination of these exploratory techniques and confirmatory A/B tests may proof particularly powerful, since hypotheses can be generated from exploring the non-experimental data and A/B tests can experimentally decide which hypotheses are indeed fertile.

Whereas the previous limitations result from technical limitations of online experimentation, we end with an ethical consideration. Although online experimentation is common (Christian, 2012), some A/B tests have been received with mixed reviews. For instance, in 2014 complaints were filed against Facebook for a large-scale experiment they performed on emotional contagion, and European privacy regulators are examining the case (Goel, 2014). Within the domain of learning there too is a longstanding debate among proponents and critics of experi-

mentation. Naturally, Institutional Review Boards watch over the boundaries of what is generally accepted. But although traditional educational institutions too are starting to acknowledge the desirable effects of experimentation (e.g., Coughlan, 2014), the controversy around the Facebook study does demonstrate a serious societal call to consider clear ethical guidelines. Since the emergence of online experimentation is recent, these guidelines need to be actively developed and shaped, and one may anticipate changes.

3.4 Discussion

Although online learning has already made its promise of being highly accessible and affordable, its unsatisfactory effectiveness seriously constrains its full potential. The discussed experimental approach, notwithstanding its limitations, provides a powerful opportunity to increase its effectiveness and thus promises to provide a much-desired leap forward. Furthermore, and most importantly, the insights and priorities that arose from this overview and synthesis of A/B tests in large-scale online learning should help advance and direct future research in this field, enabling its many benefits.

First of all, consulting the different learning sciences and translating robust interventions to online environments, may already greatly improve online education. Second, whereas too commonly so-called 'B tests' are used (i.e., blindly adopting an intervention without verifying its effectiveness in a randomized experiment, and thereby taking the risk of unknowingly introducing detrimental interventions), A/B tests help increase an intervention's impact by not only verifying its effectiveness, but also enabling subsequent tweaking in order to optimize the effect, and thus by making sure it is evidence-based. As a step beyond typical laboratory experiments external validity is maximized, since the interventions are evaluated on exactly the platform that is used for the actual learning (Brinkhuis et al., 2018). Finally, tailoring these interventions to each individual learner will most likely be one of the biggest challenges, but likewise has great potential for increasing the return on investment of online learning.

An experimental approach also provides an opportunity to answer a widely-heard call for more research on the effective components of online learning (e.g., Means, Bakia, & Murphy, 2014). Despite the rapid growth and adoption of online learning, less evidence has surfaced on what these components are. As it appears that more A/B tests are actually conducted than being reported, we encourage both academic and corporate researchers in online learning to share their data, and disseminate and replicate their experimental (null) findings (Makel & Plucker, 2014, showed that replications are scarce in educational research) (Franco, Malhotra, & Simonovits,

2014, showed that null results are often not written up and submitted). With the use of for instance the 'conceptual framework for describing online learning' (Means et al., 2014), these findings will all help build an evidence-based body of knowledge on effective components in online learning.

Moreover, as an additional benefit A/B tests are an excellent tool for triangulation. Not only can findings corroborate for instance laboratory experiments and classroom observations, their large-scale, double-blinding, and ecological validity offer a distinct means to discern causal effects in the learning sciences. Indeed, the problems encountered in offline educational RCTs are profound: a double-blind procedure and strict randomization, essential ingredients of proper randomized controlled experimentation, are often impracticable. This is particularly problematic since large sample sizes are notoriously difficult to obtain while effect sizes of learning interventions are often small. The online environment helps address these issues, as large sample sizes are relatively easy to obtain, enabling a greater sensitivity to relatively small effect sizes and generally more reliable results.

One-to-one tutoring and its attributed effectiveness might remain an eternal ambition that is never fully achieved, but it does set the ultimate challenge and surely can be approached by accurately and continuously assessing learners, and tailoring the learning interventions to each and every individual. In the current paper, we have focused on one complimentary approach, that is broadly applicable and minimally invasive yet is only to a limited extent adopted by the online learning community, and moreover receives little attention in the larger learning community. A/B testing is an online-only and within-platform approach that enables evidence-based iterative improvement of online learning. Deploying A/B tests in online learning however requires us to move away from aiming interventions at the general population of learners, and to seriously take into account the intricate nature of not only learning in general, but also the inseparable ecological online environment, and the individual and contextual differences. A/B tests, when deployed correctly, provide a powerful opportunity that may help determine the effective components in online learning and eventually contribute to an increase in return on investment.

With practice, you can make great paintings.

Damien Hirst

4

Return of effort

Summary

We report on an online double-blind randomized controlled field experiment (A/B test) in Math Garden, a computer adaptive practice system with over 150 000 active primary school children. The experiment was designed to eliminate an unforeseen opportunity to practice with minimal effort. Some children tend to skip problems that require deliberate effort, and only attempt problems that they can spontaneously answer. The intervention delayed the option to skip a problem, thereby promoting effortful practice. The results reveal an increase in the exerted effort, without being at the expense of engagement. Whether the additional effort positively affected the children's learning gains could not be concluded. Finally, in addition to these substantial results, the experiment demonstrates some of the advantages of A/B tests, such as the unique opportunity to apply truly blind randomized field experiments in educational science.

4.1 Introduction

One of the main challenges in education research is to unravel causal relations. Randomized controlled trials are widely viewed as the gold standard in studying causal effects (Athey & Imbens, 2016; Borghans, de Wolf, & Schils, 2016; Slavin, 2002). However, the use of

RCTs in education research is not uncontroversial. The main critiques are that they are expensive, take long to conduct, and only provide answers to narrowly defined questions. Moreover, while double-blinding is deemed essential to avoid experimenter effects in medical research, so far, this turned out to be near-impossible in education research (Deaton & Cartwright, 2016; Olson, 2004).

In this paper, we show that large-scale experiments in online learning environments, also referred to as A/B tests, can be used to solve some of these issues. We report on a successful application of an A/B test in a large-scale online computer-adaptive practice system for Dutch primary schools (Math Garden, with over 150 000 active users). In the A/B test we delayed the option to skip a problem. This option was used by some children to skip difficult problems and practice with minimal effort, and the delay was thus aimed at promoting more effortful practice. Before describing the experimental details, we first aim to build a basic understanding of A/B tests in relation to traditional educational experiments, introduce the online practice system that is central to the experiment, and then discuss our motivation for this particular intervention.

4.1.1 A/B tests

A/B tests, the online equivalent of randomized controlled field experiments, are widely used by internet companies. In this section, we shortly dedicate some specific attention to the method of A/B testing, as there are relatively few applications in the field of online learning, especially in comparison to the thousands of A/B tests that large internet companies perform on a yearly basis, while the methodology has opened up massive opportunities for learning research.

Because of the huge scale of online learning, A/B tests enable mass experimentation that is virtually free of charge. There are no recruitment and data-collection costs, as participants already use the system and responses are tracked. Also, randomization is effortless, and adjustments to the environments can be made readily, precisely, and homogeneously. This is the reason A/B tests are sometimes said to be minimally invasive (Heffernan & Heffernan, 2014) and enable iterative improvement (Williams et al., 2014). Using A/B tests, we can successively test changes to learning environments to find out which components are effective. One might thus argue that A/B tests combine the scale and ecological validity of RCTs and the precision of laboratory experiments.

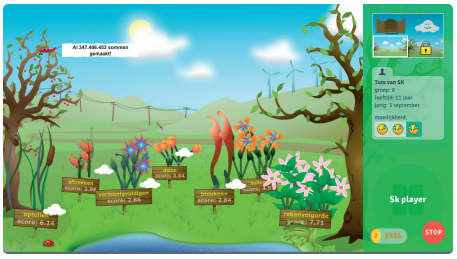
Importantly, a profound criticism of educational experiments, the practically near impossibility to satisfy a double-blinded procedure (e.g., Olson, 2004), does not pertain to A/B tests. Interventions in online learning environments neither need to rely on teacher instructions, nor

need to be necessarily noticeable for the students. Let alone that the hypotheses that drive those changes need to be known to either teachers or students. A/B tests thereby have the power to effectively eliminate experimenter effects from educational experiments. This is not to say that A/B tests are a panacea. A/B tests only suit large-scale online education and are restricted to single platforms, consequently problems like external validity still require attention.

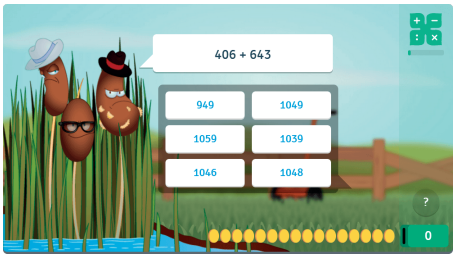
4.1.2 Math Garden

In this paper, we illustrate the method of A/B testing in the online learning environment Math Garden. We aim to reduce problem skipping and promote effortful practice. Before discussing the experiment, we first introduce the environment. Math Garden is an online environment for adaptive practice of math and math-related domains, spanning from addition and multiplication to logical reasoning and working memory. The system is used in over 1500 primary schools in The Netherlands, and currently has over 150 000 active users, that collectively respond to more than 6 million items on a weekly basis. Such scale, its numerous sister systems for languages, typing, and statistics, and the symbiotic relationship between research and practice, provides an ideal basis for scientific research and continues to result in both methodological and substantive papers. Only some of the most recent research concerns topics ranging from the development of typewriting skills (van den Bergh et al., 2015), non-formal mechanisms in cognitive development of arithmetic (Braithwaite, Goldstone, van der Maas, & Landy, 2016), and number transcoding in a language with inversion (van der Ven, Klaiber, & van der Maas, 2016), to self-adapting success rates in math practice (Jansen, Hofman, Savi, Visser, & van der Maas, 2016).

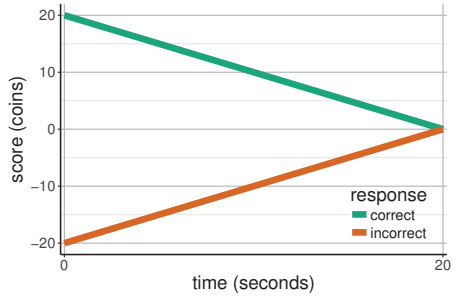
Children that use Math Garden maintain a virtual garden, with different plants representing different domains and the health of a plant reflecting the frequency of practice. By selecting a plant, the child starts to practise a set of items within that domain. The system uses item response theory to estimate child abilities and item difficulties, and uses the Elo rating system to adaptively match children to items in real-time (Klinkenberg et al., 2011). In order to aid the accurate estimation of abilities and difficulties, a scoring rule with a speed-accuracy trade-off is employed (Maris & van der Maas, 2012). A response must be given within a certain time limit, which is visualized by a diminishing number of virtual coins at the bottom of the screen. Correct responses are rewarded with the remaining coins, whereas incorrect responses are punished by subtracting the remaining coins. Failing to give a response before the deadline results in neither a reward nor a punishment. After each item, the correct answer is shown, and the child proceeds



(a) Example of a virtual garden. Plants represent domains. The smileys with different numbers of drops of sweat represent the difficulty levels.



(b) Example of an item from the addition domain. The remaining time (i.e., number of remaining coins) and the question mark button are shown on the bottom.



(c) The scoring rule. Rewards and punishments decrease linearly with time.

(d) A full trophy cabinet.

Figure 4.1. Math Garden.

to the next item.

Each successful completion of a set of items within a domain earns the child some additional coins. The collected coins can be used to buy different kinds of virtual trophies. To cater individual differences with respect to desired difficulty, children may select the difficulty level themselves. This is reflected in the expected proportion correct (0.9 for easy items, 0.75 for medium items, and 0.6 for hard items). The rewarded/subtracted coins are doubled when using the hard level, and halved when using the easy level. Children may skip items that they deem too difficult to answer by hitting a question mark button. They are shown the correct answer and neither earn nor lose coins using this strategy. However, the adaptivity of Math Garden should generally prevent matching a child with an item that is too difficult. In Figure 4.1 we show some of the above elements.

4.1.3 Effortful practice

A major aim of adaptive practice systems like Math Garden is to present problems at the level of the student. Other than in a traditional classroom environment, where in its most extreme case all students work through the same problems in the same pace, adaptive practice systems function as individual tutors. Vygotsky's theory of the zone of proximal development (Vygotskii, 1978) is central to this practice (Murray & Arroyo, 2002), and adaptive practice systems can be viewed as systems that seek to explore what a student can do with instruction or the outer boundary of what a student can do without instruction (depending on the level of instruction in the system). Specifically, Math Garden exploits the estimated difficulties of the problems, and makes sure each student receives little to no problems that are either too easy or too hard, and thus by balancing on the boundary of what a student can do without instruction.

By exploiting the zone of proximal development and delivering individual tutoring, adaptive practice systems seek to optimize learning gains. In return, this requires a serious and continuous effort from the student, as they are performing on the edges of their abilities. Not only does this take a great deal of motivation from the student (e.g., Pintrich, 1999), students do not always recognize the importance of effort for effective learning, and sometimes even falsely assume that easy problems are better for learning (Bjork, Dunlosky, & Kornell, 2013). Therefore, Math Garden aids students directly in their motivation to practice problems by means of the virtual coin incentive, and indirectly by giving students the option to move closer towards or further away from the edge of their ability by means of the difficulty level selection.

4.1.4 Problem skipping

In spite of these motivational aids, students still find ways to avoid difficult items, and for the current study Math Garden's question mark button is of particular interest. We noticed that some children use the question mark relatively often and relatively fast, which is probably best explained as strategic behaviour. Children that aim to maximize their earned coins pursue fast correct responses, and benefit from quickly skipping those items that they cannot spontaneously answer.[1] This strategy moves the child out of the zone of proximal development and severely reduces the amount of exerted effort.

[1] Math Garden already utilizes one prevention for fast incorrect or question mark responses. Children are logged off from a domain if they submit x or more incorrect and/or question mark responses within the first 3.5 seconds, where x equals the number of items in the set, divided by 3, rounded to the nearest integer, and with a minimum of 3 and a maximum of 9 of such responses.

Two reasons justify the aim to prevent this strategic behaviour. From a learning perspective, the behaviour relates at most, if at all, to surface learning. Those children do practise, but primarily by repeating known problems. Although this benefits memorization of those problems, it obscures the learning of new problems. The biggest learning gain is to be found in the zone of proximal development, and will require active and effortful learning.

Also, from a measurement perspective, the accuracy of the obtained ability and difficulty estimates increases when the children behave according to the scoring rule. Question marks do not provide clear information about children's abilities. The most accurate ability estimates can be computed for children that put in as much effort as they can and respond as soon as they think they have come up with the correct answer. Ultimately, accurate ability estimates benefit the adaptivity of the system.

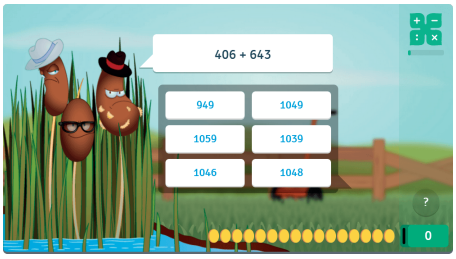
4.1.5 Minimum toil time

In order to prevent question mark misuse, we designed an A/B test to test whether a straightforward delay on the appearance of the question mark button would promote more effortful learning. For children that do not directly know the correct answer to an item, this delay can be seen as the minimum required toil time. We expect that those children will resort to more effortful strategies. After all, fast guesses are relatively expensive (an incorrect guess results in a punishment), and effortless waiting until the question mark becomes available costs time and decreases the potential reward. Following this reasoning, we expect children in a toil time condition to use the question mark button less frequently.

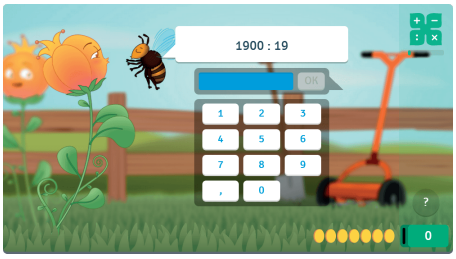
4.2 Methods

4.2.1 Experimental domains

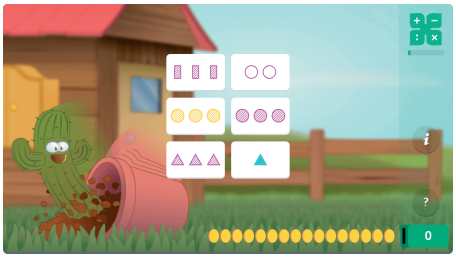
The experiment was performed in three separate game domains: addition, division, and one-two-three. One-two-three is an implementation of the popular logical reasoning game Set (e.g., Nyamsuren & Taatgen, 2013). Figure 4.2 shows an example item from each of these domains. In all three domains, a one game session contained ten items, after which the child was given the opportunity to choose the same or a different domain. In the addition and division domain each item had a deadline of 20 seconds, whereas one-two-three had a deadline of 30 seconds. Also, by default the former domains were available to all children, whereas the latter only became available after a child had sufficiently practised the base domains (this default setting could be



(a) Example of an item from the addition domain.



(b) Example of an item from the division domain.



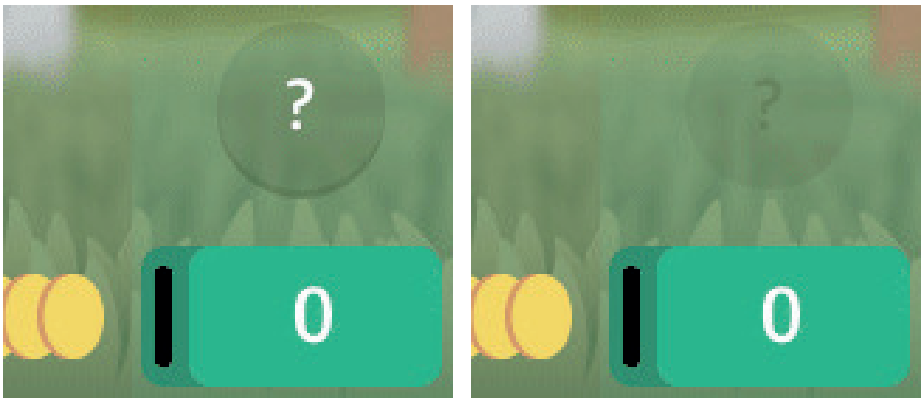
(c) Example of an item from the one-two-three domain.

Figure 4.2. Experimental domains.

changed by individual teachers for individual children). Finally, addition items had a multiple-choice format, whereas division and one-two-three items were open-ended.

4.2.2 Participants

A total of 107, 979 Math Garden users participated in the experiment, mostly children aged 4 to 12. Math Garden is used in ecological settings, at school and at home, on different devices, and during the whole day and week, but mostly during school hours. Children that indicated that they did not want to be part of the scientific research done in Math Garden were excluded from the analyses. The procedure was approved by the department's Ethics Review Board.



(a) Active question mark button.

(b) Greyed out and inactive question mark button.

Figure 4.3. Visualizations of the question mark button.

Allocation

Participants were randomly distributed across the four conditions[2]: the question mark button was either active (control) or greyed out and inactive for 3, 6, or 9 seconds. Randomization was done separately within each game domain. Figure 4.3 shows the visualizations of an active and an inactive question mark button.

Exclusion

The intervention relied on the CSS property *pointer-events*[3], which is not supported by some older Internet browsers. In all conditions, we excluded all children that used an incompatible browser ($n = 9,665$). Browsers and browser versions were recovered from the user agent id's that are recorded with each response, using the R implementation of *ua-parser*[4]. Nonetheless, a manipulation check revealed that 39 children with seemingly compatible browsers did have question mark responses before the question mark delay ended. As there is no reason to believe that the responses from children that used an incompatible browser relate in any way to the objective measures of this study (e.g., question mark use), we did not exclude these users from

[2] The user id's were transformed using a bitwise right shift of 0 in the addition domain, of 2 in the division domain, and of 4 in the one-two-three domain. We then used a modulus to transform each id into one of the four conditions.

[3] https://developer.mozilla.org/en-US/docs/Web/CSS/pointer-events

[4] https://github.com/ua-parser/uap-r

the analyses. We neither expect that the remaining 59 illegal responses (55 in the addition domain, 1 in the division domain, and 3 in the one-two-three domain) from those 39 children will have any substantial effects on the outcomes of the study.

Cross-validation & peer review

The huge scale of the experiment allowed us to cross-validate the effects. Moreover, in consultation with the journal editor, we followed a novel procedure to further improve the reliability of the results. We randomly selected half of the participants for the performed analyses (practice set; $n = 50,433$, excluding participants with an incompatible browser). The provisional report, which was solely based on the analyses on the practice set, was then subjected to formal peer review. After acceptance by the editor and reviewers, the results were verified on the other half of the participants (test set; $n = 50,267$, excluding participants with an incompatible browser). In this final report, we additionally report the results from the test set, but only if these deviate from the results from the practice set. This procedure, in the spirit of pre-registration[5], ensures that the methods need to be reviewed and assessed independent of the results, and that possible capitalization on chance during the analysis and review phase is corrected for by the cross-validation.

Distribution

In Table 4.1 we summarize the number of participants for different selections of the data, excluding participants with an incompatible browser. Be aware that children can be in different conditions for different domains.

4.2.3 Duration

The experiment was performed in 2016, from March 16 to June 22, spanning a total of 14 weeks. The period is a multiple of weeks to eliminate day-of-the-week effects.

4.2.4 Software

Analyses were performed using *R* (R Core Team, 2016) and *RStudio* (RStudio Team, 2015). Figures were created with the *R* package *ggplot2* (Wickham, 2009).

[5] https://www.apa.org/science/about/psa/2015/08/pre-registration.aspx

domain	condition	practice set	test set
addition	no delay	11866	11739
addition	3s delay	11889	11661
addition	6s delay	11600	11794
addition	9s delay	11740	11714
division	no delay	5636	5763
division	3s delay	5696	5584
division	6s delay	5594	5675
division	9s delay	5549	5622
one-two-three	no delay	7160	7012
one-two-three	3s delay	7015	7158
one-two-three	6s delay	7261	7060
one-two-three	9s delay	7134	7040

Table 4.1. Distribution of participants across domains and conditions.

4.3 Results

We used linear regression analyses, with dummy variables for the conditions, to discern the effects of the different question mark delays. First, we evaluated the decrease in question mark use and made sure the delay does not affect engagement. Second, we evaluated the speed and accuracy of substitute responses to the question mark. We report standardized beta's, such that the relative strengths of the effects can be evaluated.

4.3.1 Question mark delay decreases question mark responses

First, we evaluated the decrease in question mark responses, and thus in problem skipping. In Figure 4.4 we show the weekly proportions of question mark responses, averaged across participants and difficulty levels. We also show how these differ across the experimental domains.

A visual inspection of Figure 4.4 clearly reveals a structural decrease in the proportions question mark responses with increased question mark delay. For instance, if in the addition domain the question mark button is not delayed, children tend to skip roughly 10 to 12% of the problems. With a 3 seconds delay this percentage is reduced to roughly 8 to 10%, and with a full 9 seconds delay only roughly 3 to 4% of the problems is skipped. Interestingly, these effects seem decidedly smaller in the one-two-three domain. We'll return to this issue in the Discussion section.

We used linear regression analyses with backward difference coding in order to find the effects of the additional increases in question mark delay. Thus, the 3 seconds delay is compared to the control, the 6 seconds delay is compared to the 3 seconds delay, and the 9 seconds

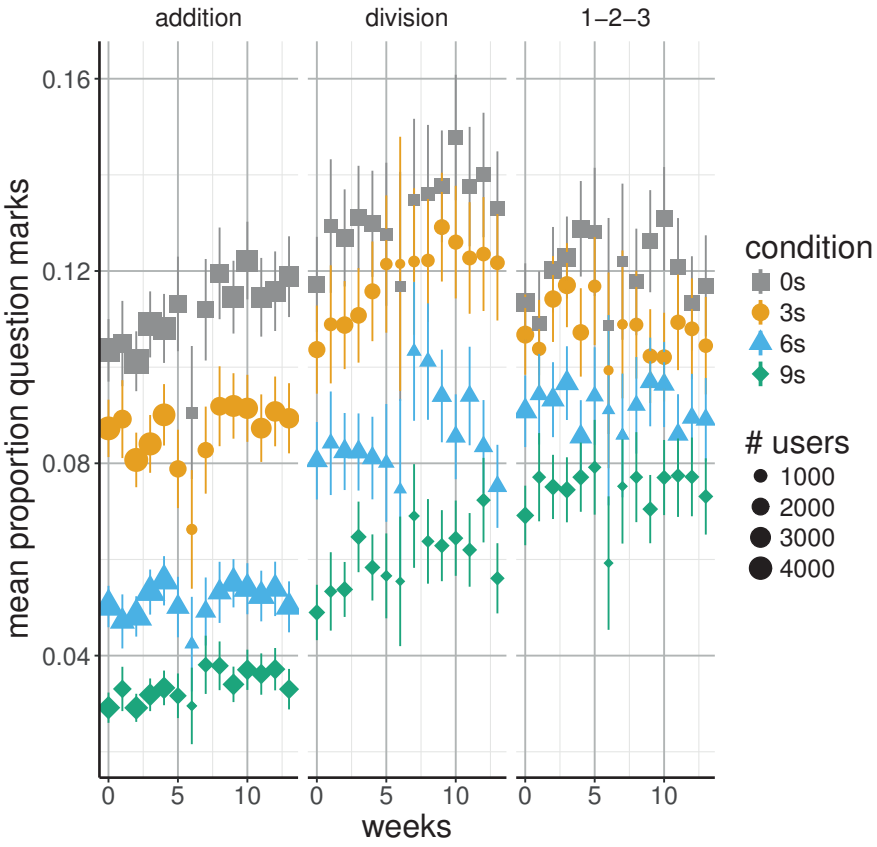


Figure 4.4. Average proportions of question mark responses across participants and difficulty levels, by week. Panels represent domains. Error bars represent 95% confidence intervals.

delays is compared to the 6 seconds delay. The analyses confirm the differences in question mark use across conditions. Table 4.2 shows that each additional question mark delay adds up significantly in decreasing the proportion of question marks used (all $p < .01$). In the test set, these results were confirmed, although the 3 seconds delay in the one-two-three domain was found to significantly decrease the proportion of question marks used with $p = .027$.

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.000	0.004	-0.000	1.000
addition	3s delay	-0.061	0.005	-13.214	0.000
addition	6s delay	-0.094	0.005	-17.492	0.000
addition	9s delay	-0.042	0.005	-9.081	0.000
division	(Intercept)	0.000	0.006	0.000	1.000
division	3s delay	-0.039	0.007	-5.512	0.000
division	6s delay	-0.074	0.008	-9.028	0.000
division	9s delay	-0.060	0.007	-8.550	0.000
one-two-three	(Intercept)	0.000	0.005	0.000	1.000
one-two-three	3s delay	-0.018	0.006	-3.002	0.003
one-two-three	6s delay	-0.037	0.007	-5.244	0.000
one-two-three	9s delay	-0.042	0.006	-6.892	0.000

Table 4.2. Linear regression results for experimental differences in the proportion question mark responses, separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the preceding delay, in order to find the effects of the additional increases in question mark delay. Standardized betas are reported ('estimate').

4.3.2 Question mark delay has no adverse effects on time on task

Preferably, the question mark delay intervention has no adverse effects on engagement. When children consider the delay annoying, they might decide to practice less. To rule out the possibility of such an adverse effect, we checked whether the question mark delay conditions differed with respect to the readily available proxy-measure time on task. First, time on task (in minutes) was computed by summing the response times separately for each participant during the experimental period. We expected no differences in time on task between conditions, and thus compared each intervention condition (i.e., 3, 6, and 9 seconds question mark delay) directly with the control (no delay).

The results of the linear regression analyses are summarized in Table 4.3. No significant differences were found, except for the 9 seconds delay conditions in the addition and division

domains. Both effects suggest that with a 9 seconds delay children spend more rather than less time on the addition and division tasks. However, we are reluctant to give these effects too much weight, as the modest standardized beta's of 0.011 and 0.027 point to negligible effects that possibly originate from the huge amount of power of the study. In the test set, these results were confirmed, as no significant differences were found.

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.000	0.005	-0.000	1.000
addition	3s delay	0.005	0.006	0.965	0.335
addition	6s delay	0.006	0.006	1.063	0.288
addition	9s delay	0.011	0.006	1.992	0.046
division	(Intercept)	0.000	0.007	0.000	1.000
division	3s delay	0.015	0.008	1.785	0.074
division	6s delay	0.006	0.008	0.720	0.472
division	9s delay	0.027	0.008	3.327	0.001
one-two-three	(Intercept)	-0.000	0.006	-0.000	1.000
one-two-three	3s delay	-0.005	0.007	-0.647	0.518
one-two-three	6s delay	-0.009	0.007	-1.204	0.229
one-two-three	9s delay	-0.008	0.007	-1.118	0.264

Table 4.3. Linear regression results for experimental differences in time on task, separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the control condition (no delay), as no differences in time on task are expected. Standardized betas are reported ('estimate').

4.3.3 Substitute responses are primarily slow

Following up the shown decrease in question mark responses, we investigated how children substitute their responses. Naturally, to know exactly which responses are substitutes for question mark responses requires counterfactual information, but the speeds and accuracies of substitute responses can nonetheless be estimated by assessing the changes to the overall response times and accuracies. In Figure 4.5 we show the weekly response time means, averaged across participants, difficulty levels, and response types. We also show how these differ across domains.

A visual inspection of Figure 4.5 reveals structural differences between conditions. In the addition and division domains, the mean response times clearly increase with increased question mark delay. For instance, if in the division domain the question mark button is not delayed, children respond in roughly 7.3 to 7.5 seconds. With a 3 seconds delay the responses slow down

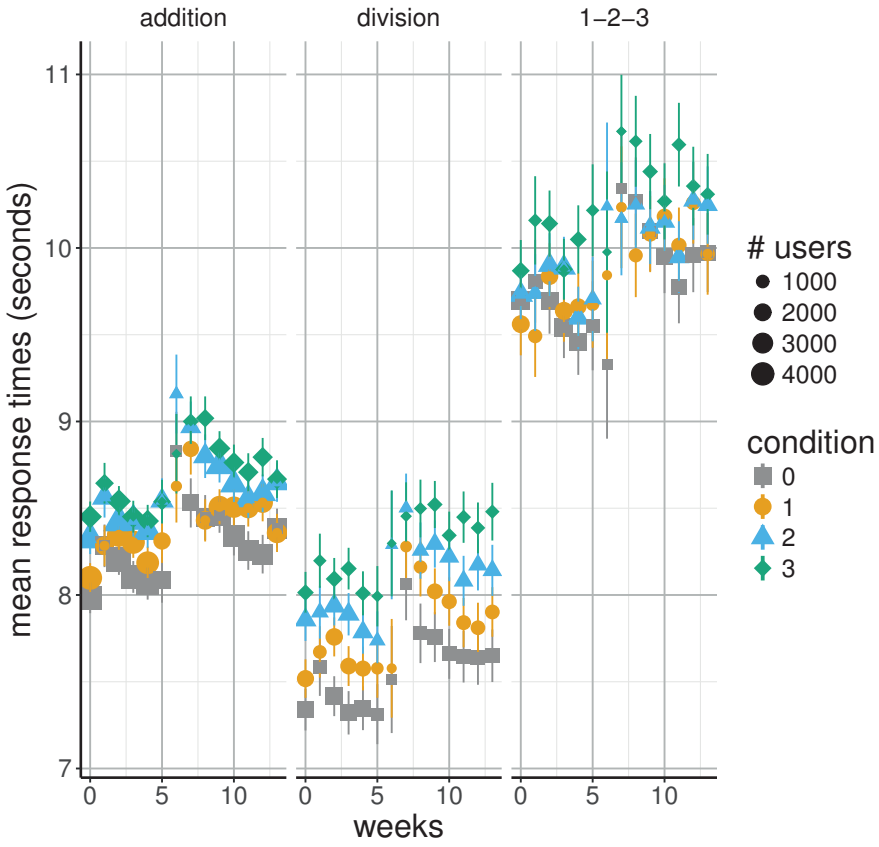


Figure 4.5. Average response times across participants, difficulty levels, and response types, by week. Panels represent domains. Error bars represent 95% confidence intervals.

to roughly 7.5 to 7.8 seconds, and with a full 9 seconds delay children respond in roughly 8 to 8.2 seconds. Interestingly, this increase is decidedly less clear in the one-two-three domain.

We used linear regression analyses with backward difference coding in order to find the effects of the additional increases in question mark delay. The analyses confirm the observed differences. Table 4.4 shows that each additional question mark delay adds up significantly in increasing the response time, except for the 3 and 6 seconds delay in the one-two-three domain, and a decrease in response times for the 9 seconds delay in the addition domain. This finding provides some evidence that the question mark delay is indeed, at least partly, used for toil time, and that fast question marks are not solely substituted by fast guesses.

In the test set, these results were largely confirmed. In all domains, each additional delay contributed to an increase in response times (all $p < .001$, except for the 3 seconds delay in the one-two-three domain, with $p = .040$). Contrary to the results in the practice set, the 9 seconds delay in the one-two-three domain resulted in a small decrease in response times ($p < .001$).

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.022	0.001	-38.339	0.000
addition	3s delay	0.002	0.001	3.534	0.000
addition	6s delay	0.015	0.001	19.164	0.000
addition	9s delay	-0.003	0.001	-4.552	0.000
division	(Intercept)	-0.161	0.001	-206.909	0.000
division	3s delay	0.005	0.001	5.018	0.000
division	6s delay	0.019	0.001	17.051	0.000
division	9s delay	0.021	0.001	21.630	0.000
one-two-three	(Intercept)	0.131	0.001	138.387	0.000
one-two-three	3s delay	-0.001	0.001	-1.011	0.312
one-two-three	6s delay	0.001	0.001	0.565	0.572
one-two-three	9s delay	0.020	0.001	17.021	0.000

Table 4.4. Linear regression results for experimental differences in response times (in seconds), separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the preceding delay, in order to find the effects of the additional increases in question mark delay. Standardized betas are reported ('estimate').

4.3.4 Substitute responses are primarily incorrect

Additionally, we investigated the accuracy of the substitute responses. In Figure 4.6 we show the weekly response accuracy proportions, averaged across participants and difficulty levels. We

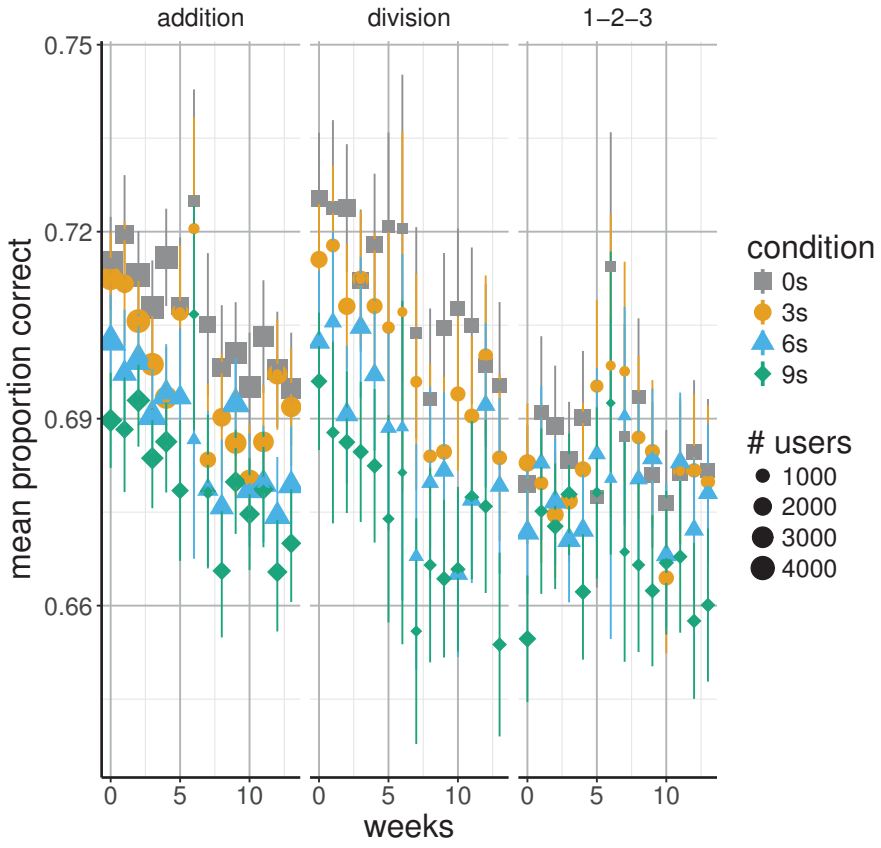


Figure 4.6. Average proportions correct responses *excluding question mark responses*, across participants and difficulty levels, by week. Panels represent domains. Error bars represent 95% confidence intervals.

also show how these differ across domains. We removed all question mark responses, since a change in question mark responses necessarily changes the proportions correct and incorrect responses with respect to all responses, yet we are interested in the mutual proportions between correct and incorrect responses.

A visual inspection of Figure 4.6 seems to reveal a decrease in the proportions correct responses with increased question mark delay, at least for the addition and division domains. For instance, if in the addition domain the question mark is not delayed, children tend to solve roughly 70 to 72% of the problems. With a full 9 seconds delay children solve roughly 67 to 69% of the problems. This decrease is much less clear in the one-two-three domain.

We used linear regression analyses with backward difference coding in order to find the effects of the additional increases in question mark delay. The analyses confirm the observed diffuse effects. Table 4.5 shows that each additional question mark delay adds up significantly in decreasing the proportion of correct responses in the addition domain and the 3 and 6 seconds delay in the division domain, but not in the 9 seconds delay in the division domain and in the one-two-three domain. This finding tentatively points out that although children take more time to formulate a response, the response is often incorrect.

In the test set, the tentativeness of these results is further emphasized. The results were confirmed for the addition domain. However, in the division domain, the 3 seconds delay did not differ significantly from the 0 seconds delay ($p = .717$), whereas the 9 seconds delay did differ significantly from the 6 seconds delay ($p = .011$). And in the one-two-three domain, both the 3 seconds delay and 9 seconds delay differed significantly from respectively the 0 seconds delay ($p = .008$) and 6 seconds delay ($p = .001$).

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.000	0.004	-0.000	1.000
addition	3s delay	-0.011	0.005	-2.372	0.018
addition	6s delay	-0.028	0.005	-5.065	0.000
addition	9s delay	-0.018	0.005	-3.807	0.000
division	(Intercept)	-0.000	0.006	-0.000	1.000
division	3s delay	-0.019	0.007	-2.609	0.009
division	6s delay	-0.026	0.008	-3.168	0.002
division	9s delay	-0.013	0.007	-1.837	0.066
one-two-three	(Intercept)	-0.000	0.005	-0.000	1.000
one-two-three	3s delay	0.001	0.006	0.143	0.887
one-two-three	6s delay	-0.012	0.007	-1.689	0.091
one-two-three	9s delay	-0.009	0.006	-1.531	0.126

Table 4.5. Linear regression results for experimental differences in proportion correct responses *excluding question mark responses*, separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the preceding delay, in order to find the effects of the additional increases in question mark delay. Standardized betas are reported ('estimate').

4.4 Discussion

The question mark delay intents to require children to exert at least some minimum amount of effort, and can thus be seen as the minimum amount of required toil time. The results clearly

demonstrate that the delay indeed ensures a decrease in the use of the question mark. Rather than waiting for the question mark button to appear, children seem to attempt the item more frequently. Also, the toil time does not seem to diminish engagement as the delay does not affect the amount of time children spent on solving items.

Naturally, whether the question mark delay indeed supports active and effortful learning is not that easily concluded. Children may for instance substitute their fast question mark responses for a fast guessing strategy. From a theoretical point of view this is unlikely however. In Math Garden, a fast guessing strategy is risky since especially *fast* incorrect answers are punished with a substantial subtraction of coins, and moreover particularly risky in domains with open-ended question such as the division and one-two-three domains. Moreover, to exclude the possibility of fast guesses, we showed that substitute responses are, although primarily incorrect, also primarily slow.

Looking into the decrease in question mark responses across different domains, one thing to notice is the seemingly smaller decrease in the one-two-three domain as opposed to the addition and division domains. Interestingly, also the substitute responses seem to show a different pattern for this domain. As opposed to the responses in the addition and division domains, the response times do not necessarily increase (except for the 9 seconds delay), and the proportion (in)correct responses is not influenced by the delay.

Multiple explanations can account for this possible difference. First, whereas in the addition and division domains children may resort to memorization strategies, in the one-two-three domain, a complex logical reasoning task, more effortful strategies are already demanded. In this case it is expected for the question mark delay to have less of an effect.

Moreover, since by default the one-two-three domain is only unlocked after frequent practice in the base domains, we might be looking at a highly motivated subset of children that are already less likely to quickly resort to effortless strategies. And lastly, the one-two-three domain has a time limit of 30 rather than 20 seconds. Possibly, since the toil time is thus relatively shorter, it could make the effect less pronounced.

Taking the above together, the strength of the intervention is expressed in its broad applicability. The minimum required toil time ensures an increase in more active and effortful practice, regardless of the complexity of the task, the response mode, or the task length. Moreover, it does not invoke other gaming strategies, such as fast guesses. And finally, it is a so-called soft intervention: it does not prevent children from skipping problems and thus from self-regulating their learning, but nudges children towards a more effortful and more effective learning strategy.

4.4.1 Conclusions

Delaying the option to skip problems in (online) learning can be beneficial. Especially in cases where students are being challenged and an enduring effort is requested, it can be a helpful nudge to exert at the very least some minimum amount of effort. Of course, to safely and conclusively generalize this finding it must be examined on other platforms and in a variety of situations. Also, establishing whether the increased effort results in actual learning gains is an important question that remains open. Nevertheless, three major strengths of the methodology used in the current study are important to highlight.

First, the current paper demonstrates some of the advantages of the A/B testing methodology in the learning domain. Importantly, it allows researchers to evaluate learning interventions on large groups of learners in their natural learning environment. We can use experiments to evaluate causal effects of changes to the system. The readily available data taps into many different aspects of the complex dynamic system of learning, and can thus reveal related patterns such as adverse or beneficial side-effects. Successful interventions can have a large and direct impact: on the basis of this study Math Garden implemented a question mark delay of 25% of a domain's deadline (e.g., 5 seconds for domains with a deadline of 20 seconds), potentially benefiting over 150 000 children. Whereas likewise, adverse interventions can be uncovered upfront rather than blindly implemented.

Second, not only does the large scale of online learning drastically improve the reliability and impact of the interventions, it enables cross-validation of the findings. We exploited this fact in order to further increase the reliability of the study, by using a novel procedure in the spirit of pre-registration. As explained in the Methods section, the findings were only verified on the test set after the editor gave formal approval for publication. This way, we ensured that the research is assessed on the basis of the methods, and we prevented capitalization on chance in both the analysis and review phases.

Finally, findings from online experiments may not only help improve online learning, but the obtained insights may as well validate traditional (offline) interventions and feed back into the various sciences they were drawn from. Generalizability may naturally vary from study to study, but A/B tests can be used for triangulation and usually have great ecological validity. Moreover, it tackles many of the problems encountered in traditional educational experimentation, most importantly the often impracticable double-blind procedure.

A nation that has no music and no fairytales is a tragedy.

Ai Weiwei

5

Tools for teachers

Summary

Virtual learning environments (VLEs), such as Massive Open Online Courses, will maintain an undeniable role in education as a whole. The opportunities of such online learning environments are ample, provided that teachers are equipped with the appropriate tools. Unfortunately, this is not always the case, as customization is often limited to the functionality of the adopted software. One important exception is Learning Tools Interoperability (LTI), which allows teachers to extend a VLE with external software, and thus increase the pedagogical range of their VLE. In this chapter I introduce a software protocol that exploits Qualtrics, popular software for creating and distributing surveys, to extend native VLE functionality with random assignment, additional educational elements, and options for personalizing educational content. The protocol can be used for any online learning environment that supports LTI. I discuss some of the major issues in MOOC research and show how the Qualtrics bridge can contribute to those issues, ultimately providing MOOC teachers with a powerful tool to make evidence-based course improvements.

5.1 Introduction

Opinions vary on the desirability of the rise of Massive Open Online Courses (MOOCs) (e.g., Baggaley, 2013; Ngambi & Bozalek, 2015; Siemens, 2015). However, regardless of whether you are a proponent, opponent, or take a stance somewhere in the middle, in the forthcoming future MOOCs will most likely keep influencing the shape of higher education. In 2016, over 700 universities were involved with MOOCs, offering about 6850 courses, and enrolling an estimated 58 million students (Shah, 2016). These figures, and the steady rise of blended learning, do not seem to herald a decrease in their popularity. With such a massive reach and potential impact, ensuring the quality of these courses is indispensable (Gamage, Fernando, & Perera, 2016). At the same time, with so many different universities involved and such a rich variety of offered courses, ensuring that quality is far from trivial.

One important complicating factor in guaranteeing quality is the inflexibility, or lack of versatility, of virtual learning environments (VLEs) such as MOOCs. A decade ago, Severance, Hardin, and Whyte (2008) already stressed that "...monolithic VLEs are too hard to customize at the individual user level, and evolve far too slowly to meet teaching and learning of users who want their teaching and learning environments to be under their personal control." In an effort to increase the versatility of MOOCs, in this paper I introduce a software protocol that enables three extensions with major importance to both teachers and researchers: questionnaires, adaptive lessons, and experimental comparisons. In the following, I first discuss some issues that obstruct the improvement of MOOCs, and finally discuss how the protocol targets those issues.

5.1.1 Issues with MOOCs

To begin with, Reich (2015) identified various pressing issues that impede leveraging the quality of MOOCs. First, the assessment of learning gains has so far been quite limited, as accurate measures of learning are difficult to obtain, and popular pseudo-measures such as measures of engagement can be deceiving. In order to sensibly evaluate the quality of a MOOC, additional measures that may not be readily provided by the MOOC provider are thus required.

Second and third, a lack of data sharing hinders the comparisons of learning interventions across various domains and circumstances, and although general domain-independent interventions may surely improve the quality of MOOCs, more complex domain-specific interventions might be required to fully optimize those courses. Thus, improving the quality of MOOCs

requires the teachers and researchers involved with a specific MOOC to adapt their teaching and research to the peculiarities of that domain.

In addition to these issues, Savi et al. (2015) stressed the need for prioritizing adaptivity in MOOCs. Indeed, rather than the current one-size-fits-all approach, a strong personalization of MOOCs is required to serve each student's individual needs.

5.1.2 Versatility of MOOCs

Notwithstanding these issues, the most rudimentary one, that of a lack of versatility, is regularly ignored. Namely, although MOOC teachers and researchers must consider the aforementioned issues, MOOC environments not necessarily provide the required functionality to do so. Indeed, the design of effective learning interventions requires a significant flexibility in adapting a MOOC to the needs of the teacher or researcher involved. However, these needs, such as for instance the adaptive assignment of problems or instructions to students, may go beyond the existing functionality of the used platform.

Furthermore, in order to discern whether an intervention improves the desired learning outcomes, one must be able to randomly assign different variants of the MOOC to different students. This approach, better known as A/B testing (Savi et al., 2017), enables a double-blind, iterative and evidence-based optimization of the MOOC in question. Fortunately, the large scale of MOOCs particularly suits experimental comparisons. However unfortunately, again the functionality may not be available, or may be too limited for ones needs. Such lack of versatility thus may severely obstruct the opportunities of teachers and researchers to optimally and decisively improve the quality of their course.

5.2 Learning tools interoperability

As this lack of versatility is the most rudimentary, in this paper I introduce and discuss one potential approach to tackling it. This approach is to use Learning Tools Interoperability (LTI) (Severance, Hanss, & Hardin, 2010), a specification for the standardized integration of learning tools, in combination with Qualtrics, which is widely used software for creating and distributing online surveys. In this section I briefly discuss the LTI specification, and in the next I discuss the potential of Qualtrics to extend MOOCs with questionnaires, adaptive lessons, and experimental comparisons, and the protocol that ultimately bridges MOOCs with Qualtrics.

The LTI specification enables a teacher or researcher to enhance his or her MOOC with student-access to additional external learning tools, while ensuring that small pieces of informa-

tion (such as user id and grade) are exchanged between the MOOC and the external tool. The benefit of this approach is that it extends the native functionality of the MOOC platform with the functionality provided by the external tool, essentially creating a modular learning system.

Many VLEs support LTI, including the widely used Coursera and edX, and it is not uncommon to see the obtained modularity utilized. A common goal in many use cases is to increase the pedagogical range of a MOOC. Examples range from the addition of educational games (Fontenla, Perez, & Caeiro, 2011; Freire, del Blanco, & Fernandez-Manjon, 2014) and Intelligent Tutoring Systems (Aleven et al., 2016; Aleven et al., 2015), to MOOClets (Williams et al., 2014), modules from Carnegie Mellon University's Open Learning Initiative (Koedinger, Kim, Jia, McLaughlin, & Bier, 2015), asynchronous peer instruction (Bhatnagar, Lasry, Desmarais, & Charles, 2016), remote laboratories (Salzmann, Gillet, & Piguet, 2016), and project-based collaborative learning (Cheng, Yu, Park, & Zhu, 2017).

Importantly, the LTI approach provides a significant contribution to tackling the aforementioned issues. First, the external learning tool may establish access to additional measures of learning. Those measures can tap into a different granularity of the learning process, and thus enrich the estimation of learning gains and processes. Second, the additional flexibility of such a tool creates new opportunities for the implementation of more complex domain-specific interventions. This way, teachers and researchers are better served in deploying their (technological) pedagogical content knowledge (Mishra & Koehler, 2006) for the enhancement of their course. And third and fourth, the tool can offer opportunities to further personalize the learning experience, adapt it to the needs of individual learners, and importantly, enable random assignment for experimental comparisons and thus aid in making evidence-based improvements to the course.

5.3 The Qualtrics LTI bridge

5.3.1 Qualtrics

Qualtrics in its turn redeems these promises to a large extent, making it a great candidate for the LTI integration. The software's functionality includes random assignment, and moreover offers an additional layer of flexibility in course design, including functionality that can be of great use in learning environments such as MOOCs. Here, I summarize a few of the most valuable features.

Qualtrics has an intuitive point-and-click and drag-and-drop interface, which enables new

users to quickly master the software, yet gives plenty opportunities for adapting the survey to ones needs. Crucially, it supports the inclusion of typical educational elements, such as instructions, embedded videos, and quizzes. If necessary, it is even possible to include custom-built elements. For instance Barnhoorn, Haasnoot, Bocanegra, and van Steenbergen (2014) created a tool for online reaction time experiments, which could be used to introduce students to behavioral research.

As the software is optimized for surveys, it offers a large flexibility in question design. It not only supports a rich variety of question types, such as rank order questions, constant sum questions, and timed questions, it moreover supports randomizing the order of both questions and answers. Also, the supported Likert scales and visual analogue scales can greatly benefit MOOC research that requires the use of questionnaires.

Interestingly, the software also enables the random assignment of course components. This feature enables experimental comparisons and is key in providing an evidence-based method to iteratively optimize the course and increase its effectiveness. To further facilitate this, Qualtrics provides many opportunities for recording learner activity, creating an opportunity for additional measurements of learning gains and learning processes.

Finally and importantly, the assignment of course components does not need to be random. Qualtrics also supports carrying forward the answer of one question to a subsequent question, and it can present questions, instructions, or complete blocks of educational material conditional on for instance a previous answer or an external user id (such as the id used by the MOOC provider)[1]. These options enable a straightforward rudimentary form of adaptivity and can thus benefit personalization of the course material.

Qualtrics also has two major limitations. First, a clear restriction of Qualtrics is that it is proprietary software that requires a subscription fee. Although a free version is available, this does not offer the capacity required for the typical large number of learners in MOOCs. That said, being one of the more popular software packages for surveys, many universities already provide subscriptions to their staff.

Second, Qualtrics currently does not provide native LTI support. Although without such support MOOC teachers can still send students to the additional material in Qualtrics, it is not possible to match the responses from both environments. Such matching of responses often is a requirement for both teaching and research purposes. Teachers may want grades from Qualtrics quizzes to be returned to the MOOC environment, while researchers may want to combine information from both sources in their analyses. Therefore, in order to circumvent the lacking

[1]Qualtrics calls this functionality *Branch Logic* and *Display Logic*.

LTI support, an LTI bridge was developed.

5.3.2 Qualtrics LTI bridge

In order to circumvent Qualtrics' lacking LTI support, the Qualtrics LTI bridge appends the URL that is used to send the student from the MOOC environment to Qualtrics with a query string containing the relevant information. Such passing of information is necessary for identifying the MOOC learner in the Qualtrics environment, and matching it to his or her responses on the additional course elements.

The software protocol that bridges the MOOC environment and Qualtrics is available on GitHub[2] (Poesse & Wiles, 2016). The protocol works with any MOOC platform or virtual learning environment that supports LTI. Importantly, the protocol includes detailed instructions for teachers and researchers on how to implement the protocol and adapt it to ones personal needs. The protocol is open-source, and thus free for everyone to use and build upon.

From the viewpoint of a learner, the LTI bridge integrates Qualtrics by creating an URL that directs the learner from the MOOC environment to the Qualtrics environment. The MOOC teacher naturally decides on the location and timing of the link. As the URL is appended with the user id that is used by the MOOC provider, the learner can be identified from within the Qualtrics environment. Once the learner enters the Qualtrics environment he or she can start studying the material that was created by the teacher or researcher. This can be a rich variety of materials, including the examples given in the previous section. In case of experimental comparisons, the user can be randomly assigned to one of multiple conditions. Once the learner has finished studying the educational material, the responses are recorded in Qualtrics and the learner can return to the MOOC environment. The response data can be easily exported by the teacher or researcher involved. And interestingly, Qualtrics allows for customization of the page design, such that the layout of the Qualtrics environment can be matched to that of the MOOC environment.

Finally, it is possible to grade assignments and quizzes in Qualtrics, and return the grades to the MOOC environment. This is important in case one wants the performance in Qualtrics to

[2]The protocol is available at https://github.com/renspoesse/qualtrics_lti_bridge. The protocol was adapted from an earlier bridge developed by Simon Wiles at Stanford University, which was generously shared at https://github.com/cognitivesciencelearning/qualtrics_lti_bridge. Instructions for using and adapting the protocol are available at https://github.com/renspoesse/qualtrics_lti_bridge/wiki, where you can also contribute to the instructions or provide usage examples. An issue or feature request can be filed at https://github.com/renspoesse/qualtrics_lti_bridge/issues.

count towards the achievement in the MOOC. Importantly however, as the LTI workaround fails to take advantage of the full LTI specification, it does not support a fool-proof two-way interaction between the MOOC environment and Qualtrics. Although for instance Coursera accepts incoming grades, and Qualtrics provides automatic grading, the workaround cannot guarantee an accurate and secure transfer of those grades.

The instructions on GitHub are updated as soon as changes are made to the protocol. As the GitHub environment enables anyone to contribute to the protocol, researchers are encouraged to implement and test the protocol, contribute to the instructions, provide usage examples, or adapt the protocol. Feature requests or bugs can be filed directly at the protocol-repository on GitHub, or be sent to the author.

5.4 Discussion

The limited functionality of a virtual learning environment, such as a MOOC, severely restricts the pedagogical range of the teacher. With VLEs here to stay, it is important to support teachers with sufficient opportunities for customization. The software protocol introduced in this paper aims to do exactly that. It uses the LTI standard to enable the integration of VLEs and Qualtrics, creating additional pedagogical opportunities for teachers.

The Qualtrics LTI bridge extends the pedagogical range of MOOCs, and enables teacher-driven experimental comparisons on platforms that do not, or insufficiently, facilitate such comparisons. The added educational elements enrich and diversify not only the learning experience, benefiting the possibilities for creating more domain-specific learning interventions, but also the learning measurements, enabling a more fine-grained evaluation of learning gains and processes. In combination with experimental comparisons teachers can iteratively and decisively improve their course.

Moreover, the Qualtrics bridge is not confined to MOOCs. Since the LTI protocol ensures a connection between any two online learning environments that support the protocol, the bridge can also be used by teachers and researchers that use a different type of learning environment. Those other environments might also benefit from additional pedagogical flexibility and learning measures, or neither provide the tools for teacher-driven experimental comparisons, thus for those teachers and researchers the protocol can also come in helpful (e.g., Henrick, 2012).

The power of experimental comparisons in MOOCs, and other online learning environments, must not be underestimated. Reliable experiments in education are challenging, since

the conditions of randomization and double-blinding are difficult to satisfy. Yet in online learning those conditions are much more easily satisfied, and its large scale provides the power to reliably discern the effects of different interventions (Kizilcec & Brooks, 2017; Savi et al., 2017). Experimental comparisons in MOOCs thus create a unprecedented opportunity for improving their quality, ultimately ensuring a positive impact of MOOCs on higher education.

Finally, the significance of MOOC teachers that can themselves deploy experimental comparisons can neither be overemphasized. Moving away from the aforementioned domain-independent research on learning interventions requires teachers to have sufficient educational and pedagogical flexibility, including ways to thoroughly and reliably evaluate their course. The Qualtrics bridge provides MOOC teachers with the desired experimentation and measurement functionality, and thus potentially helps levitate the quality of MOOCs.

Things are beautiful or ugly only in time and space. The new man's vision being liberated from these two factors, all is unified in one unique beauty.

Piet Mondriaan

6

Misconceptions unmasked

Summary

In learning, errors are both ubiquitous and inevitable. It is widely understood that these errors may provide a clue about a person's misconceptions. In this article we propose and investigate a model that aims to identify misconceptions from observed errors. We apply the method to single digit multiplication; a domain that is very suitable for the method, is well-studied, and allowed us to analyze over 25 000 error responses from 335 actual learners. The model, derived from the Ising model popular in physics, makes use of a bigraph that links possible errors to possible misconceptions. The error responses were taken from Math Garden, a computerized adaptive practice environment for arithmetic that is widely used in The Netherlands. The results show that the model outperforms a random selection from the observed errors' possible causes, and correctly predicts the possible cause of a person's subsequent error up to over 75% of the time. Finally, we discuss the model, the findings, and the implications.

6.1 Introduction

Students' error responses provide a unique window into the mind. Or—less poetically—students' error responses may reflect their applied strategies, or cognitive processes,

when solving problems. This fundamental understanding has spawned decades of research, from classifications of errors (e.g., Ben-Zeev, 1998; Straatemeier, 2014), and cognitive models aimed at explaining errors (e.g., Buwalda, Borst, van der Maas, & Taatgen, 2016), to the identification of misconceptions from observed errors (e.g., Taraghi et al., 2015; Taraghi, Saranti, Legenstein, & Ebner, 2016). In this contribution to the field of errors in learning, we investigate a method for the latter—a new approach to detecting the latent causes of an individual student's manifest errors.

6.1.1 Challenges in error diagnostics

Errors come in many shapes. One straightforward classification is the separation into systematic and unsystematic errors. An unsystematic error is usually termed a mistake or slip (D. A. Norman, 1981)—"the error that occurs when a person does an action that is not intended." This type of error may originate from sloppiness, carelessness, or inattentiveness. On the systematic end of the spectrum are so-called misconceptions, or rational errors (Ben-Zeev, 1995)—that is "students ...correctly following incorrect rules, rather than incorrectly following correct ones." Rational errors are sometimes described as bugs—incorrect perturbations of correct procedures (J. S. Brown & Burton, 1978)—and Ben-Zeev (1998) discusses various hypothesized origins of such errors, among which an incorrect induction of examples (VanLehn, 1986). Finally, a decidedly different category of systematic errors stems from heuristics (e.g., Reber, Brun, & Mitterndorfer, 2008) and biases (e.g., Shaki & Fischer, 2017), which became of increased interest more recently.

The ultimate promise of being able to diagnose error responses, is the guidance it may provide in adapting education to students' individual needs. However, inferring the cause of an error—and ensuring that the student can benefit—poses serious challenges. The first challenge is to map all possible causes of an error. Straatemeier (2014) gives an elegant example from arithmetic: the case of $9 \times 9 = 18$. It is easily seen that this individual might have used the wrong operator, adding both operands rather than multiplying them. But, for all we know, this person might have actually performed the correct calculation, but reversed the decade and the unit in the answer. Such a decade-unit inversion exists, for instance in the Dutch language (van der Ven et al., 2016), where 81 is pronounced 'one-and-eighty', and which might contribute to the error. A third option is the operator-relevant error, where 18 is the correct answer to the incorrect problem: 2×9 or 9×2. And finally, in addition to these systematic misconceptions, it simply could have been an unsystematic slip.

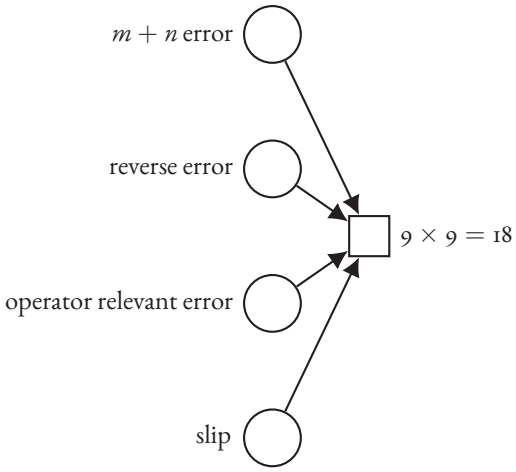


Figure 6.1. The bipartite graph for $9 \times 9 = 18$. The left column with circles depicts the latent causes and the right column with the square depicts the manifest effect. The arrows reflect the theory on what are the causes of which effect.

Graphically, the first challenge can be visualized by a so-called bipartite graph—or bigraph—which links causes to effects. The graph for the above problem is shown in Figure 6.1. This graph, which we will call the theoretical model, links the selected errors to all of their possible causes (insofar that those causes are known). Figure 6.1 thus serves as a very minimal example, but nonetheless conveys the structure of a full theoretical model, which may include all single digit multiplication errors and their known possible causes.

Additionally, the graph beautifully expresses the need to adapt education to the individual. Although the theoretical model nicely summarizes the many different causes of observed errors across individuals, the actual causes may naturally differ from individual to individual, and from time to time. Whereas the one might retrieve the correct answer to an incorrect problem from memory, the other might have difficulty with transcoding. In many situations, such differences require different interventions.

Importantly, the observed error $9 \times 9 = 18$ does not reveal the actual cause for one particular individual. Therefore, the second challenge, and the topic of this article, is inferring the latent causes that drive the manifest errors of an individual. Again, this challenge can be displayed as a bipartite graph, shown in Figure 6.2. However, this time the error is represented by a black square, reflecting the fact that the error is observed, and the possible causes are represented by dashed circles, reflecting the fact that the actual cause for this individual is unknown.

Finally, the observed error responses pose yet another challenge. Error responses are not

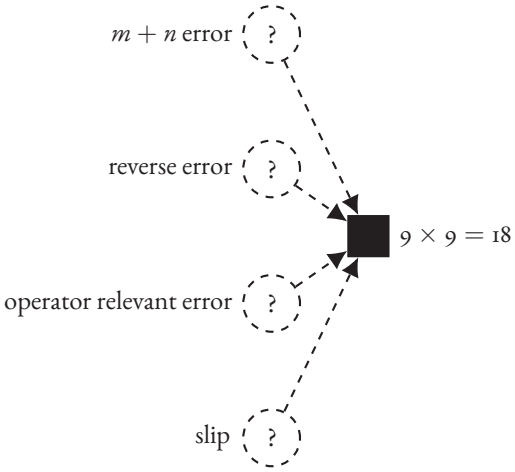


Figure 6.2. The problem of identifying the cause of an error for a particular individual, represented in the bipartite graph for $9 \times 9 = 18$. The left column with dashed circles depicts the possible latent causes for a particular individual and the filled square in the right column depicts the observed effect.

ubiquitous, and their causes may change over time. For the promise of adapting education to the individual, the inference of possible causes thus relies on limited data. This is not only a problem for inference, but also for evaluating the accuracy of a model. In this study, the latter problem is solved by using a response-intensive longitudinal data set, which allows us to investigate the robustness of the former.

6.1.2 Cognitive diagnostic models

One approach to the challenges of diagnosing errors is the use of Cognitive Diagnostic Models (CDMs). CDMs are latent class models developed to identify the presence or absence of specific skills that are required to correctly answer a set of items. This is in contrast to traditional item response theory models, which measure ability on a unidimensional scale. Instead of measuring a single unidimensional rating for each person, CDMs maintain a profile $\alpha = (\alpha_1, \ldots, \alpha_K)$ where $\alpha_k = 1$ if the kth skill has been mastered and $\alpha_k = 0$ if the kth skill is not mastered for $k = 1, \ldots, K$.

A key construct in CDMs is the Q-matrix which specifies which skills are required by which items (Tatsuoka, 1983). Various different CDMs exist, and differ in how they relate the latent class profile and the Q-matrix to the observed responses. Examples of CDMs include the DINA model (Haertel, 1989) and the DINO model (Templin & Henson, 2006).

In recent literature, CDMs have been extended to diagnose misconceptions as well as ability. The first extension was a joint model which incorporated concepts from item response theory (IRT), as well as CDMs for modeling unidimensional ability and diagnosing misconceptions (Bradshaw & Templin, 2013). An extension of the DINO model was then developed by Kuo, Chen, Yang, and Mok (2016). Finally, a model which diagnoses the presence of skills as well as misconceptions was developed (Kuo, Chen, & de la Torre, 2017).

The limitation of CDMs for diagnosing misconceptions is in the use of the Q-matrix to relate misconceptions to items. All three of the models that have used CDMs to diagnose misconceptions map the misconceptions to specific items. This limits the kind of misconceptions that can be analyzed. From the previously discussed bigraph, it is seen that misconceptions are often not related to a particular *item*, but rather to a specific *error*. For example, consider the misconceptions; operand related error, reverse error, and addition error. All three of these could be made on the item 4×5. On the other hand, if we know the specific error that was made was $4 \times 5 = 25$ we know an operand related error occurred.

6.1.3 Bayesian misconception tracing

In this manuscript, we introduce and evaluate another approach that can be used to diagnose an individual's misconception: Bayesian Misconception Tracing (BMT). This method is simple and intuitive, allows easy implementation in online learning systems, and exploits the known theoretical relations between misconceptions and errors as summarized in a bigraph. In this simple but nontrivial method, discussed in detail in the Methods section, we utilize the proportion of theoretical relations between a misconception and a set of observed errors as an indication of a misconception's probability.

In demonstrating the method, we apply it to the problem of single digit multiplication. Although in principle it works in any domain, as long as clear errors can be identified and the possible causes can be mapped—that is, a bigraph like in Figure 6.1 can be created—not all domains lend themselves that easily. Multiplication serves as a great illustration for a couple of reasons. First, the very procedural nature of multiplication allows for the identification of clear bugs, which in turn has motivated scholars to identify a multitude of causes. Second, adding to this convenience, software algorithms can easily and automatically detect multiplication errors. And finally, many online learning systems exist that provide multiplication education, and could thus readily benefit from the method.

To illustrate and evaluate the method, we use data from Math Garden, a computer adaptive

practice environment primarily aimed at Dutch primary school children. Thus far, their single digit multiplication domain has generated over 25 million responses from over 170 thousand different primary school children, spanning 3 years, and with over 5 million errors made (May, 2018). This learning environment originates from an innovation that enables the adaptive assignment of problems to children, based on on-the-fly updated general measures of ability and difficulty, by means of an adapted Elo rating system (Brinkhuis et al., 2018; Klinkenberg et al., 2011).

Finally, in our example the error taxonomy created by Straatemeier (2014) serves as the point of departure for the theoretical bigraph. Conveniently, their taxonomy is based on the rich Math Garden data, which enabled them to identify a great many causes. We adapted their theoretical model to suit single digit multiplication. The adapted model, with definitions for the used set of multiplication misconceptions, is too provided in the Methods section.

6.2 Methods

In this section, we lay out the general framework of the approach. We provide the employed error classification, describe the data, introduce the model that we use to induce the cause of observed errors, and discuss the procedure we used to evaluate the model. Note that we interchange the use of the terms misconceptions, error categories, and causes, as well as students, children, and users.

6.2.1 Error classification

As discussed in the introduction, theory links the errors to their possible causes. This theory is captured in a bigraph, of which Figure 6.1 gives an illustrative example. For this bigraph, we used the classification scheme proposed by Straatemeier (2014), and adapted it for single digit multiplication. The classification scheme consists of a great variety of error types. Importantly, we consider all categories for which Straatemeier identified *systematic* responses. Table 6.1 provides the error categories and definitions.

cause	description
operator.relevant	answer to a different single digit multiplication problem, from the tables 2 to 9
operand.related	answer to a problem with one matching operand, and one operand that is 1 or 2 units smaller/larger (except when operand becomes zero or negative)
double.half	double or half the correct answer
same.decade	answer has the correct decade
miss.1	correct answer plus/minus 1
miss.10	correct answer plus/minus 10
miss.100	correct answer plus/minus 100
miss.power	correct answer with the decimal point misplaced by 1 to 5 positions (i.e., correct answer multiplied/divided by 10 to the power of 1 to 5)
m.div.n	the first operand divided by the second operand
m.minus.n	the first operand minus the second operand
m.plus.n	the first operand plus the second operand
typo	correct answer with the repetition or omission of a digit (omission only when correct answer has 2 digits or more)
reverse	the digits of the correct answer are reversed (only for problems with a solution that consists of 2 digits)
zero	0 is (incorrectly) provided as the answer

Table 6.1. The error categories considered in this study. Adapted from Straatemeier (2014) for single digit multiplication.

6.2.2 Data

Math Garden

Math Garden hosts a variety of domains, primarily related to arithmetic, that can be practiced in isolation. One such domain is the *multiplication table* domain—which provides the data for this study—and the 22 other domains include problems ranging from word problems to logical reasoning tasks. Importantly, the previously mentioned adaptive algorithm matches students to problems. Students can be viewed as competing with the problems in a domain, and the outcome—both in terms of speed and accuracy (Maris & van der Maas, 2012)—feeds into the adapted Elo algorithm, to continuously update student ability estimates and problem difficulty estimates.

In the multiplication table domain, users practice the multiplication tables of one to ten. For 5, 10, 15, or 20 seconds, a problem is presented, and during this time the user can provide a solution by means of a visualized numeric keypad. For each second that there is time left to solve

the problem, the user receives a virtual coin for a correct response, and loses a virtual coin for an incorrect response. This scoring rule represents the implemented speed-accuracy trade-off, and is visualized by coins disappearing from the screen. Too difficult problems can be skipped by using a question mark button, which is without consequence.

Importantly, Math Garden is primarily used in natural learning settings. Both schools and families can buy subscriptions, and the system is used in and outside of the formal school setting. This property, along with the sheer amount of problems solved, creates a unique data set with very diverse error responses (previously analyzed by Straatemeier, 2014).

Selection criteria

These data properties require a careful selection procedure. We applied the method to single digit multiplication problems, thus we removed all responses to the multiplication table of ten. As the method only takes error responses into account, we disregarded correct responses. Also, we removed question mark responses, non-responses (time-outs), and responses within 1 second. We restricted selection to a three months period (January to March, 2017), users that allowed scientific research based on their responses, and users in Dutch grades 3 to 8 (comparable to grades 1 to 6 in the US, and approximately age 6 to 12). We removed error responses that could not have been caused by any of the misconceptions in the used classification. Also, we only considered users with no more than 50% error responses. Depending on the analyses, we selected users with a minimum of 80 responses in total and a minimum of 40 error responses, or users with a minimum of 40 responses in total and a minimum of 20 error responses. In the Model Evaluation section we explain these differences.

Error distribution

The selected data, for the stricter data selection with a minimum of 80 responses in total and a minimum of 40 error responses, is summarized below. Table 6.2 gives the error response frequencies by school grade. To put these numbers in perspective, the table also gives the frequencies of correct responses in the selected data. The table shows an increase in the amount of responses across school grades, both correct and incorrect. Then, Figure 6.3 shows the numbers of students by their error response frequencies, with a total of 335 students in the selected data. One should note that the distribution is highly skewed, with a few students providing a large number of error responses.

response accuracy	grade	# responses
0	3	1137
0	4	2392
0	5	4535
0	6	4762
0	7	5997
0	8	7134
1	3	7091
1	4	14638
1	5	30573
1	6	45604
1	7	55624
1	8	90306

Table 6.2. Total response frequencies by grade of the student and accuracy of the response, for the stricter data selection with a minimum of 80 responses in total and a minimum of 40 error responses.

6.2.3 Model

We propose a model to induce the latent misconceptions that cause the selected systematic error responses. To begin with, an intuitive understanding of the primary mechanism of the model is easily obtained. Given a subset of observed errors, one simple method to calculate the probability of a cause, is to determine the number of errors associated with the cause of interest, proportional to the total number of associations between causes and errors. Figure 6.4 exemplifies this method, showing that it comprises of no more than computing the proportion of edges for each of the causes in the observed bigraph. The model we introduce hereafter, a type of Ising model, shares the idea that the relative number of associations a cause has with the observed errors, provides a proxy for the plausibility of the cause.

Let $G = \{C, E, \mathcal{A}\}$ be a bipartite graph where C is the set of nodes related to causes, E is the set of nodes related to errors, and $\mathcal{A}$ is the set of (weighted) edges relating the causes and errors. We can represent $\mathcal{A}$ in matrix form such that the i, j element, a_{ij}, is the weighted edge from cause i to error j for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. A node $e \in E$ can either be $e = 1$ or $e = 0$ to indicate whether the error has been observed or not observed. A node $c \in C$ can likewise be either $c = 1$ or $c = 0$ to indicate the presence or absence of the respective cause.

We model the joint distribution of causes and errors as a type of Ising model (Ising, 1925). The Ising model is a simple model for jointly modeling the distribution of a set of dichotomous variables. It was originally formulated to model ferromagnetism in physics. The standard Ising model is for variables that are coded with ± 1 whereas we use 1/0 encoded variables. The joint

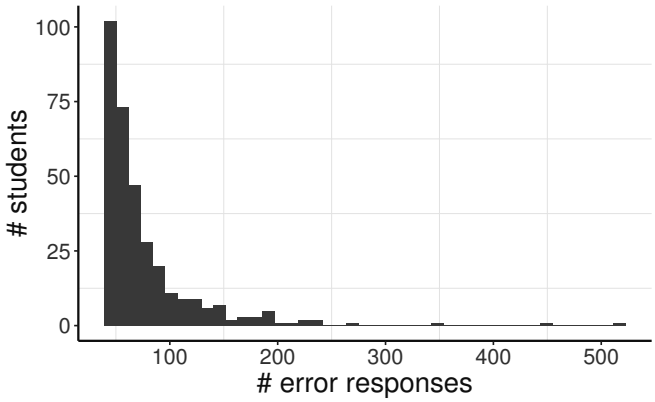


Figure 6.3. Distribution of students across the number of error responses, for the stricter data selection with a minimum of 80 responses in total and a minimum of 40 error responses.

distribution can be expressed as

$$p(\mathbf{x} = (\mathbf{c}, \mathbf{e})|\mu, \beta) = \frac{1}{Z} \exp(\beta \mathbf{x}^\intercal \mu + \beta \mathbf{x}^\intercal \Sigma \mathbf{x}) \tag{6.1}$$

where $\mathbf{x}^\intercal = \begin{pmatrix} \mathbf{c}^\intercal & \mathbf{e}^\intercal \end{pmatrix}$, μ are the parameters associated with causes and errors (the external magnetic field in the Ising literature), Z is the normalizing factor, and Σ is the interaction effect matrix where the i, j element of Σ, σ_{ij}, corresponds to the interaction strength between x_i and x_j.

Because G is bipartite, Σ is of the form

$$\Sigma = \begin{pmatrix} \mathbf{0} & A \\ A^\intercal & \mathbf{0} \end{pmatrix}$$

so we have

$$p(\mathbf{x} = (\mathbf{c}, \mathbf{e})|\mu, \beta) = \frac{1}{Z} \exp\left(\beta \sum_{i=1}^{n} \mu_i c_i + \beta \sum_{j=1}^{m} \mu_{n+j} e_j + 2\beta \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} c_i e_j \right) \tag{6.2}$$

Provided that we are interested in the probability that one misconception is present, given

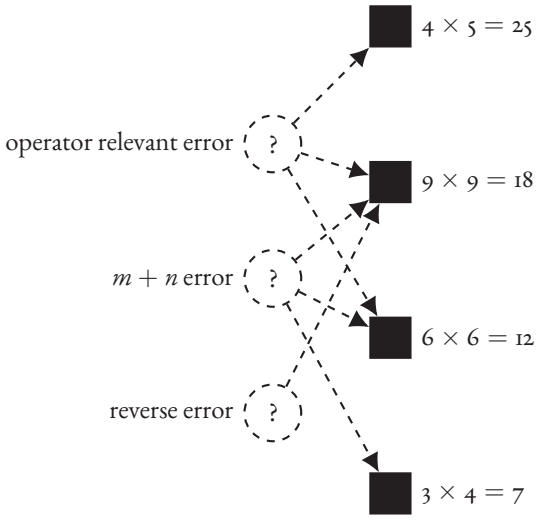


Figure 6.4. An example bigraph with 3 causes and 4 errors.

that we observed all considered errors in the data, we define

$$p_i = p(c_i = 1 | \textstyle\sum_{i=1}^{n} c_i = 1, \mathbf{e} = \mathbf{1}, \mu, \beta) \tag{6.3}$$

$$= \frac{p(c_i = 1, \mathbf{c}_{-i} = \mathbf{0}, \mathbf{e} = \mathbf{1} | \mu, \beta)}{\sum_{k=1}^{n} p(c_k = 1, \mathbf{c}_{-k} = \mathbf{0}, \mathbf{e} = \mathbf{1} | \mu, \beta)} \tag{6.4}$$

$$= \frac{\exp\left(\beta\mu_i + \beta \sum_{j=1}^{m} \mu_{n+j} + 2\beta \sum_{j=1}^{m} a_{ij}\right)}{\sum_{k=1}^{n} \exp\left(\beta\mu_k + \beta \sum_{j=1}^{m} \mu_{n+j} + 2\beta \sum_{j=1}^{m} a_{kj}\right)} \tag{6.5}$$

$$= \frac{\exp\left(\beta\mu_i + 2\beta \sum_{j=1}^{m} a_{ij}\right)}{\sum_{k=1}^{n} \exp\left(\beta\mu_k + 2\beta \sum_{j=1}^{m} a_{kj}\right)} \tag{6.6}$$

Figure 6.4 provides an example. With $\mu_i = 0$ for $i = 1, \ldots, 7$, and $\beta = 1/2$, we have

$$p_1 = \frac{\exp(3)}{\exp(3) + \exp(3) + \exp(1)} \qquad p_2 = \frac{\exp(3)}{\exp(3) + \exp(3) + \exp(1)}$$

$$p_3 = \frac{\exp(1)}{\exp(3) + \exp(3) + \exp(1)}$$

In addition to being intuitive, the model has several benefits. For one, it neatly accounts for the fact that a slip could have caused the error, which we show in Appendix A. Also, the model can easily be generalized to the probability given that any number of misconceptions are present.

For example, the probability that two misconceptions are present is defined as

$$p_{i_1,i_2} = p(\text{misconception } i_1 \text{ and } i_2 \text{ are present} | \text{exactly two causes are present}) \tag{6.7}$$

$$= p(c_{i_1} = 1, c_{i_2} = 1 | \textstyle\sum_{i=1}^{n} c_i = 2, \mathbf{e} = \mathbf{1}, \mu, \beta) \tag{6.8}$$

$$= \frac{\exp(\mu_{i_1} + \mu_{i_2} + \sum_{j=1}^{m} a_{i_1,j} + \sum_{j=1}^{m} a_{i_2,j})}{\sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \exp(\mu_{k_1} + \mu_{k_2} + \sum_{j=1}^{m} a_{k_1,j} + \sum_{j=1}^{m} a_{k_2,j})} \tag{6.9}$$

Importantly, assigning the obtained probabilities to the considered misconceptions is in accordance with Luce's choice axiom (Luce, 2005). This axiom states that the probability of selecting one item over another from a pool, should not be affected by which items are present in the pool. Such probabilities are said to have independence from irrelevant alternatives.

6.2.4 Model evaluation

In this article, we aim to evaluate this proposed model. For this purpose, we took two different approaches. First, we assessed the model's prediction accuracy, by predicting users' errors from previous error responses. For each individual in the data, we calculated the model's expected probabilities for each of the causes. We calculated these probabilities from a moving window consisting of a predetermined number of error responses. Using these probabilities, we predicted the observed error directly following the considered window. We only considered the misconceptions that—according to the bigraph—could cause an error in considered window.

We compared three methods of selecting a predicted cause from the obtained probabilities: the cause with the highest predicted probability, a sampled cause given the causes' probability distribution, and a random cause. Each time we determined whether the predicted cause could have caused the observed error. Based on these windowed predictions, we then calculated the proportion of correct predictions for each user in the data. We evaluated five different window sizes, consisting of 1 to 30 errors. To make sure to have enough error responses for each user and each window, we used the stricter parameters for data selection, with a minimum of 80 responses in total and a minimum of 40 error responses.

This approach requires two important remarks. First, predictions we considered incorrect, are not necessarily incorrect. An incorrect prediction tells us that the predicted misconception could not have caused the newly observed error. Indeed, the student might still suffer from this particular misconception, but could have used a different incorrect strategy for this particular item (such as an incorrect memorization). Second, predictions we considered correct, are not necessarily correct. Although we know that the observed error can in principle be caused by the

predicted cause, it might as well have been caused by a different cause—either a cause that we too identified, or a cause that is not in the model. These remarks should be taken into account when interpreting the evaluations of the model.

Second, we evaluated the model's capacity to capture developmental changes in error responses. Here, we computed the probabilities for each of the misconceptions from all error responses of each user. To increase the number of users, and thus power, we used the less strict parameters for data selection with a minimum of 40 responses in total and a minimum of 20 error responses. To investigate developmental patterns, we analyzed differences in expected probabilities across school grades.

6.3 Results

We first evaluate the model by means of its prediction accuracy. In all analyses, we considered the most parsimonious model, with $\mu = 0$, $\beta = .5$, and $a_{ij} = 1$ for causes that can produce a particular misconception ($a_{ij} = 0$ for causes that cannot produce a particular misconception). The approach is best described on the basis of Figure 6.5. The figure shows the predictions for successive errors of twenty students, based on the fifteen preceding errors of the student. We picked the causes with the highest predicted probabilities. Downward nudges represent incorrect predictions, whereas upward nudges represent correct predictions.

Already some interesting patterns can be observed in this specific example. Standing out is the dominance of the operator relevant error—both in incorrect and correct predictions—where the student's response is the correct answer to a different problem in single digit multiplication (a dominant error by definition). Other interesting patterns can be observed by focusing on specific individuals. For instance, in case of the fifth student from the bottom, the model is relatively uncertain, but primarily correct. Or in case of the first student from the top, the model is relatively certain, and primarily correct. On average, the model correctly predicts roughly four out of five errors.

In the remainder of this section, we investigate the model systematically and in greater detail. First, we compare different methods of selecting a predicted cause from the model's probability estimates, vary the size of the window, and compare predictions of one and two causes. For each of these scenarios we determine how it affects the model's predictive power. Finally, we briefly look into developmental patterns of the predicted misconceptions across school grades.

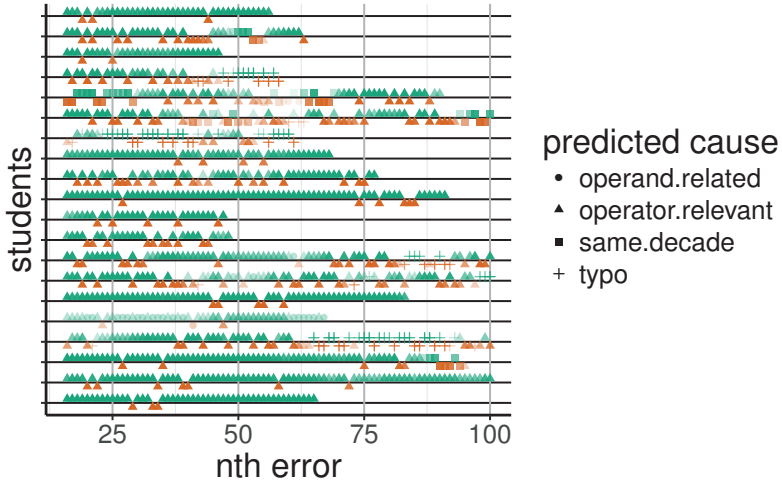


Figure 6.5. Model predictions for successive errors (*x*-axis) of 20 students (*y*-axis). Predictions are based on a moving window of 15 errors, hence the start of the *x*-axis at the 16th error. The shape of the point represents the predicted error cause. A downward nudge (red) represent an incorrect prediction, whereas an upward nudge (green) represents a correct prediction. The colour density represents the model's certainty (predicted probability).

6.3.1 Prediction accuracy

We determined the model's prediction accuracy for the three different methods of selecting the predicted cause, different window sizes (the number of observed errors provided to the model, either 1, 3, 7, 15, or 30), and different numbers of predicted causes (1 or 2 causes[1]). Only the causes that are linked to the observed errors in the error window were considered, to allow a reasonable comparison with the random method of selecting a predicted cause.

Figure 6.6 shows box plots for the average prediction accuracy for each student, for each of these scenarios. It shows the proportions of correct predictions for taking the highest predicted probability ('max'), a sampled cause based on the distribution of predicted probabilities ('sample'), and a random cause ('random'). We look into the predictive power of a single cause (one), or two causes (two).

First, it is easily observed that the prediction accuracies vary from user to user. For most scenarios, the highest proportion of correct predictions is about .25 to .5 higher than the lowest.

[1]Errors or sets of errors, with bigraphs that contained a single cause, were not considered in the scenarios where more than one cause was requested.

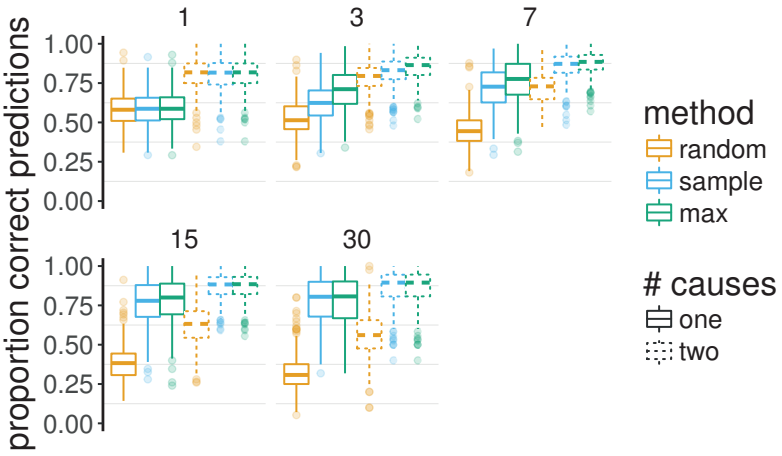


Figure 6.6. Box plots for the proportions of correct predictions (calculated for each student). The panels show these proportions for different window sizes (1, 3, 7, 15, or 30 errors), the colors for different methods (random prediction, highest prediction, or sampled prediction), and the box border types for different numbers of predictions per student (one or two).

This difference seems higher for single predictions than double predictions.

Then, the predictions for the window size of a single error are noteworthy. The selection methods seem to perform equally, and prediction accuracies already are considerable. These two observations might be explained by the dominance of the operator relevant error, and the fact that the bigraph of a single error will contain very limited causes.

Importantly, the remaining panels in the figure suggest that predictions generally get better with increased window size, except for the random selection method. Indeed, with larger window size it is expected that the bigraph contains more possible causes, for which the random selection method has no preference.

The improved prediction accuracy of the Bayesian Misconception Tracing method (by either sampling a prediction or picking the most probable) over randomly selecting a cause from the bigraph, is confirmed by linear regression analyses. Table 6.3 shows the results of these analyses, separately for one and two predictions, and the various window sizes.

Then, the decrease in performance of the random selection method with increasing window size is also confirmed using linear regression analyses. Table 6.4 gives the results, and shows that although the method—with either sampling a prediction or picking the highest—benefits of a larger error window, the errors additional to a window of seven to fifteen errors do not seem to

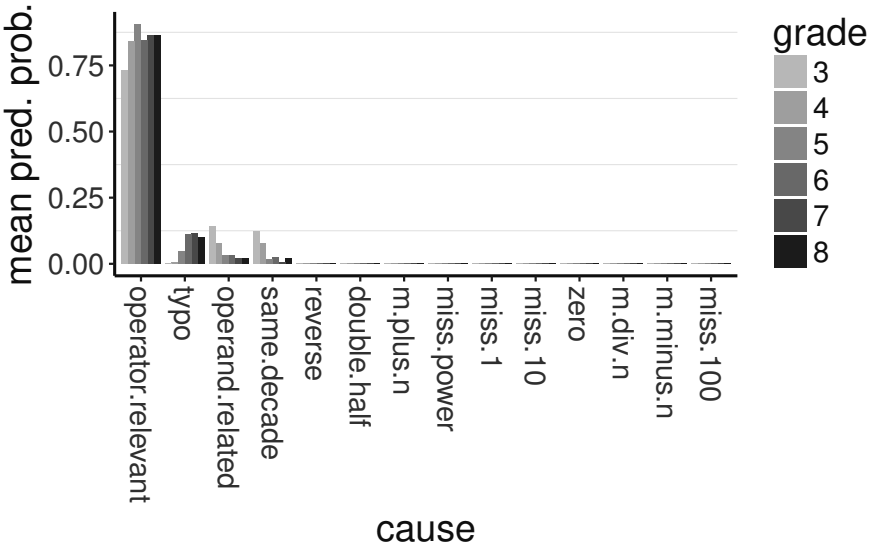


Figure 6.7. Predicted probabilities averaged across users, grouped by grade and cause. Only causes that are related to the errors in the data set are considered. Within each misconception in the figure, the bars read from left to right, with the grades going from 3 to 8.

increase prediction accuracy.

6.3.2 Theoretical validation

Our final objective is a substantive verification of the method. Notably, the rich longitudinal Math Garden data is the preeminent data for studying developmental patterns. We examined the model's predicted probabilities, averaged across users, but grouped by school grade and misconception. For each misconception, this gives us the average change in predicted probability across grades. Users that were identified in two or more different grades were removed. Figure 6.7 shows the average probabilities.

Notably, the model predicts four dominant causes: operator relevant errors, operand related errors, same decade errors, and typos. And importantly, the figure suggests that these causes follow clear developmental patterns. The expected probabilities for both the operand related error and same decade error decrease across school grades, whereas the expected probabilities for typos increase. The operator relevant error shows the least clear pattern, but apart from the probability in grade 5 seems to suggest an increase as well.

These observations were confirmed by linear regression analyses. Table 6.5 shows the results of these analyses, evidencing significant differences between grades for each of the four misconceptions, although only modestly for operator relevant errors.

6.4 Discussion

Learners' error responses can provide insights into their applied strategies and cognitive processes. To benefit from this opportunity, the causes of the error responses must be determined. In this article, we proposed and investigated the accuracy of the Bayesian Misconception Tracing method. Importantly, we showed that the model outperforms a random selection from the misconceptions that may cause a set of observed errors. Also, we showed that a relatively small amount of observed errors is sufficient to correctly determine a possible cause of a subsequent error, in fifty to almost a hundred percent of the cases, depending on the student.

Moreover, when examining the performance of the model across school grades, the model identified clear developmental patterns. The model's predictions suggest that students' operand related and same decade misconceptions gradually decrease, whereas typos gradually increase. The prominence of operator relevant errors seems to slowly increase with school grade.

Altogether, these findings indicate that the model is very well suited for the identification of misconceptions. And it has important benefits. Not to be underestimated is the fact that it is intuitive. The primary source of the model is the relative number of observed errors each cause can explain. Although from a prediction perspective one might not be concerned about the model being intuitive or not, in an educational context it is a clear advantage. Both students and teachers generally value an understanding of the origin of inferences like these, rather than having to deal with black box analytics.

Also, the model suits relatively easy implementation in online learning environments. Given that a theoretical bigraph that links causes to errors is present, the actual calculations are lightweight, and may depend on a limited number of errors. And finally, in addition to being intuitive and lightweight, the model carries substantial weight as it is embedded in the well-understood Ising model (e.g., Kruis & Maris, 2016).

Clearly, as outlined in the introduction, important challenges exist in identifying student's misconceptions. The Bayesian Misconception Tracing method requires a theoretical model that links misconceptions to errors; a labour-intensive task to create, and we are not aware of areas in which this process was automated. On the other hand, many misconceptions in many areas have been identified in the literature, and this subject-independent method allows one to collect

these misconceptions and—given the availability of appropriate data—calculate predictions for each for the misconceptions.

Learning creates a challenge too. Misconceptions may naturally change over time, and indeed the whole purpose of diagnosis is to eliminate misconceptions. In other words, learning defeats stable misconceptions, and the suitability of the chosen window of observed errors crucially depends on this stability. We solved this issue by continuously tracking misconceptions over time, using a reasonably small window, and we showed that windows larger than seven to fifteen errors did not substantially benefit predictions. However, to better understand the influence of the chosen window, a theoretical understanding of developmental trends in misconceptions and the typical severity of different misconceptions across students would be valuable.

Being aware of these benefits and challenges, a thorough understanding of the model and its implications demands a discussion of four important issues. To begin with, we analyzed the model with error responses from an adaptive learning environment. The primary reason for using this data was the large number of both students and responses, and we see no reason to believe that the model would perform differently with data from a non-adaptive learning environment. In addition, although adaptive data was used, the chosen domain contains a very homogeneous set of items, and the bigraph that captures the theoretical relations between misconceptions and errors was specifically designed for the studied items. Within such a confined domain, one might safely assume that a student has a very limited amount of misconceptions. However, when analyzing responses from a variety of domains, or when multiple misconceptions are plausible for a single student, the method might be less appropriate.

Additionally, one important aspect of the model deserves note. This dynamic may occur when two or more misconceptions are present. Figure 6.8 shows three misconceptions and two observed errors. The two filled misconceptions are active for this particular student, each causing one of the errors. However, in this particular instance the model would predict the inactive misconception to cause the two observed errors, as it outweighs the other two misconceptions in its proportion of edges. This dynamic illustrates a dominance of misconceptions that are related to many different errors over misconceptions that are related to a few errors. Having said that, in most cases this appeal to the majority makes perfect sense. Given that one can assume a single misconception, the one cause that explains the most errors simply serves as the simplest theory.

Third, the proposed model cannot take correct responses into account. One may view this as a shortcoming, arguing that correct responses can carry counter-evidence for certain misconceptions. This is however not as straightforward as it may seem. A simple intuition is that a

Figure 6.8. A bigraph with three misconceptions (circles) and two observed errors (black squares). The black circles represent active misconceptions, whereas the white circle represents an inactive misconception. The method incorrectly predicts the inactive misconception to cause the two errors.

correct response invalidates any misconception in the domain of interest. However, in the case of single digit multiplication this could mean that a single correct response would invalidate all misconceptions. Obviously, this is not realistic, and forces us to acknowledge that students may have localized misconceptions, where some items are susceptible to their misconception, whereas others are not (i.e., students may use different strategies for different groups of items, such as correct memory recovery for some, and an erroneous procedure for others). Determining these clusters for individual students is an interesting avenue for future research.

Lastly, one might argue that the payoff of error analyses is limited. If a student makes errors, a teacher could simply provide additional instruction about the correct procedure, without the need to understand the student's specific misconception. Yet interestingly, Muller, Bewes, Sharma, and Reimann (2007) and Muller, Sharma, Eklund, and Reimann (2007) argue that, in the domain of science learning, this method can have an undesirable effect. They first show that correct explanations may sometimes actually *reinforce* students' misconceptions, and then show that discussing the misconception as part of the instruction can make students aware of it. Although it is unclear to which domains these findings generalize, it serves as an important warning to not just blindly assume the benefit of instructions on solely the correct procedure. On top of that, identifying the exact misconception can help select specific problems that target the misconception, and provide additional tailored practice.

Following up on the findings of Muller et al., we suggest viewing the diagnosis and treatment of errors as an actual instructional design principle. Contrary to errorless learning—the idea that learning does not benefit from errors—it should be accepted that misconceptions are inevitable, and targeted diagnosis and treatment of errors might really benefit the student. In such

a diagnose-and-treat model of learning, learning can be described in terms of the elimination of misconceptions. Also, because of this focus on misconceptions, the instruction and practice that is subsequently provided will target what the student does not understand, rather than what the student already knows.

6.4.1 Conclusions

The completed analyses are essential in understanding the optimum performance of the model. Next, learning interventions can be executed on the basis of its estimations. Interventions are not only an ultimate goal of error analyses—tailoring instruction or practice to the misconception of a specific student—but too are a great tool in further determining its accuracy. Given that an intervention is effective for a given student with a given misconception, its success reflects the accuracy of the model. The proof is in the pudding.

n_causes	window_size	term	estimate	std.error	statistic	p.value
one	1	(Intercept)	0.582	0.006	101.829	0.000
one	1	sample	0.004	0.008	0.475	0.635
one	1	max	0.008	0.008	0.964	0.335
one	3	(Intercept)	0.522	0.006	83.509	0.000
one	3	sample	0.101	0.009	11.393	0.000
one	3	max	0.185	0.009	20.962	0.000
one	7	(Intercept)	0.451	0.007	67.798	0.000
one	7	sample	0.267	0.009	28.379	0.000
one	7	max	0.313	0.009	33.236	0.000
one	15	(Intercept)	0.385	0.007	54.888	0.000
one	15	sample	0.382	0.010	38.453	0.000
one	15	max	0.396	0.010	39.852	0.000
one	30	(Intercept)	0.320	0.008	42.308	0.000
one	30	sample	0.459	0.011	42.959	0.000
one	30	max	0.461	0.011	43.142	0.000
two	1	(Intercept)	0.806	0.005	146.874	0.000
two	1	sample	-0.002	0.008	-0.290	0.772
two	1	max	-0.002	0.008	-0.258	0.796
two	3	(Intercept)	0.786	0.005	162.576	0.000
two	3	sample	0.037	0.007	5.357	0.000
two	3	max	0.064	0.007	9.433	0.000
two	7	(Intercept)	0.717	0.005	151.510	0.000
two	7	sample	0.143	0.007	21.303	0.000
two	7	max	0.158	0.007	23.578	0.000
two	15	(Intercept)	0.630	0.005	115.374	0.000
two	15	sample	0.243	0.008	31.511	0.000
two	15	max	0.244	0.008	31.629	0.000
two	30	(Intercept)	0.560	0.007	82.211	0.000
two	30	sample	0.303	0.010	31.486	0.000
two	30	max	0.304	0.010	31.583	0.000

Table 6.3. Linear regression results for differences in prediction accuracy across methods of selecting the predicted cause, separately for the predictions of one or two causes, and five window sizes. Results show the differences with the at random selected predicted cause.

n_causes	method	term	estimate	std.error	statistic	p.value
one	max	(Intercept)	0.725	0.003	231.526	0.000
one	max	three	0.118	0.010	11.951	0.000
one	max	seven	0.056	0.010	5.669	0.000
one	max	fifteen	0.017	0.010	1.722	0.085
one	max	thirty	-0.000	0.010	-0.018	0.985
one	sample	(Intercept)	0.695	0.003	223.922	0.000
one	sample	three	0.038	0.010	3.831	0.000
one	sample	seven	0.095	0.010	9.690	0.000
one	sample	fifteen	0.049	0.010	4.981	0.000
one	sample	thirty	0.012	0.010	1.198	0.231
one	random	(Intercept)	0.452	0.003	167.594	0.000
one	random	three	-0.059	0.009	-6.960	0.000
one	random	seven	-0.071	0.009	-8.351	0.000
one	random	fifteen	-0.066	0.009	-7.718	0.000
one	random	thirty	-0.066	0.009	-7.691	0.000
two	max	(Intercept)	0.854	0.002	382.030	0.000
two	max	three	0.046	0.007	6.574	0.000
two	max	seven	0.025	0.007	3.519	0.000
two	max	fifteen	-0.001	0.007	-0.112	0.911
two	max	thirty	-0.010	0.007	-1.419	0.156
two	sample	(Intercept)	0.845	0.002	365.428	0.000
two	sample	three	0.019	0.007	2.577	0.010
two	sample	seven	0.038	0.007	5.130	0.000
two	sample	fifteen	0.014	0.007	1.851	0.064
two	sample	thirty	-0.010	0.007	-1.376	0.169
two	random	(Intercept)	0.700	0.003	248.672	0.000
two	random	three	-0.020	0.009	-2.251	0.025
two	random	seven	-0.069	0.009	-7.699	0.000
two	random	fifteen	-0.087	0.009	-9.800	0.000
two	random	thirty	-0.070	0.009	-7.869	0.000

Table 6.4. Linear regression results for differences in prediction accuracy across five window sizes, separately for the predictions of one or two causes, and three methods of selecting the predicted cause. Results show the difference with the preceding window size, to see the effects of the additional increases in window size.

cause	term	estimate	std.error	statistic	p.value
operand.related	(Intercept)	0.147	0.015	9.618	0.000
operand.related	grade	-0.018	0.003	-6.986	0.000
operator.relevant	(Intercept)	0.788	0.033	24.230	0.000
operator.relevant	grade	0.011	0.005	2.133	0.033
same.decade	(Intercept)	0.136	0.016	8.419	0.000
same.decade	grade	-0.017	0.003	-6.389	0.000
typo	(Intercept)	-0.070	0.030	-2.338	0.020
typo	grade	0.024	0.005	4.994	0.000

Table 6.5. Linear regression results for grade differences in users' average predicted probabilities of the four most common causes. Results show whether predicted probabilities are stable across grades.

Art is a lie that makes us realize truth, at least the truth that is given us to understand.

Pablo Picasso

7

Idiographic intelligence

Summary

The positive manifold of intelligence has fascinated generations of scholars in human ability. In the past century, various formal explanations have been proposed, including the dominant *g*-factor, the revived sampling theory, and the recent multiplier effect model and mutualism model. In this article we propose a novel idiographic explanation. We formally conceptualize intelligence as evolving networks, in which new facts and procedures are wired together during development. The static model, an extension of the Fortuin-Kasteleyn model, provides a parsimonious explanation of the positive manifold and intelligence's hierarchical factor structure. We show how it can explain the Matthew effect across developmental stages. Finally, we introduce a method for studying growth dynamics. Our truly idiographic approach offers a new view on a century-old construct, and ultimately allows the fields of human ability and human learning to coalesce.

7.1 Introduction

Formal models of intelligence have greatly evolved since Spearman's (1904) fundamental finding of the positive manifold: the robust pattern of positive correlations between

scores on cognitive tests (Carroll, 1993). In explaining this manifold, contemporary models have diverged from the popular reflective latent factor models (e.g., Spearman, 1927), to various proposed mechanisms of emergence (Conway & Kovacs, 2015). Models that have been key in expanding the realm of explanatory mechanisms include sampling models (Bartholomew, Deary, & Lawn, 2009; Kovacs & Conway, 2016; Thomson, 1916; Thorndike, Bregman, Cobb, & Woodyard, 1926), gene-environment interaction (GxE) models (Ceci, Barnett, & Kanaya, 2003; Dickens, 2007; Dickens & Flynn, 2001, 2002; Sauce & Matzel, 2018), and network models (van der Maas et al., 2006). We embrace this trend, as exploring alternative mechanisms for the positive manifold may significantly aid us in our understanding of intelligence (Bartholomew, 2004).

Dickens and Flynn's and van der Maas et al.'s recent contributions have been serious attempts to encapsulate development into the theory of general intelligence. Here, we combine ideas from both their GxE and network approaches, to conceptualize general intelligence as dynamically growing networks. This creates a completely novel conception of the shaping of intelligence—idiographic and developmental in nature—that uncovers some of the complexity thus far obscured. Our proposed formal model not only explains how idiographic networks can capture intelligence's positive manifold and hierarchical structure, but moreover opens new avenues to study the complex structure and dynamic processes of intelligence at the level of an individual.

The paper is divided into two parts. In the first part, we briefly review current formal models of intelligence, and discuss the desire to give idiography and development their deserved place within this tradition of formal models. In the second part, we introduce an elaborate developmental model of intelligence. We explain how the model captures various stationary and developmental phenomena, and portray an individual's complex cognitive structure and dynamics. Finally, in the Discussion section we explore the model's implications and limitations.

7.2 Formal models of intelligence

In this first part, we begin with a discussion of the primary modeling traditions in intelligence research. This discussion is followed by an analysis of what we call idiographic and developmental blind spots in formal models of intelligence: the failure of in particular factor models to seriously consider idiography and development. Finally, we discuss a formal mechanism for development, called Pólya's urn scheme, to elucidate how surprisingly simple growth mechanisms can create phenomena that are key in the development of intelligence.

7.2.1 The positive manifold and its explanations

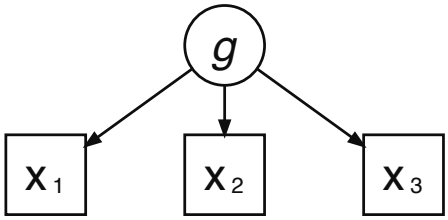
The first challenge for theories of general intelligence is to explain the pattern of positive correlations, the positive manifold, between scores on cognitive tests across individuals. Thus far, the proposed explanations form a colorful palette of diverse conceptions. We summarize four influential explanations—captured in Figure 7.1—that were formalized in various theories of intelligence. This summary requires two remarks. First, we use the terms model and theory interchangeably. However, whereas models provide a conceptual representation (e.g., the factor model), strictly speaking they carry no theoretical load. Theories on the other hand, add theoretical interpretation to a model (e.g., the—rather vague—theory that the factor named *g* represents mental energy). Here, we consider conceptually different *models*, that have been used for serious *theories* of intelligence. Second, we do not consider explanations of the positive manifold that were not formalized—such as the explanations by Tryon (1935) and Ferguson (1954)—even-though they are by no means less interesting. In the following, we will shortly introduce each of the four models one by one, and discuss their differences and similarities.

Factor models

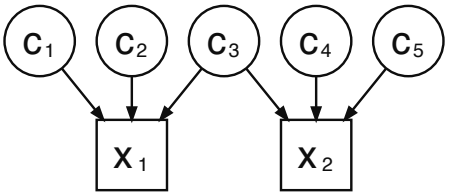
Spearman (1904) not only discovered the positive manifold, but also gave it an elegant explanation. In his two-factor model, Spearman (1927) introduced the general factor *g*, assuming the existence of an underlying common source that explains the scores on multiple cognitive tests. Although lacking a formal explanation, Spearman primarily hypothesized it to be some source of mental energy. Importantly, whereas intelligence generally is viewed as an intra-individual characteristic, *g* stems from an inter-individual observation, and must be understood alike (Borsboom, Kievit, Cervone, & Hood, 2009). Spearman's factor-analytic approach has inspired many scholars to propose models in the same tradition.

Among the most influential contributions is Thurstone's (1938) theory of primary mental abilities. Thurstone initially argued that Spearman's unitary trait is a statistical artifact, and proposed a multi-factor model consisting of seven distinct latent constructs. Other theorists followed this approach, with Guilford (1967, 1988) pushing the limits by ultimately including 180 factors in his influential structure of intellect model. Thurstone, on the other hand, eventually had to climb down, as verifying his model on a new empirical sample compelled him to add a second-order unitary factor to his model. This set the stage for various hierarchical models of intelligence (Ruzgis, 1994).

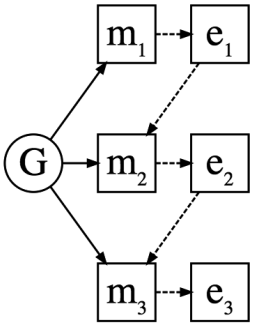
The marriage between a multi-factor theory and a hierarchical theory has evolved into what



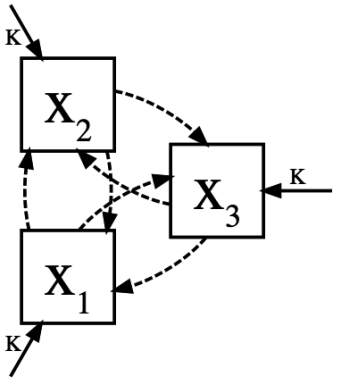
(a) The factor model. The positive manifold stems from a single factor (g) that influences scores on cognitive tests (x_i).



(b) The sampling model. The positive manifold stems from the unavoidable overlap in cognitive tests. Cognitive tests (x_i) are insufficiently specific to measure individual cognitive units (or bonds, c_i).



(c) The multiplier model. The positive manifold emerges from gene-environment interactions. Measurements of IQ over time (m_i) are influenced by both genetic endowment (G) and interactions with the environment over time (e_i).



(d) The mutualism model. The positive manifold emerges from positive local interactions between cognitive processes (x_i). Each process is constrained by genetic endowment and environmental resources (K).

Figure 7.1. Four explanations of the positive manifold (simplified). Circles represent unobserved entities, whereas boxes represent observed entities. Dashed lines represent relations that have an influence over time.

is currently the most widely supported factor-analytic model of intelligence: the Cattel-Horn-Carrol (CHC) theory (McGrew & Flanagan, 1998). The first theory, Cattell and Horn's G*f*-G*c* model (Cattell, 1963; Horn & Cattell, 1966), postulates eight or nine factors, including the well-known fluid and crystallized intelligence (derived from Hebb's intelligence A and B; R. E. Brown, 2016; Hebb, 1949). The second theory, Carroll's three-stratum hierarchy (Carroll, 1993), postulates a hierarchy of three levels, or strata, consisting of a general ability, broad abilities, and narrow abilities. In CHC theory, the broad stratum consists of Cattell and Horn's primary abilities.

Sampling models

In the last decade, after a century long dominance of factor theories of general intelligence, three alternative theoretical approaches to explaining the positive manifold have been (re-)introduced. The first, the sampling (or bonds) model, was originally advocated by Thomson (1916, 1951) and Thorndike et al. (1926) as an alternative to Spearman's *g* theory. In the sampling model the positive correlations between test scores origin from the shared underlying basic processes (or bonds) those tests tap into. That is, cognitive tests are insufficiently specific, and the overlap in shared processes will necessarily result in positive correlations between tests.

Bartholomew et al. (2009), Bartholomew, Allerhand, and Deary (2013), and more recently Kovacs and Conway (2016), re-introduced the sampling theory of general intelligence. The former generalized Thomson's model to account for multiple latent factors, and the latter further extended sampling theory, in order to account for the effects of domain-general executive processes, identified primarily in research on working memory, as well as more domain-specific processes.

Gene-environment interaction models

A decidedly more recent explanatory mechanism for the positive manifold was introduced by Dickens and Flynn (2001, 2002). In aiming to solve the paradox of both high heritability estimates in IQ and large environmental influences on IQ, they hypothesized a gene-environment interaction, where through reciprocal causation, IQ influences ones close environment and that environment in turn influences ones IQ, creating a multiplier effect. Moreover, rises in the IQ of others may also affect ones IQ, a so-called social multiplier. Dickens (2007) extended the multiplier model to include multiple abilities.

In an effort to reconcile genetic and environmental claims on cognitive ability, Ceci et al. (2003) nicely summarized a variety of models that are build upon such a multiplier principle. Among Dickens and Flynn's model for intelligence, they distinguished four other areas where models with similar dynamics have been proposed, such as in dynamical systems theory and bio-ecological theory. Each of these areas provide a compelling case for multiplier effects in cognitive development.

Network models

The final new explanation of the positive manifold, based on network modeling, was introduced by van der Maas et al. (2006). Inspired by dynamical explanations of the health of shallow lakes, the idea of their mutualism model is that the cognitive system consists of many basic processes that are connected in a network with primarily positive interactions. During development, the initially uncorrelated basic processes become correlated due to these positive reinforcements. And indeed, these mutual positive reinforcements too exhibit a multiplier effect.

This network approach has particularly resonated in the domain of psychopathology, resulting in a recent surge of research (see Borsboom, 2017, for a comprehensive overview). In intelligence, van der Maas, Kan, Marsman, and Stevenson (2017) extended the mutualism model, allowing for test sampling, mutualistic relations and multiplier effects, and central cognitive abilities.

Differences and similarities

What unites these discussed models, is that they all explain the positive manifold equally well. Yet, there is a lot that sets them apart. In *g* factor models, the correlations are due to a common source of cognitive performance in many domains. The *g* factor is understood as a so-called reflective latent variable. That is, in theorizing the nature of intelligence, the general factor is understood as a causal entity. Spearman's notion of mental energy is an example of that. Importantly, this notion of a psychological *g*, reflective and causal in nature, is a hypothesized one and not uncontroversial.

In the mutualism model there is no such common source. Rather, the positive manifold emerges from the network structure. The nevertheless apparent statistical *g* factor is interpreted as formative variable, as an index variable of the general quality of the cognitive system, akin to economical indexes such as the Dow Jones Industrial Average. Contrary to psychological *g*, this psychometric *g* is well-established and non-controversial (e.g., Carroll, 1993).

In sampling theory, the statistical *g* factor should also be interpreted as a formative variable. But that is not to say that sampling theory and the mutualism model are very similar. In sampling theory, the positive manifold is essentially a measurement problem. If we would be able to construct very specific tests, targeted at the fundamental processes, the overlap in measurement would disappear, and so will the correlations between tests. In the mutualism model on the other hand, the correlations are real, created during development, and will not disappear when IQ tests become more specific.

In both the multiplier effect model and mutualism model, the positive manifold emerges from positive reciprocal reinforcements. However, the two models differ in several key respects. Most importantly, the mutualism model proposes an internal developmental process, whereas the multiplier model depicts development through an interaction with the external environment.

Finally, and importantly, as for instance Bartholomew et al. (2009) and Kruis and Maris (2016) note, *g* theory and sampling theory, and factor models and network models, cannot be statistically distinguished on the basis of correlation indexes alone, nor do they necessarily contradict one another. Van der Maas et al. (2017) illustrate this in their unified model of general intelligence. However, this is not to say that these models are equivalent with respect to their explanatory power. Each conception might tap into a different granularity of general intelligence, ultimately aiding us in our understanding of the construct. Also, it is not to say that the models cannot be distinguished. Time-series data and experimental interventions may very well distinguish between the models. Marsman et al. (2018) describe these issues in more detail.

7.2.2 The developmental blind spot

Intelligence cannot be understood in isolation. It is a product of genetic, environmental, and developmental factors, and must be considered within this complex context. Nonetheless, particularly the *development* of intelligence has long been an afterthought in its formal modeling tradition. We briefly provide two possible reasons for this unfortunate fact, and in the process aim to convey the importance of exploring formal developmental notions of intelligence.

One dominant model

A first reason for the developmental blind spot is the dominance of *g* theory. Although *g* does not necessarily nail down its origin, be it genetic, environmental, or both, it does not naturally capture development. Ackerman and Lohman (2003) concisely summarized this, by explaining

that "[o]ne of the most intractable problems in evaluating the relationship between education and *g* is the problem of development and age. As near as we can tell, *g* theories have failed to provide any account of development across the lifespan." On top of that, Nisbett et al. (2012) observe that "[t]he high heritability of cognitive ability led many to believe that finding specific genes that are responsible for normal variation would be easy and fruitful." Yet, on the contrary, intelligence is a genuinely complex construct, and this presumed reification fallacy has thus far produced limited insight other than the realization that many genes may be involved with small effects (e.g., Lee et al., 2018), and a gained understanding of the importance of environmental factors. Indeed, intelligence is not solely nature's responsibility, and it should not come as a surprise that Nisbett et al.'s conclusion, directly following the previous quote, is as ironic as unsettling: "So far, progress in finding the genetic locus for complex human traits has been limited."

The closest factor models have come to providing an account of development across the lifespan, is in Cattell's (1987) investment theory. In his landmark book, Cattell hypothesizes that one develops a pool of crystallized intelligence by the 'investment' of fluid intelligence, in conjunction with the "combined result of the form of the school curriculum, and of the social, familial, and personal influences which create interest and time for learning simultaneously in any and all forms of intellectual learning." This idea, derived from Hebb's (1949) two intelligences (intelligence A, "an *innate potential*, the capacity for development", and intelligence B, "the functioning of a brain in which development has gone on") (R. E. Brown, 2016), was never formalized, but is argued to explain the Matthew effect (Schalke-Mandoux, 2016)—a key developmental phenomenon discussed in the next section.

In his investment hypothesis, Cattell thus explicitly sketched an evident role for the environment, where genes and the environment are united to explain individual differences in intelligence. More recent insights however, demand a further integration of the two. In discussing the puzzling heritability increase, Plomin and Deary (2014) explain: "Genotype-environment correlation seems the most likely explanation in which small genetic differences are magnified as children select, modify and create environments correlated with their genetic propensities. This active model of selected environments—in contrast to the traditional model of imposed environments—offers a general paradigm for thinking about how genotypes become phenotypes." This developmental notion of a gene-environment interaction (Tabery, 2007) suggests a causal mechanism between the two that may give rise to the phenotype IQ.

Dickens and Flynn's novel formal multiplier model capitalizes on such a developmental gene-environment interaction, giving the high heritability of IQ a convincing explanation. Accord-

ing to their model (see Figure 7.1c), children not only actively select their environment in accordance with their genetic endowment, but additionally this environment influences their IQ, creating reciprocal causal relations between the phenotype and the environment. This way, Dickens and Flynn arrive at a truly developmental model of intelligence.

Even more recently, and following a decidedly different track, van der Maas et al. showed how interactions between cognitive processes are capable of explaining high heritability. Whereas Dickens and Flynn broke down *g* into genetic and environmental factors, van der Maas et al. proved that a single underlying factor is no intrinsic requirement for explaining some of the most important phenomena in intelligence.

One dominant phenomenon

A second reason for the developmental blind spot is the primary focus on the positive manifold. Thanks to the work of particularly Spearman (1904) and Carroll (1993), the positive manifold is an undisputed phenomenon. In turn, static factor models have provided an elegant parsimonious explanation of this phenomenon. Yet, the positive manifold lacks a similarly strong developmental companion that can function as a yardstick for the proposed models. Cattell (1987) beautifully stresses the importance of such a phenomenon: "The theorist who wants to proceed to developmental laws about abilities – who wants to be 'dynamic' in his explanations of the origin, growth, and nature of intelligence – must be patient to make and record observations first. He can no more focus meaningful movement without this 'description of a given moment' than a movie director can get intelligible movement in a film without the individual 'static' frames themselves presenting each a clearly focused 'still.'"

One phenomenon, the Matthew effect, results from exactly those descriptions of given moments: static frames that have been put in chronological order, to give a description of the development of cognitive abilities. The Matthew effect is characterized by initially diverging yet increasingly stable patterns of development, as illustrated in Figure 7.2a, and may serve as a primary candidate for the role of developmental companion to the positive manifold.

Originally coined by Merton (1968) to describe the widening gap in credit that scientists receive during their career, the term Matthew effect refers to the popular catchphrase 'the rich get richer and the poor get poorer', and is named after the biblical figure Matthew. Although the Matthew effect does not necessarily involve the poor getting poorer, it does involve a widening gap between the rich and the poor, where the rich and the poor can be metaphors (e.g., for the skilled and the unskilled).

The Matthew effect is in no way an isolated phenomenon. Although the preferred term varies between (and within) disciplines, including cumulative advantage and preferential attachment, the intended process is essentially the same. Other related terms on the other hand, such as the fan-spread effect and power law, may refer to the observed effect rather than the underlying process (Bast & Reitsma, 1998; Perc, 2014). Perc (2014) provides a comprehensive overview of the Matthew effect in empirical data.

Stanovich (1986) was probably the first to link the Matthew effect to education, in an attempt to conceptualize the development of individual differences in reading. In this field, he argued, initial difficulties with reading acquisition can steadily propagate through reciprocal relationships with related skills, ultimately creating more generalized deficits.

Yet, the effect is not undisputed. For instance, Shaywitz et al. (1995) found a Matthew effect for IQ, but not for reading, when controlling for regression to the mean. Moreover, there is also evidence for the opposite developmental trajectory, the so-called compensation effect. This effect, for instance found by Schroeders, Schipolowski, Zettler, Golle, and Wilhelm (2016), describes a closing rather than widening gap.

Complicating things further, it is often hypothesized that both the factors driving and combating the gap influence development. This is for instance clearly explained by Schroeders et al. (2016): "[i]t seems that the compensation effect of a formalized learning environment counteracts the effect of cumulative advantages that is present in a non-formalized setting." This at least provides an explanation for the more ambiguous status of these two developmental phenomena, especially when compared to the positive manifold.

It should not come as a surprise that Protopapas, Parrila, and Simos (2014) suggest to focus on the reciprocal relations that drive the gap, rather than on estimating the gap itself. And intriguingly, one deceptively simple mechanism—driven by such reciprocal relations—can actually explain the Matthew and compensation effects. However, before we introduce the mechanism, we first briefly discuss a second blind spot: idiography.

7.2.3 The idiographic blind spot

Confusingly, whereas it is generally understood that intelligence is a property of a single individual, many key phenomena in intelligence—including the positive manifold, and Matthew and compensation effect—reflect structural differences between multiple individuals. Jensen (2002) warns us for this confusion, by explaining that

> "[i]t is important to keep in mind the distinction between *intelligence* and *g*.

> [...] The psychology of intelligence could, at least in theory, be based on the study of one person, just as Ebbinghaus discovered some of the laws of learning and memory in experiments with $N = 1$. [...] The *g* factor is something else. It could never have been discovered with $N = 1$, because it reflects *individual differences* in performance on tests or tasks that involve any one or more of the kinds of processes just referred to as *intelligence*. The *g* factor emerges from the fact that measurements of all such processes in a representative sample of the general population are positively correlated with each other, although to varying degrees."

Jensen's warning is far from frivolous, and concerns the broader field of psychological science. For instance, Molenaar (2004) and Borsboom et al. (2009) have expanded on this cautionary tale, by showing that intra-individual (idiographic) interpretations of inter-individual (nomothetic) findings lead to erroneous conclusions, only exempting cases where very strict assumptions are met. Consequently, models of individual differences should be based on models of the individual—a message that is reinforced by Molenaar's urgent call for an idiographic approach to psychological science: explaining nomothetic phenomena with idiographic models.

One promising approach to idiography are network models, briefly discussed in relation to van der Maas et al.'s mutualism model. Indeed, networks are an ideal (and idealized) tool for modeling the individual. These networks, or graphs, are a general and content-independent method for representing relational information. It graphically represents entities, typically visualized as circles called *nodes* or *vertices*, and their relations, typically visualized as lines called *edges*, or *links*. As networks are content-independent, few sciences, if any, fail to appreciate their value. Notable applications span from social networks, describing relations between individuals (e.g., Duijn, Kashirin, & Sloot, 2014), to attitude networks, describing relations between attitudes across individuals (e.g., Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017), and psycho-pathological networks, describing relations between symptoms within individuals (e.g., Kroeze et al., 2017).

In the next section, we discuss a simple yet powerful mechanism that explains idiographic development. Although the mechanism is too simplistic for our ultimate aim—the formal model of intelligence introduced in second part—it convincingly conveys the power of simple developmental mechanisms. We show how it explains the previously discussed Matthew and compensation effect, as well as the third source phenomenon. Finally, we use a simple network transformation to illustrate the benefit of networks in idiographic science.

7.2.4 Idiographic development: the case of Pólya's urn

Before we proceed with the second part—the introduction of the model—we briefly discuss an elegant abstraction of a growth process. The Pólya-Eggenberger urn scheme (Eggenberger & Pólya, 1923), or simply Pólya's urn, intuitively mimics a system that grows dynamically by means of preferential attachment, and hence gives us a convenient tool to illustrate a basic mechanism of development. Moreover, we use the obtained surface understanding of Pólya's urn to clarify not only the Matthew and compensation effects, but also the third source phenomenon.

A brief example may clarify the mechanism. Imagine a child receiving a tennis racket for her birthday. Before her first tennis lesson she practices the backhand twice at home, incorrectly unfortunately. Then, during the first lesson, her trainer demonstrates her the correct backhand. She now has three experiences, two incorrect and one correct. Now, suppose her backhand development is based on a very simple learning schema. Whenever a backhand return is required she samples from her earlier experiences, and the sampled backhand is then added to the set of earlier experiences. How will she develop? That is, how will her backhand develop on the long term? And what is the long term expectation for her equally talented twin sister with the same trainer?

Pólya's urn gives us an important intuition. It is represented as an urn that contains two different-colored balls, say black and white. One ball is randomly drawn from the urn and replaced by two balls of the same color, a procedure that is repeated n trials. Interestingly, the time course of this process is rather counter-intuitive. One might expect this process to diverge to extreme values, but it rather progresses towards a random number between zero and one[1]. As can be deduced from this process, it ensures dynamical growth by means of preferential attachment: the urn grows each trial, with a preference towards the most abundant color.

The applicability of Pólya's urn is endless, and various modifications have been proposed to accommodate a diversity of issues. In Eggenberger and Pólya's original paper, the number of replaced balls can be of any positive value, and Mahmoud (2008) describes a number of

[1]Intuitively, the process can be understood as follows. At $t = 0$, the space of probable outcomes after an infinite amount of trials is $0 < p(white) < 1$. One might imagine that after each trial this space of probable outcomes becomes smaller, as either the upper or lower limit becomes less probable. Ultimately, the space of probable outcomes becomes infinitely small, and the process is said to have stabilized. Indeed, approaching the limit of 1 or 0 in the long run remains a possibility at any point in time, but the bandwidth of *probable* outcomes becomes ever smaller. The outcomes follow a uniform distribution in case of an equal distribution of color at the start of the process, or a beta-binomial distribution otherwise. Finally, the earliest trials weigh the most in determining the probable outcomes, as those most severely alter the subsequent proportion of balls.

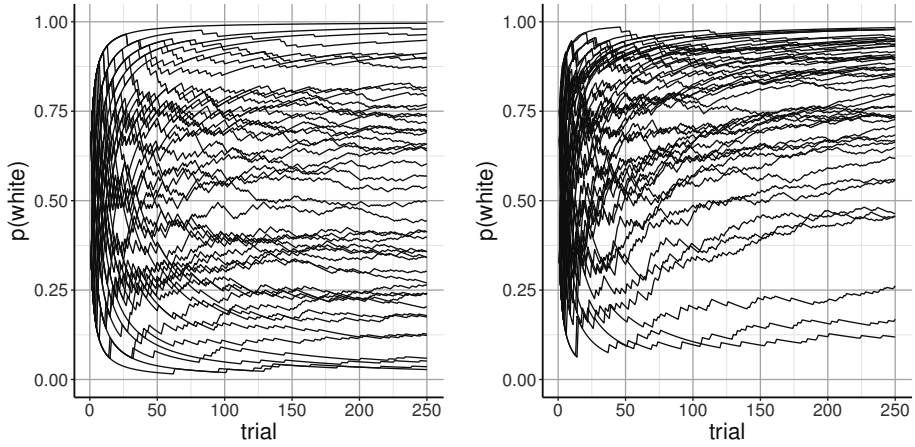
modifications to this basic scheme. For instance, Friedman's (1949) urn scheme allows one to replace a ball with not only balls of the same color, but also of the other color, and Bagchi and Pal (1985) further generalized this such that the number of balls to replace can depend on the color of the drawn ball. Other examples include urns where the probability of drawing a ball depends on how long it has been in the urn (Arnold, 1967), urns where more than two colors are involved (Athreya & Karlin, 1968), urns where new colors can randomly arise (Hoppe, 1984), and urns where multiple balls can be drawn at once (M.-R. Chen & Wei, 2005). Important applications range from evolution of species (Hoppe, 1984) to unemployment (Heckman, 1981).

Crucially, models of contagion—such as Pólya's urn—cannot be statistically distinguished from factor models. This is analogous to the previously discussed incapacity of statistical models to distinguish between factor models, sampling models, and network models. Greenwood and Yule (1920) provided the probability distributions that result from contagious processes and latent causes, and later it was realized that both distributions can be rewritten into the beta-binomial distribution. The importance of this fact for epidemiology, where it is well-known, can hardly be overlooked: imagine combating Ebola from an entirely genetic perspective rather than preventing contagion. However, in intelligence research this realization is just as fundamental: the fit of a statistical model cannot illuminate the actual underlying causal processes.

Pólya's urn and the Matthew effect

Here, we are concerned with the urn's ability to simulate typical developmental patterns of cognitive ability. We consider the unmodified version of Pólya's urn that we described previously. Figure 7.2a shows the proportions of white balls for 250 trials from 50 independent urns. As you can see, the urn compositions quickly diverge. The earlier trials have the largest effects, with gradually decreasing influence over time. Indeed, this pattern closely resembles the general Matthew effect, where at the start within-person variance is high, gradually decreasing over time, and between-person variance is low, gradually increasing over time.

Pólya's urn might thus be conceived as a model for a developmental process that produces a Matthew effect. The initial configuration of the urn depicts the genetic component, whereas the trials represent the environmental experiences. The white balls can for instance represent skills that reinforce advantageous experiences, whereas the black balls might represent misconceptions that reinforce disadvantageous experiences. By means of a strict random process, the urns' configurations diverge in a similar vein as in Dickens and Flynn's multiplier process. In this Pólya process, skills and misconceptions are reinforced, ultimately growing towards a stable



(a) Pólya urn demonstration of the Matthew effect. Regardless of its color, each drawn ball is replaced with two balls of the same color.

(b) Pólya urn demonstration of the compensation effect. Drawn white balls are replaced with either two or three white balls ($p = .5$), whereas drawn black balls are always replaced with two black balls.

Figure 7.2. Pólya urn demonstrations of the Matthew effect and compensation effect. Starting with an urn that contains a white and a black ball, in each trial the drawn ball is replaced with two or three balls of the same color, depending on the desired effect. The figures show the development of the proportion of balls for 50 independent urns.

state.

Importantly however, one significant additional property over the multiplier mechanism must not be overlooked. As the distribution of color of each independent urn is exactly equal at the start, it is shown that Pólya's urn does not require initial genetic differences for the multiplier process to do its work. This illuminating effect is not necessarily an artificial property of Pólya's urn: for example, Freund et al. (2013) show that individual differences actually can emerge in genetically identical mice.

Pólya's urn and the compensation effect

Naturally, environmental influences are less random than assumed here. One strong systematic influence is formal education, which is hypothesized to create compensatory effects that counteract the Matthew effect. Figure 7.2b shows what this effect could look like. To obtain the effect, we slightly adapted Pólya's urn to allow for one possible effect of education. Rather than reinforcing an advantageous experience with one extra ball, we now reinforce it with two extra balls, with $p = .5$, while keeping the rule for reinforcing disadvantageous experiences the same. An adaptation like this can for instance be conceptualized as the beneficial effect of practice and instruction in education. Similarly, remediation of disadvantageous experiences such as errors may too create a compensation effect, and can be modeled by not reinforcing such an experience with an extra ball.

Pólya's urn and the third source

The third source of developmental differences (Kan, Ploeger, Raijmakers, Dolan, & van der Maas, 2010; Molenaar, Boomsma, & Dolan, 1993) refers to phenotypic variability that cannot be attributed to either genetic or environmental factors. To explain this phenomenon, both papers proposed rather complicated nonlinear models. Conveniently, the third source becomes directly apparent in Pólya's urn. In the two examples of Pólya processes in Figure 7.2, both the genetics (the initial urn configurations) and the environment (the rules for drawing and replacing the balls) are identical. Yet, the developmental trajectories vary greatly.

Pólya's urn as a network

By transforming the example of Pólya's urn to a network representation, the benefit of networks in an idiographic science is clearly shown. Conveniently, Pólya's urn permits a simple

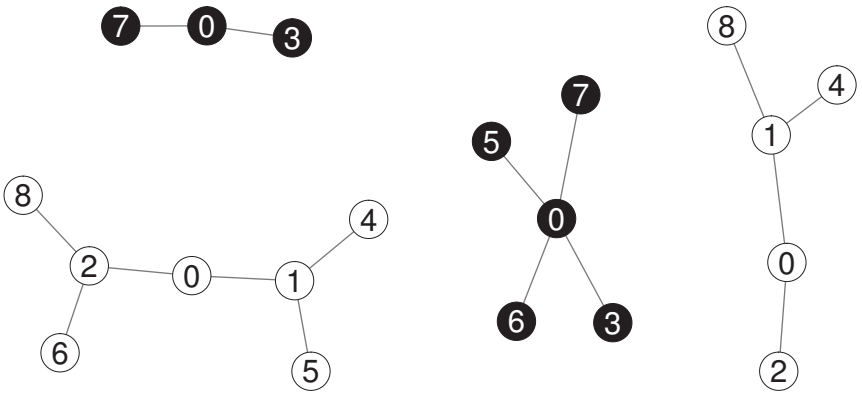


Figure 7.3. Two instances of a Pólya's urn network. Both networks started with a single black node and a single white node ($t = 0$). Also, the networks share an identical growth mechanism: a new node is randomly connected to one of the existing nodes and copies its color. The numbers show the time points at which the nodes were added.

transformation to such a network representation. Imagine a network with two types of nodes; black nodes that represent some kind of misconception, and white balls that represent some kind of correct conception. Analogous to the urn example, the initial network may consist of a disconnected black and white node. Now, on each trial, one node is randomly attached to one of the existing nodes in the network, copying its color. This simple mechanism ensures that the probability of a new node receiving a certain color is proportional to the number of existing nodes with that color, essentially a preferential attachment mechanism.

The Pólya's urn networks in Figure 7.3 (graphs created in *R* with *qgraph*; Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) each represent a different individual, and already illustrate their benefit in an idiographic science. However, the unidimensional structure of this simple urn example is too simplistic for the ultimate objective to describe a multidimensional intelligence. While preserving the network perspective, in the next part we introduce a new, formal, and multidimensional model of intelligence—a theory that explains both stationary and developmental phenomena, an abstraction that concretely describes an individual's skills and knowledge on the level of specific educational items, and an avenue for separating the role of genetics and the environment.

7.3 The wiring of intelligence

At the intersection of the issues discussed in the previous part, we propose a novel formal model of intelligence[2]. In this second part, we first describe a static model, hereafter referred to as *wired cognition*, and clarify how it explains two key stationary phenomena: the positive manifold and intelligence's hierarchical structure. We then describe a dynamic model, hereafter referred to as *wiring cognition*, and clarify how it explains developmental phenomena such as the Matthew effect and the age dedifferentiation hypothesis. The model's composition of a static and dynamic part reflects its twofold aim—explaining stationary and developmental phenomena—and stresses the poor balance in substantiation of the phenomena in both categories. Moreover, it enables the static and dynamic part to be assessed and further developed in relative isolation.

7.3.1 Statics: *wired* cognition

We conceptualize intelligence as a network of interrelated cognitive skills or pieces of knowledge. In this network, $G = (V, E)$, the set of p distinct cognitive skills or pieces of knowledge (used interchangeably in the remainder of the text) are represented as labeled nodes V, and their possible relations as edges E. Unless otherwise stated, it is assumed that the set E contains all $p(p-1)/2$ possible relations between the p nodes of the network. To each node i in the network we associate a random variable that takes one of two values,

$$x_i = \begin{cases} +1 & \text{if the skill or knowledge } i \text{ is obtained} \\ -1 & \text{if the skill or knowledge } i \text{ is unobtained} \end{cases}$$

Furthermore, we associate to each edge $e = \langle i, j \rangle$ in E a random variable ω_e that also takes one of two values,

$$\omega_e = \omega_{\langle i,j \rangle} = \begin{cases} 1 & \text{if a direct connection between skills or knowledge } i \text{ and } j \text{ is present} \\ 0 & \text{if a direct connection between skills or knowledge } i \text{ and } j \text{ is absent} \end{cases}$$

This assembly of dichotomous nodes and edges thus forms our abstraction of idiographic intelligence. Additionally, two remarks must be made regarding the nodes. First, the model is

[2] Although mathematical notation is unavoidable, we keep it to a minimum and put it at the service of comprehensibility. Mathematical proofs that are not key to a basic understanding of the model are provided in Appendix B.

ignorant with respect to their exact substance, that is, the 'cognitive skills or pieces of knowledge' they represent. Second, besides the obtained and unobtained knowledge that the nodes represent, it is important to consider that the majority of possible nodes is absent from the cognitive network. To illustrate this, in the network conception of six-year-old Cornelius, nodes that reflect concepts like integrals are most likely unobserved. Thus, the actual presence of nodes depends on factors like maturation and education.

Fortuin-Kasteleyn

The definitions of nodes and edges give us a minimal description of the wired cognition network. Now, the model aims to describe the probabilities with which skills are either obtained or unobtained, and how they are related. It is this description of probabilities that enables us to explain the established stationary phenomena.

The following model, proposed by Fortuin and Kasteleyn (1972, hereafter referred to as FK) in the statistical physics literature, forms the basis of this approach,

$$p(x, \omega) = \frac{1}{Z_F} \prod_{e \in E} \left\{ \vartheta \, \delta_{(\omega_e, 1)} \delta_{(e)}(x) + (1 - \vartheta) \, \delta_{(\omega_e, 0)} \right\}, \tag{7.1}$$

where ϑ is a parameter of the model that describes the probability that any two skills become connected. The function $\delta_{(a,\, b)}$ is known as Kronecker's delta,

$$\delta_{(a,\, b)} = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}$$

and $\delta_{(e)}(x) = \delta_{(x_i,\, x_j)}$ for an edge $e = \langle i, j \rangle \in E$. Z_F is a normalizing constant.

An important property of the model, is that whenever two skills are connected to one another, they are necessarily in the same state, that is, they are either both present or both absent. Consequently, whenever two skills are in different states, that is, one skill is present while the other is absent, then these two skills cannot be connected to one another. With this simple rule, and the single parameter ϑ, the model can describe the joint probability distribution of both the nodes (i.e., skills or knowledge) and their relations.

Idiography

This unaltered FK model already has highly beneficial properties for the study of intelligence, making it a convenient point of departure. Notably, the properties of the FK model can also be studied by inspecting the marginal distribution that it implies on the nodes x (Grimmett, 2006, p. 9),

$$p(x) = \sum_{\omega \in \Omega} p(x, \omega).$$

In Appendix B, we show that for the FK model this marginal $p(x)$ characterizes a fully-connected network known as the Curie-Weiss model (Kac, 1968). Marsman, Maris, Bechger, and Glas (2015) and Epskamp, Maris, Waldorp, and Borsboom (2016) showed that this Curie-Weiss model generates data that are consistent with an item response theory model known as the Rasch model. In fact, all versions of the model relate to a specific instance of the multidimensional item response theory model (Marsman et al., 2018, also see Appendix B), which to some may provide a more familiar way of studying certain properties of the model.

Figure 7.4 illustrates a few of the properties thus far discussed. To begin with, we can ensure idiography, the first modeling principle. As the model is characterized by both random nodes and edges, both skills and their relations may vary across individuals. This is clearly seen in the differences between the individuals in the figure, Cornelius and Pete, which are instances of the exact same model. In Pete's network considerably more knowledge is obtained than in Cornelius' network, while at the same time it is less densely connected (both within domains and between domains). Also, interesting differences between domains exist, with Pete clearly performing differently on two of the four domains. The careful eye spots that the connected nodes only form clusters with either obtained knowledge or unobtained knowledge, a homogeneity that is dictated by the model.

Positive manifold

Then, the next property of this model we turn to, is the positive manifold it produces. As discussed in the introduction, the uncontested significance of this phenomenon renders a plausible explanation a burden of proof for any serious theory of intelligence. Intuitively, this property of the FK model is shown by the fact that the correlation between any two nodes x_i and x_j in the network is positive. This is the case whenever the probability that the two nodes are in the same state, $p(x_i = x_j)$, is larger than the probability that the two nodes are in a different state,

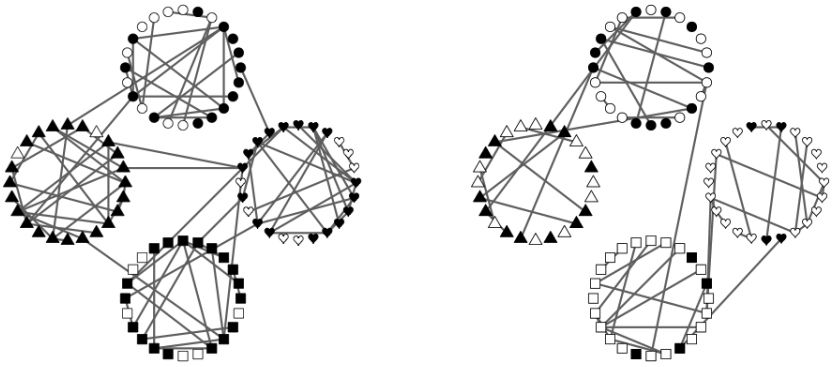


Figure 7.4. Cornelius and Pete—two instances of the FK model. The cognitive networks of both Cornelius and Pete consist of 96 nodes, equally distributed across four domains (represented by differently shaped nodes). Cornelius has 25 pieces of obtained knowledge (white nodes) and Pete has 65 pieces of obtained knowledge. Networks were generated with $\vartheta_W = .07$, $\vartheta_B = .005$, and $\mu = .03$.

$p(x_i \neq x_j)$. The following expression by Grimmett (2006, p. 11) confirms this,

$$p(x_i = x_j) = \frac{1}{2} + \frac{1}{2}p(i \leftrightarrow j) \geq \frac{1}{2},$$

where $p(i \leftrightarrow j)$ is the probability that nodes i and j are connected by an open path (i.e., are in the same cluster). Since $p(i \leftrightarrow j)$ is nonzero, $p(x_i = x_j)$ is strictly larger than 0.5, and the positive manifold emerges.

Figure 7.5 visualizes the positive manifold by means of a heatmap. Each of the patches in the figure represents a correlation between two nodes. The positive manifold can be deduced from the fact that all patches indicate a positive correlation. For this figure, we considered 1000 idiographic networks, construed from the FK model, with some important extensions laid out in the following sections.

Hierarchical structure

Arguably the second most important stationary phenomenon in intelligence is its hierarchical structure. Although the debate on whether g is organized in a bi-factor or higher-order structure continues to keep some intelligence researchers occupied, the fact that some cogni-

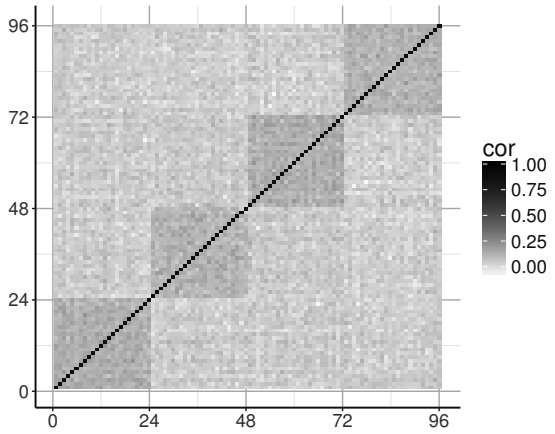


Figure 7.5. Heatmap of the correlational structure of the nodes of the FK model. Analogues to Spearman's very first observation of the positive manifold in the correlational structure of his cognitive tests, the exclusively positive patches illustrate the positive manifold as a constraining property of the FK model. Also, the hierarchical structure of intelligence is clearly reflected in the block structure. Networks were generated with $\vartheta_W = .07$, $\vartheta_B = .005$, and $\mu = .03$.

tive domains form clusters, with higher correlations within clusters than between clusters, is uncontested (e.g., Carroll, 1993; Spearman, 1904). This phenomenon is reflected in the typical block structure seen in correlation matrices of intelligence tests (see Figure 7.5).

Crucially, although the block structure clearly is present, the blocks are not fully isolated. Indeed, the small but meaningful correlations outside the blocks indicate interactions between the blocks. Simon (1962) termed this property *near decomposability*, and demonstrated its ubiquitous presence across a multitude of complex hierarchical systems. In his words, "[i]ntracomponent linkages are generally stronger than intercomponent linkages. This fact has the effect of separating the high-frequency dynamics of a hierarchy—involving the internal structure of the components—from the low frequency dynamics—involving interaction among components." The presence of both a general factor and a hierarchical structure can be seen to reflect this.

The human brain serves as a convenient illustration. The functional specialization of our brain can cause different cognitive tasks to tap into structurally dispersed brain areas (e.g., Fodor, 1983; Spunt & Adolphs, 2017), making within-community connectivity more likely, and between-community connectivity less likely. An example is the (increasing) functional specialization of arithmetic (e.g., Dehaene, 1999; Rivera, Reiss, Eckert, & Menon, 2005). Naturally, this

does not exclude the possible existence of processes that play a more general, or maybe central, role in cognitive functioning, such as executive functions.

In network models, the block pattern is generally referred to as a community structure. In the model proposed here, we impose such a structure by creating communities of nodes that have a higher probability of connecting with nodes within their community, than with nodes in other communities. Importantly, the model is ignorant with respect to the exact substance, or content, of a component, as for the theoretical purpose of the model the levels of the hierarchy are irrelevant. Yet, for illustrative purposes, in the specification of the model hereafter, we assume that the communities are known.

Suppose that there are two communities, say skills that are related to mathematics and skills that are related to language. We partition the set of nodes V into two groups, $V = (V_M, V_L)$, one associated to each community. Similarly, we partition the set of edges E into three parts, $E = (E_M, E_L, E_{ML})$, where E_M are the relations between different mathematics skills, E_L the relations between different language skills, and E_{ML} are all the relations between a mathematics-related skill and a language-related skill. In principle, we may associate to each community of skills c a unique probability ϑ_c to connecting its members, and associate to each pair of communities c and d a unique probability π_{cd} to connect members from the community c to members of the community d. However, for now it is sufficient to have one probability ϑ_W to connect the skills within a community and one probability ϑ_B to connect skills between two different communities.

The extended model is now as follows,

$$
\begin{aligned}
p(x, \omega) = & \frac{1}{Z_F} \prod_{e \in E_W} \left\{ \vartheta_W \, \delta_{(\omega_e, 1)} \, \delta_{(e)}(x) + (1 - \vartheta_W) \, \delta_{(\omega_e, 0)} \right\} \\
& \times \prod_{e \in E_B} \left\{ \vartheta_B \, \delta_{(\omega_e, 1)} \, \delta_{(e)}(x) + (1 - \vartheta_B) \, \delta_{(\omega_e, 0)} \right\},
\end{aligned}
$$

where E_W denotes the set of edges relating skills *within* a community, and E_B denotes the set of edges relating skills *between* communities. In the two-community structure that we described above, the within community edge set E_W is the union of the edges within the mathematics community and the edges within the language community, $E_W = E_M \cup E_L$, and the between edge set is simply E_{ML}. If $\vartheta_W > \vartheta_B$, it follows that $p(i \leftrightarrow j)$ is larger for any two skills i and j within the same community than for two skills i and j that are not a member of the same community. As a result, a hierarchical pattern of correlations emerge from the model, with higher correlations between skills within a community than between skills from different

communities.

Figure 7.4 shows two networks that arose from the extended model. The four communities in the networks are denoted by the various shaped nodes. Importantly, communities should not be confused with clusters—the nodes, or groups of nodes, that are isolated from the rest of the network. In these networks, the within-community connectivity was set to $\vartheta_W = .07$ and between-community connectivity to $\vartheta_B = .005$. Of course, these probabilities can be seen as an empirical estimation problem, which we do not consider here. Figure 7.5 shows how the community structure is reflected in the correlational structure of 96 nodes across 1000 extended FK models.

General ability

Notably, the FK model has no preference towards obtained or unobtained pieces of knowledge. However, in the proposed model, we assume that there actually is a preference, and indeed *towards general ability*. Reminded of the previously mentioned six-year-old named Cornelius, this preference reflects the facts that for instance education usually is at the level of the student and aims at attainable goals, and interactions with the environment fit the individual to a large extent. Moreover, anticipating the growth perspective that is introduced in the next section, it is evident that individuals tend to become *more* able rather than less able (that is, if we ignore cognitive decline due to for instance aging and degenerative diseases).

To account for this bias towards aptitude, we impose a so-called *external field* that is minimally positive. External fields are used in physics to represent some outside force that acts upon variables in a network, and to understand this idea, magnetism provides a clarifying illustration. In the study of magnetism, variables in the network may represent the electrons in a piece of iron that either have an upward 'spin' (i.e., $x_i = 1$) or a downward 'spin' ($x_i = -1$). When the spins align, the iron is magnetic. One way to magnetize a piece of iron is by introducing an external field—i.e., holding a magnet close to the object—that pulls the electrons in a particular direction.

By applying a minimally positive external field, we thus ensure that the nodes in the network have a slight preference towards general ability. Consequently, on average, knowledge is more often obtained than unobtained in the population. And for an individual network, this implies that it is more likely that the knowledge in a particular cluster is all obtained rather than all unobtained. By introducing the external field—using the approach of Cioletti and Vila (2015)—the model extends into,

$$
\begin{aligned}
p(x, \omega) = & \frac{1}{Z_F} \prod_{e \in E_W} \left\{ \vartheta_W \, \delta_{(\omega_e, 1)} \, \delta_{(e)}(x) + (1 - \vartheta_W) \, \delta_{(\omega_e, 0)} \right\} \\
& \times \prod_{e \in E_B} \left\{ \vartheta_B \, \delta_{(\omega_e, 1)} \, \delta_{(e)}(x) + (1 - \vartheta_B) \, \delta_{(\omega_e, 0)} \right\} \\
& \times \prod_{i \in V} \exp\left(\mu \left[\delta_{(x_i, 1)} - \delta_{(x_i, -1)} \right] \right),
\end{aligned}
$$

where μ denotes the external field. In the networks used to create Figures 7.4 and 7.5 we set $\mu = .03$.

Moreover, as is shown in Appendix B, the probability that a cluster consists of obtained knowledge is proportional to the size of the cluster. This means that the larger the cluster, the more likely that the nodes in the cluster reflect obtained knowledge. Importantly, this ensures with high probability that the giant component (the largest cluster) consists of pieces of knowledge that are obtained rather than unobtained. Note that the external field could be negative too, in the rare situation that the environment elicits misconceptions.

7.3.2 Dynamics: *wiring* cognition

The static wired cognition model provides a solid basis for the second aim. We conceptualize intelligence as evolving networks, in which new facts and procedures are wired together during development. In this section, we therefore explore the model from such a developmental point of view. We discuss three scenarios.

Scenario 1: development towards equilibrium

Although we do not know the exact causal mechanisms that drive development, we can observe the model during development. In the first scenario, we started the network in an undeveloped state, with solely unobtained pieces of knowledge and no edges. We then used a Gibbs sampling procedure to grow the network towards its equilibrium state: the extended FK model with its desirable properties. The positive manifold and hierarchical structure displayed in Figure 7.5, discussed in the previous section, necessarily follow from this approach as they are properties of the FK model.

Here, we are concerned with how the model develops over time. As the Gibbs sampler rapidly converge the networks to an equilibrium state, we slowed down the process by updating

the nodes and edges in each iteration with $p = .15$. This way, we aimed to get insight in how the model behaves towards its equilibrium state. In Figure 7.6 we illustrate, for 1000 networks, the growth in the number of obtained pieces of knowledge across the first 30 iterations of the decelerated Gibbs sampler. We considered 96 nodes across four domains, and set the external field to .03, the within-community connectivity to .07, and the between-community connectivity to .005.

Interestingly, the figure provides a clear indication of the Matthew effect. First, Figure 7.6a shows the fan-spread effect that characterizes the Matthew effect. Although all simulated persons were conceived with the exact same cognitive networks, early differences in the number of obtained pieces of knowledge become more pronounced over time, until they stabilize. Figure 7.6b shows that the variance in obtained pieces of knowledge across networks indeed increases and ultimately stabilizes.

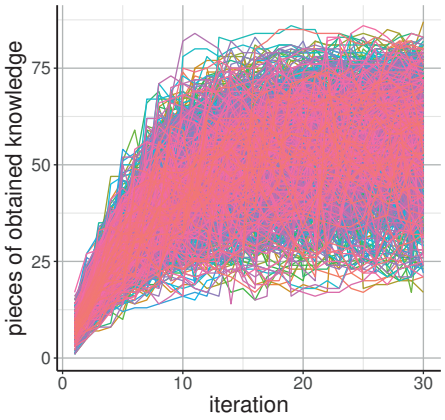
Moreover, Figure 7.6c shows that the variance in obtained pieces of knowledge across subsequent states of individual networks decreases over time, an effect that is also observed in the example of Pólya's urn that is displayed in Figure 7.2a. However, as opposed to the Matthew effect in the Pólya's urn example, the positive external field in the wired cognition model ensures that a general ability prevails.

This scenario, in which a cognitive network grows from an undeveloped state into an equilibrium state, was also used by van der Maas et al. in their description of the mutualism model. A criticism of this approach is that from the onset of development until the moment it reaches its equilibrium state, the exact properties of the network are unknown. In the next scenario, we avoided this issue by inspecting the networks solely in equilibrium, but across different sizes.

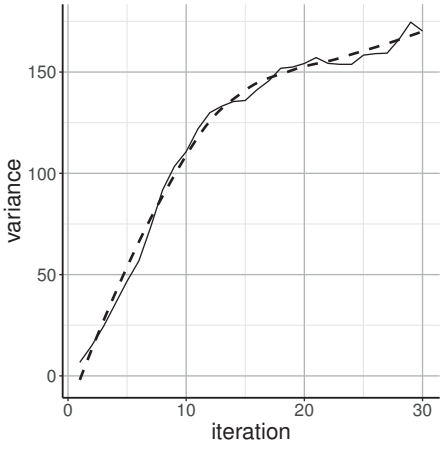
Scenario 2: development in equilibrium

Rather than observing the model during the sampling dynamics, in the second scenario we investigated the development of the model across equilibrium states of differently sized networks. Like—in the previously cited words of Cattell—the movie director that inspects the stills. To this end, we sampled networks ranging in size from 20 to 300 nodes, in steps of four nodes. We set the external field to .005, the within-community connectivity to .07, and the between-community connectivity to .01.

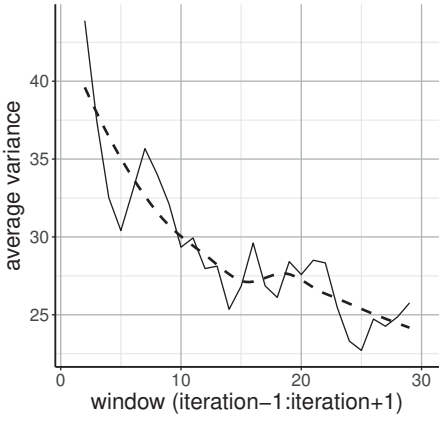
In Figure 7.7 we illustrate the growth in obtained pieces of knowledge across the increasing number of nodes in the respective networks. Naturally, the number of obtained pieces of knowledge cannot be higher than the total number of nodes in a network, hence the fact that



(a) The typical fan-spread effect observed in the Matthew effect. Each line represents the development of the number of obtained pieces of knowledge (y-axis) in a wired cognition network, with a total of 92 nodes in each network.



(b) The variance in the number of obtained pieces of knowledge across networks increases with time, and ultimately seems to stabilize. The dashed line represents the Loess (local regression) curve.



(c) The average variance in the number of obtained pieces of knowledge in a window of three subsequent iterations decreases with time. The dashed line represents the Loess (local regression) curve.

Figure 7.6. The Matthew effect appears when the Gibbs sampler—used to sample a network from the model—is decelerated.

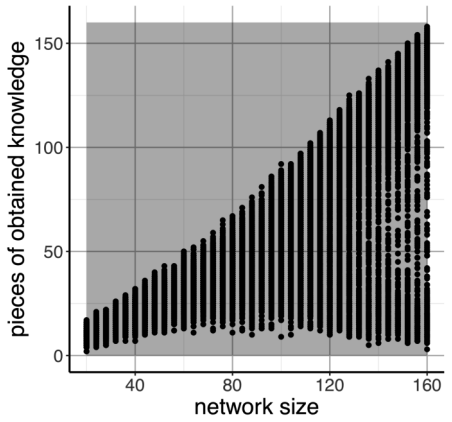
all observations are in the lower right triangle. Figure 7.7a again shows the Matthew effect. This time, all observations are completely independent from one another, hence no lines are shown.

Interestingly, when the networks continue to develop they start to bifurcate. In Figure 7.7b it is shown that this is the case for networks that contain roughly 100 or more nodes. Notably, this pattern bears similarities with the Matthew effect found in science funding (Bol, de Vaan, & van de Rijt, 2018), where a similar divergence is seen for scholars that are just below and just above the funding threshold. In the case of science funding, the effect is partly attributed to a participation effect: scholars just below the funding threshold may stop applying for further funding. In education, this participation effect is institutionalized through stratification: students just below an ability threshold will receive education on a different level.

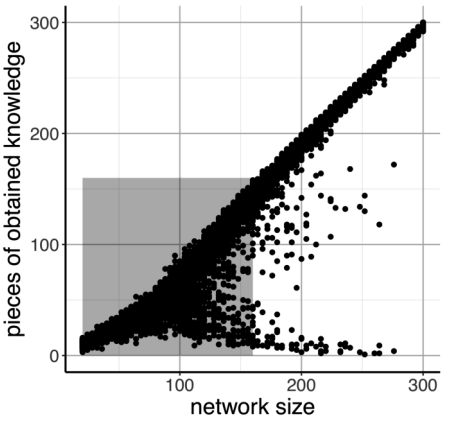
Here, the bifurcation arises from the preference of the FK model to form clusters, and the fact that only nodes that are in the same state can become connected. As soon as all nodes are connected in a giant component, the networks contain either obtained or unobtained pieces of knowledge. Additionally, the positive external field ensures that the larger a cluster, the higher the probability that it contains obtained pieces of knowledge, and hence the large number of observations near the diagonal. Finally, the low connectivity in the simulated networks creates some observations in between the two forks of the bifurcation.

The growth of cognitive networks can also shed a new light on the positive manifold. In Figure 7.8 the positive manifold is shown again, but this time across four different sizes of networks. We considered networks with 40, 80, 120, 160 nodes, and show that the positive manifold steadily increases with more nodes. This property of the model reflects a much discussed phenomenon in intelligence: the age dedifferentiation hypothesis. Dedifferentiation is the gradual increase of the factor *g*, or put differently, the increasingly common structure in intelligence across individuals. The hypothesis states that such dedifferentiation takes place from adulthood to old age. Importantly, age dedifferentiation's antagonist, the age differentiation hypothesis, posits that differentiation takes place from birth to early maturity.

Evidence for these hypotheses is both poor and problematic. Many scholars have tried to summarize the evidence, but all come to the conclusion that the evidence for either of the hypotheses is inconclusive. Methodological problems such as selection effects and measurement bias are commonly mentioned to account for the inconclusive evidence (e.g., van der Maas et al., 2006). In the model proposed here, the strength of the positive manifold is primarily determined by the size of the giant component, which increases with the size of the network.



(a) The typical fan-spread effect observed in the Matthew effect. Thousand networks were simulated for each network size from 20 to 160 nodes, in steps of four nodes.



(b) Bifurcation with increasing network size. Hundred networks were simulated for each network size from 20 to 300 nodes, in steps of four nodes. From 130 nodes onward, the networks start to bifurcate. The networks are either attracted to a fully able or a fully inable state.

Figure 7.7. The Matthew effect and bifurcation in developing networks. Points represent the number of obtained pieces of knowledge (*y*-axis), across differently sized networks (*x*-axis). Each point represents an independent observation. The grey rectangles show the parts of the figures that overlap.

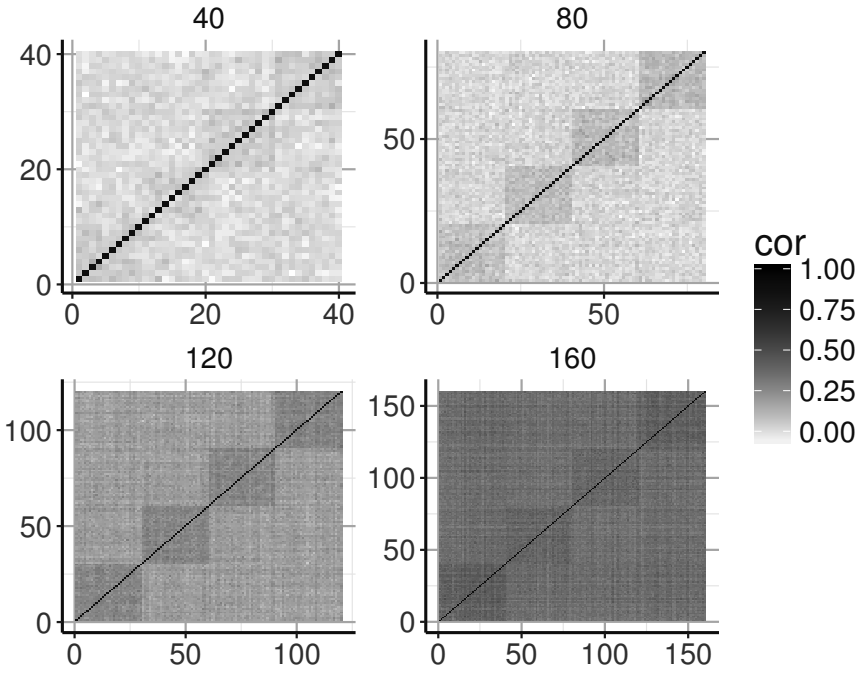


Figure 7.8. Heatmaps of the correlational structure of cognitive networks with 40, 80, 120, or 160 nodes (note the scale of each heatmap), across four different domains. Each heatmap is based on 1000 networks. The exclusively positive patches illustrate the positive manifold and the block structures illustrate the hierarchical structure. Importantly, with networks increasing in size, the positive manifold increases too.

Scenario 3: a growth mechanism

Ultimately, the formulation of a formal developmental theory of intelligence requires the identification of mechanisms of growth, and possibly decline. Although we are unaware of growth mechanisms that keep the FK structure intact, we nonetheless end this section with a third approach. We first describe a simple growth mechanism, and since it may force the network out of equilibrium, we then briefly discuss an additional method that repairs the network in order to ensure that the discussed stationary phenomena remain guaranteed during development.

In this scenario, we conceptualize growth as the addition of previously absent nodes and edges to the cognitive network, where those new nodes may represent obtained as well as unobtained pieces of knowledge[3]. Effectively, in the growth model edges are sampled from the finite set of possible edges that constitute the full network. Nodes connected by a sampled edge—if previously absent—are added to the network. Then, the communities of the nodes connected by the edge are determined, and the nodes are actually connected in the cognitive network with a probability respective to the determined communities. We further explicate this growth mechanism below.

Growth mechanism Let us start with the sampling mechanism for the edges. In our approach, we focus on growing the network *topology*—the wiring of skills and knowledge in a cognitive network—and let the states of the skills follow this process. To do so, we make use of the following factorization of the model

$$f(x, w) = f(x \mid w) f(w),$$

where $f(w)$ is the model for the topology, known as the Random Cluster model (Fortuin & Kasteleyn, 1972; Grimmett, 2006), that describes the wiring of the cognitive network.

This idea can be summarized as follows. Suppose that there is a full theoretical network $G = (V, E)$ that includes all potential skills, knowledge, and their relations. At conception, an individual may start with an empty network, or a small initial subset the network that may represent her or his genetic endowment. As time proceeds, skills, knowledge, and their relations

[3]Formally, two equivalent interpretations may apply. One may interpret absence of a node as a state of nodes that reside *within* the current network, or as nodes that reside *outside* of the current network. In the former sense, growth is interpreted as a change in the state of the node, whereas in the latter sense it is interpreted as the addition of a node not previously present in the network. Having said that, the model is not concerned with this subtle interpretative distinction.

are added to the network, by sampling edges and their attached skills and knowledge. An edge $e = \langle i, j \rangle$ between skills i and j is included in the individual's cognitive network, with simply the previously discussed probability ϑ_W to connect the nodes of the same community, and probability ϑ_B to connect the nodes of two different communities.

This very basic growth model allows for substantial variation across individual networks, thus satisfying the idiography principle. Edges may or may not be sampled, and once sampled, the attached nodes may become connected or disconnected, directly or indirectly via paths. The nodes may become obtained or unobtained pieces of knowledge, and may end up isolated, or connected with nodes of the same or other communities, in small or large clusters. And this may all vary across development.

Restrained freedom Two important remarks must be made with regard to this growth process, one on the considerable amount of freedom in this approach, and one on a self-imposed restriction. First, and importantly, we do not prescribe a sampling model. That is, we conceptualize the sampling mechanism as an empirical fact; a process that can be simulated by ones preferred theoretical model. To give some examples, edges may be added following an educational model (curricula determine the (order of the) sampled edges), using a genetic model (the state of the initial network determines the sampled edges), or for instance reflect the multiplier effect model (both the subsequent states of the network and the environment determine the sampled edges). The model proposed in this article thus provides a unique opportunity to study the effects of such diverse sampling models.

Second, and not easily observed, is the fact that the suggested growth mechanism cannot guarantee that the properties of the static model, such as the positive manifold, will continue to hold. Therefore, in order to keep the model tractable during its growth, we impose a restriction that helps retain those properties. Basically, we repair the network if it is observed to deviate from the static model, by means of re-pairing the most recent set of added nodes. This rewiring is sometimes required when two clusters are joined in the cognitive network. The procedure is described by Fill and Huber (2000).

One way to interpret the rewiring that is part of Fill and Huber's approach, is that it inspires a change in obtained knowledge of a newly joined cluster. That is, unobtained pieces of knowledge could be relatively static on a cluster over time, but might switch states when two clusters are joined, reflecting a new insight. Since the giant component is increasingly likely to represent obtained knowledge, a newly connected component is likely to turn into a component of obtained knowledge. This way, learning occurs gradually in the cognitive network, one skill

at a time, but also through phases, growing obtained knowledge on entire clusters at once.

A second consequence of this effort to retain the properties of the static model in a continuously evolving network, is the fact that in the growth mechanism we do not determine the states of the nodes during development. This means that, although at each point in time the network can be frozen and the states of the nodes determined, subsequent states are independent evaluations of the model. Although from a developmental point of view this might be seen as problematic, the justification is twofold. The first is a feature: small clusters—such as clusters of a single node—represent unconnected knowledge or skills for which instability can be an actual property. Across the independent evaluations, these small clusters may flicker accordingly. On top of that, the positive external field discussed in the static model section ensures that the larger a cluster, the higher the probability that the nodes represent obtained knowledge. This thus ensures that the larger a cluster, the more stable its state.

To conclude, it must be stressed that the discussed restriction—although not necessarily problematic—primarily provides us with a mathematical convenience, rather than that it reflects an empirical fact. And although we believe it is a welcome convenience for this initial suggestion of a growth mechanism, it may as well be abandoned in future suggestions. For now, the growth and repair mechanisms—along with the static model—give us a minimal description of a *wiring* cognition network.

7.4 Discussion

Since Spearman's first attempt to explain the positive manifold, it has been the primary aim for formal theorists of intelligence. Although many scholars followed his factor-analytic footsteps, an approach that is dominant as of today, we now also know that it is only one of many possible explanations. Recent contributions to scholarly intelligence, such as the contemporary mutualism model and multiplier effect model, have greatly aided the field by providing novel explanations of a much-debated construct. In this paper, we took those new directions two steps further by providing another alternative explanation. First, we introduced a truly idiographic model that captures individual differences in great detail. In doing so, it bridges the two disciplines of psychology, by explaining nomothetic phenomena from idiographic network representations. Second, the model provides a formal framework that particularly suits developmental extensions, and thus enables the study of both genetic and environmental influences during the development of intelligence.

The static *wired* cognition model proposes a parsimonious and unified explanation of two

important stationary phenomena: the positive manifold and the hierarchical structure. It does so without a need for mysterious latent entities, and with an opportunity to study individual differences. Indeed, many more—yet less robust—phenomena have been identified in the past century. Although in its current form the model does not aim to explain all of these, it may very well serve as a point of departure for exploring or adding more of intelligence's complexities. Then, the dynamic *wiring* cognition model is a much more modest contribution; a specimen of the potential of developmental mechanisms, and an explicit call for increased inquiry into both developmental mechanisms and phenomena. Nevertheless, it may too serve as a point for departure for subsequent theorizing. Importantly, and self-evident, both parts can be further built upon, by subjecting them to empirical facts.

In this Discussion section, we first explain three modeling principles, and then discuss how the model interprets and explains the positive manifold, hierarchical structure, and developmental effects, from an idiographic perspective. Finally, we illustrate how this approach provides a unique opportunity to relate micro-level phenomena to the macro-level phenomena prominent in intelligence research.

7.4.1 Modeling principles

In building the proposed model, we aimed to follow three important principles. First, a scientific theory should be *formal*. That is, it should be formulated as a mathematical or computational model. The traditional factor models of general intelligence are statistical models of individual differences. They do not specify a (formal) model of intelligence in the individual. In contrast, the multiplier effect model and the mutualism model have been formulated mathematically. The advantages are that these models are precisely defined, predictions can be derived unambiguously, and unexpected and undesirable by-effects of the model can be detected, for instance in simulations.

The second principle is that a theory of intelligence should be *idiographic*. With the network approach we intend to bridge two separate research traditions; on the one hand experimental research on cognitive mechanisms and processes, and on the other hand psychometric research on individual differences in intelligence. Cronbach's (1957) famous division of scientific psychology into these two disciplines is still very true for the fields of cognition and intelligence. In the words of Ferguson (1954), "[t]his divergence between two fields of psychological endeavour has led to a constriction of thought and an experimental fastidiousness inimical to a bold attack on the problem of understanding human behaviour." The model proposed in this article brings

these fields together, by enabling explanations of individual differences from hypothesized cognitive mechanisms.

Similarly important as the previous idea, the third idea is that a theory of intelligence should be *psychological*. This idea is expressed in Box' (1979) famous argument that "all models are wrong but some are useful". Our aim was a model that is indeed "illuminating and useful", by carefully weighing mathematical convenience and psychological plausibility, and by ensuring it creates novel predictions about for instance the structure of intelligence, and the role of education in the shaping of intelligence. As such, the model proposed in this article acknowledges the need for explanatory influences of the environment, education, and development.

7.4.2 A new perspective

The introduced model strongly adheres to these principles, and introduces a novel conception of intelligence. Other than by a unified factor (e.g., *g* models), a measurement problem (e.g., sampling models), or positive interactions (e.g., mutualism model and multiplier effect model), we explain the positive manifold by the wiring of knowledge and skills, or facts and procedures, during development.

Additionally, in the model the hierarchical structure of intelligence has an incredibly straightforward explanation: knowledge and skills that are more related, have a higher probability of becoming connected. This idea follows a simple intuition. If student Cornelius is trying to learn a new word, this word will attach with high probability to related words, and with low probability to distant words. The richer Cornelius' vocabulary, the higher the chance that this new word will stick. In cognitive science, this principle is dealt with in the study of schemata (e.g., Bartlett, 1932; van Kesteren, Rijpkema, Ruiter, Morris, & Fernández, 2014).

Finally, developmental phenomena are an unusual suspect in formal models of intelligence. Although decidedly less straightforward than the discussed stationary phenomena, we do believe that developmental trends—such as the Matthew and compensation effect—must play a key role in the study of developmental intelligence. As argued by Protopapas et al. (2014) and Schroeders et al. (2016), and as illustrated in the example of Pólya's urn, mechanisms that can provide an explanation for both phenomena can turn out to be worthwhile in understanding key drivers of development.

In the proposed model, we show that a Matthew effect may not only spontaneously appear when an undeveloped network is grown towards its equilibrium state, but too if cross-sections of networks in development are inspected. This observation of the Matthew effect is in line

with Schroeders et al.'s hypothesis for non-formalized learning environments, as discussed in the introduction. In the model, a non-formalized learning environment can simply be seen as the natural growth that occurs without an educational intervention, whereas formalized learning would comprise of small beneficial interventions, such as illustrated with Pólya's urn.

A second insight from the model is that the continuous differences observed in the Matthew effect can at some point start to bifurcate into clearly discrete groups. This pattern can be compared to the effect of stratification in education. An even deeper insight of this bifurcation process is that growth mechanisms that satisfy the FK model, must contain a degenerative component, such as a forgetting mechanism. Only then can some networks grow into a less able state. In the next section we discuss such aspects of a growth mechanism.

Finally, what is particularly intriguing about the idiographic network perspective, is that the discussed phenomena are observed by aggregating specific representations of many individuals. Here, we like to echo Jensen's remark that "[t]he psychology of intelligence could, at least in theory, be based on the study of one person". The proposed model overtly distinguishes intelligence from *g*, and it is exactly this fact that makes the idiographic approach such a powerful one. Although the study of the individual goes beyond the scope of the current research, in the following we do give some context.

7.4.3 An idiographic approach

On top of the model's capacity to explain two of the—mainly nomothetic and macroscopic—phenomena in intellingence, its idiographic nature additionally enables straightforward interpretations of microscopic phenomena. Thus, rather than the previously discussed aggregated phenomena that are at the forefront of intelligence research, here we mean the phenomena that reside at the level of the individual. The phenomena that originate from the long traditions of experimental and cognitive psychology. Ebbinghaus' (1913) law of forgetting is one great example that we will first turn to.

Cognitive networks, such as in the proposed model, allow for intuitive mechanisms of forgetting. Both nodes and edges may be forgotten, either randomly or dependent on the time a node resides in the network. Moreover, forgetting mechanisms can easily become more interesting, for instance by taking the degree of a node—the number of edges it is attached to—into account, as it may be viewed to signal robustness. In the proposed model, the external field already creates a naive form of forgetting: the state of a cluster of nodes is evaluated at each iteration, and the probability that the nodes in a cluster represent obtained knowledge is a function of the size

of the cluster. The states of poorly connected nodes may therefore vary considerably, whereas well-connected nodes tend to stay in the same state.

Very much related to forgetting are the well-established testing and spacing effects (e.g., Karpicke & Roediger, 2008). Again, the model allows one to consider these effects within a formalized theory of intelligence. In the proposed growth mechanism, testing and spacing are reflected in the sampling of the edges. Intuitively, testing and spacing help dense the edge structure of a network, which leads to both a more resilient network and larger clusters, and hence more obtained knowledge. In combination with a forgetting mechanism, testing and spacing effects thus help determine the birth and death processes underlying the development of a network.

The network topology itself also allows for interpretations of the individual. The number and size of clusters, or the degree distribution of a network, may for instance resonate the robustness of the cognitive network. Interestingly, isolated pieces of unobtained knowledge may be viewed as slips, whereas clusters of unobtained knowledge may be viewed as structural misconceptions. Importantly, the preference for clustering in the model, combined with the positive external field, creates a very natural dynamic. At first, isolated pieces of unobtained knowledge such as slips are quite probable. The preference for clustering will then cause some of these pieces of unobtained knowledge to evolve into straight misconceptions. Finally, the external field ensures that misconceptions that become too pronounced are repaired.

7.4.4 A developmental approach

In discussing the idiographic approach, we already alluded to developmental mechanisms. Indeed, we believe that to understand intelligence, its development must be understood too. In discussing the model, we used the first two scenarios to get a grip on its developmental predictions. In the second scenario, we followed the development of the networks as they increased in size, yet while they remained in an equilibrium state. Interestingly, this allowed us to derive the developmental predictions of the model, such as the Matthew effect and ultimately the bifurcation, without the need to know the growth mechanism.

Importantly, these predictions hold in case the networks indeed grow within the limits of the equilibrium. Following from this, one evident challenge is to study growth mechanisms that keep the FK properties intact. However, the assumption that development takes place in equilibrium may just as well turn out to be too restrictive. Conveniently, the network approach allows one to consider diverse growth mechanisms and study the effect on development. Having

said that, it may also give one too many degrees of freedom. We solved this dilemma, which is inherently tied to formal modelling approaches, by suggesting a straightforward growth and repair mechanism.

Most importantly though, the true restrictions must come from developmental phenomena. This bring us to an important question: what are the key developmental phenomena in intelligence? As we aimed to reflect in the introduction, this is not an unexplored area. Phenomena such as the Matthew and compensation effect, or the age differentiation and dedifferentation hypotheses, are actively studied. Nevertheless, much confusion still exist about these phenomena. As evidenced in the introduction, contradictory phenomena are being observed, and some argue that the causal mechanisms must be studied in order to come to grips with it.

This brings us to a second question, namely, what are the important developmental mechanisms? The Pólya's urn example convincingly illustrates the explanatory power of a seemingly simplistic growth mechanism. In the field of intelligence, the mutualism and multiplier effect models give other examples of such mechanisms. In addition to that, we sense that approaching the contradictory developmental phenomena with causal mechanisms might indeed turn out to be fruitful. The fact that the same mechanism, under different circumstances, can explain contradictory phenomena, is intriguing and should be further explored. We believe that the fact that the proposed model provides a framework to incorporate such developmental mechanisms is one of its major strengths, and studying developmental phenomena and mechanisms should be a primary concern.

8

Summary & conclusions

8.1 Summary

This book covered various topics in the domains of learning, intelligence, and educational technology. Central to and shared among those topics, is the idiographic approach to the science of education. Whereas in the introduction I discussed the three main themes, in this conclusion I discuss each of the chapters in the light of three approaches we used in this book: experimentation, measurement, and theory building. But before I do so, a brief summary of each of the chapters helps us get the various topics to the forefront of our minds. If it is there already, you can safely skip to the next section.

In Chapter 2, we presented a decade of experience with analyzing and improving the computer adaptive practice environment Math Garden. We presented the methods used to steer and analyze the system in real-time, by means of a scoring rule for accuracy and response time, and an adaptive engine that matches students to problems. Importantly, we determined the quality of fit of the adaptive engine, and explored various sources of misfit, such as violations of the unidimensionality assumption. In synthesizing our experiences, we suggested that learning analytics should actively help pursue the educational objective of interest, which can be achieved

by minimizing the time of the cycle from the actual analytics to the interventions that capitalize on them.

In Chapter 3, we gave a brief overview of the field of online learning in 2014. We discussed its success in attracting learners worldwide, and its failure to deliver on its promise. Importantly, we suggested a way forward: large-scale online randomized controlled field experiments. We argued that such tests should take into account the typical nature of (online) learning, and encourage the use of knowledge from the various learning sciences to identify interventions that promise improved learning. Finally, we identified both limitations and promises of these so-called A/B tests, and showed how they can ultimately contribute to learning that is tailored to each individual learner.

In Chapter 4, we reported on a grapeshot A/B test in Math Garden. The experiment was designed to eliminate an unforeseen opportunity to practice with minimal effort. Some children tended to skip problems that require deliberate effort, and only attempted problems that they could spontaneously answer. Our intervention delayed the option to skip a problem, thereby promoting effortful practice. The results revealed an increase in the exerted effort, without being at the expense of engagement. We could not conclude whether the additional effort positively affected the children's learning gains. Importantly, we additionally introduced and illustrated the holdout principle: a method to increase the reliability of exploratory research.

In Chapter 5, I argued that the opportunities of online learning environments are ample, provided that teachers are equipped with the appropriate tools. Unfortunately, this is not always the case, as customization is often limited to the functionality of the adopted software. I discussed Learning Tools Interoperability as an important exception, as it allows teachers to extend a virtual learning environments (VLEs) with external software, and thus increase their pedagogical range. In the chapter, I introduced a software protocol that exploits Qualtrics, popular software for creating and distributing surveys, to extend native VLE functionality with random assignment for experimental comparisons, a range of additional educational elements, and options for personalizing educational content.

In Chapter 6, we proposed and investigated a model that aims to identify a student's misconceptions from the errors she or he makes. We applied our method to single digit multiplication; a domain that is very suitable for our method, is well-studied, and allowed us to analyze over 25,000 error responses from 335 actual learners. The results show that the model outperforms a random selection from the observed errors' possible causes, and correctly predicts the possible cause of a person's subsequent error up to over 75% of the time.

Finally, in Chapter 7, we introduced a novel model for the development of intelligence, with

strong theoretical implications. In the past century, various formal models for intelligence have been proposed, including the dominant *g*-factor model, the revived sampling theory, and the recent multiplier effect model and mutualism model. We proposed a novel idiographic model and conceptualized intelligence as evolving networks. The static model, an extension of the Fortuin-Kasteleyn model, provides a parsimonious explanation of the positive manifold and intelligence's hierarchical factor structure. On top of that, we showed how it explains the Matthew effect across developmental stages, and we suggested a method for studying growth dynamics.

8.2 The scientific trident

The topics in this book encompass the scientific trident: experimentation, measurement, and theory building. In the following I discuss the chapters in this book in the light of these three approaches, and in doing so I aim to illustrate their interconnectedness.

Experimentation A major topic in this book is online experimentation. Across multiple chapters (2, 3, and 4), various challenges and opportunities were discussed. I belief one observation is very crucial for understanding the role of A/B tests in online learning: the fact that relatively few A/B tests in the domain of online learning are reported in the scientific literature, especially considering its omnipresence in large internet companies. Several possible reasons explain this gap. For one, due to their uncontrolled nature, conducting A/B tests reliably requires a tremendous amount of users (Kohavi et al., 2014). Although successful providers of online learning do reach such scale, it is not easily achieved for the majority of the providers.

A second reason are the additional demands on the providers of online learning. Not only do they most likely need to accommodate a tremendous growth in a short period of time, the payoffs of A/B tests are neither directly obvious. Whereas for commercial websites like Google and Booking.com A/B tests have a clear manifest payoff (e.g., revenue), return on investment with respect to learning outcomes is diffuse and difficult to quantify. Moreover, whereas those commercial websites may find effects of seemingly trivial interventions (e.g., changing font color, Kohavi et al., 2014), interventions that improve learning with respect to invested time are more complex.

Finally, many of these providers are commercial companies, complicating direct involvement of researchers or lacking a strong incentive to publish results. A welcome exception are Massive Open Online Courses (MOOCs), expressed in its active research community (Kizilcec & Brooks, 2017; Reich, 2015). Outside the realm of MOOCs, the few providers known to run A/B

tests (e.g., crowd-sourced A/B-tests in the intelligent tutoring system ASSISTments, Heffernan & Heffernan, 2014) clearly illustrate the remaining vacuum.

There is no quick fix. First and foremost, a mutual understanding of the symbiotic relation between providers and researchers of online education is key. The former provide the scale and infrastructure, whereas the latter must ensure that this scale translates into actual impact. Math Garden, which was discussed in three chapters, is exemplary as it not only provides serious scale, but moreover has a serious track record of scientific reports based on their system. Finally, evaluation of effectiveness is crucial, but is possibly also one of the most challenging missions of education research. As discussed in Chapter 4, the model used for clinical treatments—clinical trials—arguably fails, and thus other objective methods need to be devised to guarantee effectiveness.

Presumably, the future of experiments in online learning follows the development of A/B tests in internet companies. Naturally, Bayesian approaches allow one to continuously accumulate and assess the evidence for a particular intervention, rather than having to use a predefined period and a single chance of analysis. Additionally, the increasingly popular multi-armed bandit approach enables one to adaptively change assignment probabilities to various interventions on the basis of real-time evaluations of their effectiveness. Although promising, Rafferty, Ying, and Williams (2018) show that the already large amount of participants required in A/B tests, in some cases needs to doubled in order to retain sufficient power.

Measurement Multi-armed bandits bring us to the second topic: measurement. Evidently, in order to assess an intervention, an outcome must be evaluated. In Chapter 6 we proposed a method for identifying the misconceptions that cause a particular student's errors on a task. In the context of multi-armed bandits, such measurements can be used to evaluate the effect of interventions targeted at treating the misconceptions. Additionally, these bandit algorithms can be easily extended to take the identified misconception into account when determining the next intervention, which creates de facto personalization (e.g., Segal, David, Williams, Gal, & Shalom, 2018). Finally, they are an elegant opportunity to move away from the grapeshot method that was illustrated in Chapter 4.

Math Garden's computer adaptive practice algorithm, reviewed in Chapter 2, beautifully illustrates the challenges with not only real-time measurement, but also with the assumption of unidimensional ability. Problematic tasks and incorrect assumptions of how the software is used can cause the ability estimates to be inaccurate, and individual differences in misconceptions and response speed invalidate the strong assumption of unidimensionality. Interest-

ingly, a diagnostic approach aimed at treating particular misconceptions, such as discussed in Chapter 6, can counteract such a violation of the unidimensionality assumption. Indeed, just as formal education is hypothesized to shift the Matthew effect into a compensation effect, it too promotes unidimensional ability. This is beautifully reflected in Leo Tolstoy's popularized phrase: "Happy families are all alike; every unhappy family is unhappy in its own way."

The fact that educational and psychological measurement affect the studied processes (e.g., Kuhn, 1995) provides another challenge. Although in a lab setting these effects are often thought to be small and tend to be ignored, it is easily seen that this viewpoint is not tenable for in vivo learning. Many feedback loops are encountered: students, parents, teachers, and policy makers can act on or interact with the educational system. Moreover, the active analytics advocated in Chapter 2 explicitly call for direct feedback loops from analytics to educational interventions. Ultimately, fundamental theory is required to guide such measurement and experimentation. And although challenging, such theory must convincingly deal with the actual complex dynamics of in vivo learning; a topic I discuss next.

Theory building Finally, it is a theory of cognitive ability that this book ended with. Readers that were overwhelmed by the various theories and models discussed in Chapter 7, and unfamiliar with formal theory formation, may benefit from the following metaphor:

> Explaining psychological phenomena is much like exploring an unknown cave. When Spearman (1904) aimed to explain the positive manifold—a prominent phenomenon in intelligence—he suggested *g* theory. Where the cave may represent scientific intelligence, *g* theory can be viewed as the first room that was discovered in the cave of intelligence. Now, a century later, every inch of this room has been meticulously inspected, and is filled with fluorescent light. However, unlike caves, theories are not set in stone. As such, some scholars have dared to go deeper into the cave, discovering arguably more spectacular rooms. In Chapter 7, we guided you through the rooms that have been discovered thus far, but more importantly, we take you on an expedition to a room that was previously undiscovered. And although at this stage we only have the light of our torch, its reflections are very promising.

This theory, as I argued in Chapter 1, provides an elegant formal connection between human ability and human learning. It was achieved by scaling the theory of intelligence to a new level of detail, the general idea of a process called renormalization. This process showed us that under-

standing the positive manifold on the one scale does not necessarily translate to understanding it on another, and that causal interpretations—such as in the factor modeling tradition—must be met with utmost vigilance. Indeed, correlation does not imply causation, a fact that is sometimes poorly understood in intelligence research (as evidenced by the habit of drawing arrows in graphical representations of statistical factor models, as opposed to the undirected edges in factor graphs from the field of probabilistic graph models). Very much related is the apparent statistical equivalence of models with very dissimilar implications, discussed in Chapter 7. The fact that (experimental) interventions are a primary means to distinguish these models, closes the circle in this section.

Although the model we introduce is by no means complete—as if such models exist—it provides many clues on where to direct our efforts next. Many clues can be found in a buzzing field called complexity science. This interdisciplinary patchwork field is kept together by a shared objective: explaining macroscopic phenomena from descriptions of microscopic interactions. Indeed, this is exactly our approach to modeling intelligence, and one that on this very scale (inter-individual phenomena and intra-individual interactions) can be called idiographic. Moreover, it is no coincidence that networks are being used as a primary means to model complex systems.

Various properties of the models we discussed in relation to our approach to intelligence, signal that explaining cognitive ability from a complex systems perspective is everything but futile. We showed that the global positive manifold may *emerge* from local interactions, we saw that in our current model intelligence becomes *self-organized* by means of bifurcation, key to both the multiplier effect approach and mutualism approach are *feedback loops*, and Pólya urn trees are highly *self-similar* in their root structure. Importantly, these are properties that either define, or are regularly found in, complex systems across many disciplines.

More clues can be found in developmental biology (indeed a primary reason for the plant metaphor in the opening of this book). I belief that in theories of psychological development, biological systems are highly overlooked. Both organisms, plants and humans, share many important attributes, such as their growth, decline, variation, and clear genetic and environmental influences. Contrary to psychology however, as I discussed in Chapter 1, developmental biology benefits from the fact that it is much easier to control. And unsurprisingly, it has much experience with modeling development.

One example is the cellular Potts model (CPM, Graner & Glazier, 1992). Interestingly, the Potts model is a generalization of the Ising model that we discussed in Chapter 7. In developmental biology, CPMs are used to model various developmental mechanisms and to explain a

multitude of phenomena. For instance, Grieneisen, Xu, Marée, Hogeweg, and Scheres (2007) use it for root growth of plants, Li and Lowengrub (2014) for the growth of tumor cell clusters, and N. Chen, Glazier, Izaguirre, and Alber (2007) for morphogenesis. The young field of developmental intelligence may greatly benefit from the techniques and mechanisms used in these and other applications (Prusinkiewicz & Runions, 2012, give an overview of computational models of plant development).

8.3 Science and society

Whereas in this book topics like educational technology and learning analytics are discussed from a primarily scientific perspective, these technologies too have an undeniable societal impact. And, in any domain that is affected by technology, a disruption is not neutral and has both pros and cons, proponents and opponents, confidences and concerns. Although this book is not the place to discuss these in detail, I do think it is appropriate to briefly discuss three ideas aimed at making sure society benefits most.

First, for society to adequately adopt educational technologies, the technologies must be trusted, and to be trusted, they must be understood. A major asset of two technologies discussed in this book, the computer adaptive practice environment Math Garden and the diagnostic model discussed in Chapter 6, is that they are intuitively understood. In the former, students compete with items, and when the student wins she or he is paired with a slightly more difficult competitor. In the latter, a straightforward function of the frequency that a certain misconception could have been the cause of a set of observed errors determines the estimated probability that the student has that misconception.

Machine learning, in this book a bit of an elephant in the room, does not benefit from that characteristic. Although it is a hugely popular family of techniques for use with big data, it suffers from a decisive trade-off. Its superiority with respect to prediction comes at the cost of its explanatory value. I belief one should be wary with decisions based on machine learning techniques that have considerable impact.

Second, learning analytics needs validation mechanisms. This is an important, but surely not original idea. In jurisprudence, a single source of evidence is hardly ever accepted. In high-quality journalism, a single source of information detains publication. Likewise, if decisions based on learning analytics have significant consequences, the analytics cannot stand on their own. This is especially true for the black-box approaches like machine learning, but too for white-box approaches discussed in this book.

Third, educational technology demands fundamental research, such as presented in this book. As I mentioned in Chapter 5, a future with educational technology is unstoppable and, what is more, desirable. But if we want students and teachers to retain their central role in education, we cannot sit back and relax. The rise of educational technology must be accompanied with fundamental research on its desired role in education, on the aspects that optimally benefit students and teachers, and on the aspects that do not. If done right, educational technology can help strengthen the role of students and teachers.

Bibliography

Ackerman, P. L. & Lohman, D. F. (2003). Education and g. In *The scientific study of general intelligence* (pp. 275–292). Elsevier BV. doi:10.1016/b978-008043793-4/50052-0

Aleven, V., Sewall, J., Popescu, O., Ringenberg, M., van Velsen, M., & Demi, S. (2016). Embedding intelligent tutoring systems in moocs and e-learning platforms. In *Intelligent tutoring systems* (pp. 409–415). Springer Nature. doi:10.1007/978-3-319-39583-8_49

Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, D., Baker, R., ... Gasevic, D. (2015). The beginning of a beautiful friendship? intelligent tutoring systems and moocs. In *Lecture notes in computer science* (pp. 525–528). Springer Nature. doi:10.1007/978-3-319-19773-9_53

Allen, I. E. & Seaman, J. (2014, January). Grade change: Tracking online education in the united states. Retrieved July 8, 2014, from http://onlinelearningconsortium.org/publications/survey/grade-change-2013

Arnold, B. C. (1967). A generalized urn scheme for simple learning with a continuum of responses. *Journal of Mathematical Psychology*, *4*(2), 301–315. doi:10.1016/0022-2496(67)90054-5

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review*, *2*(3), 213–236. doi:10.1016/0273-2297(82)90012-0

Athey, S. & Imbens, G. (2016, July 3). The state of applied econometrics - causality and policy evaluation. *arXiv*. arXiv: 1607.00699v1 [stat.ME]

Athreya, K. B. & Karlin, S. (1968). Embedding of urn schemes into continuous time markov branching processes and related limit theorems. *The Annals of Mathematical Statistics*, *39*(6), 1801–1817. Retrieved from http://www.jstor.org/stable/2239282

Aud, S., Hussar, W., Johnson, F., Kena, G., Roth, E., Manning, E., ... Notter, L. (2012, May 24). The condition of education 2012. Retrieved July 8, 2014, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012045

Bagchi, A. & Pal, A. K. (1985). Asymptotic normality in the generalized polya–eggenberger urn model, with an application to computer data structures. *SIAM Journal on Algebraic Discrete Methods*, *6*(3), 394–405. doi:10.1137/0606041

Baggaley, J. (2013). Mooc rampant. *Distance Education*, *34*(3), 368–378. doi:10.1080/01587919.2013.835768

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), 612–637. doi:10.1037/0033-2909.128.4.612

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2014). Qrtengine: An easy solution for running online reaction time experiments using qualtrics. *Behavior Research Methods*, *47*(4), 918–929. doi:10.3758/s13428-014-0530-7

Bartholomew, D. J. (2004). *Measuring intelligence*. Cambridge University Press. doi:10.1017/cbo9780511490019

Bartholomew, D. J., Allerhand, M., & Deary, I. J. (2013). Measuring mental capacity: Thomsons bonds model and Spearmans g-model compared. *Intelligence*, *41*(4), 222–233. doi:10.1016/j.intell.2013.03.007

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

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press. Retrieved from http://www.mpi.nl/publications/escidoc-2273030

Bast, J. & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology*, *34*(6), 1373–1399. doi:10.1037/0012-1649.34.6.1373

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. doi:10.1016/0022-2496(79)90004-x

Batchelder, W. H., Bershad, N. J., & Simpson, R. S. (1992). Dynamic paired-comparison scaling. *Journal of Mathematical Psychology*, *36*(2), 185–212. doi:10.1016/0022-2496(92)90036-7

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). Mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, *32*(6). doi:10.18637/jss.v032.i06

Ben-Zeev, T. (1995). The nature and origin of rational errors in arithmetic thinking: Induction from examples and prior knowledge. *Cognitive Science*, *19*(3), 341–376. doi:10.1207/s15516709cog1903_3

Ben-Zeev, T. (1998). Rational errors and the mathematical mind. *Review of General Psychology*, *2*(4), 366–383. doi:10.1037/1089-2680.2.4.366

Bhatnagar, S., Lasry, N., Desmarais, M., & Charles, E. (2016). Dalite: Asynchronous peer instruction for moocs. In *Adaptive and adaptable learning* (pp. 505–508). Springer Nature. doi:10.1007/978-3-319-45153-4_50

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*(1), 417–444. doi:10.1146/annurev-psych-113011-143823

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. doi:10.3102/0013189x013006004

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, *115*(19), 4887–4890. doi:10.1073/pnas.1719557115

Borghans, L., de Wolf, I., & Schils, T. (2016). Experimentalism in dutch education policy. In T. Burns & F. Köster (Eds.), *Governing education in a complex world*. OECD Publishing. doi:10.1787/9789264255364-en

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. doi:10.1002/wps.20375

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In *Dynamic process methodology in the social and developmental sciences* (pp. 67–97). Springer US. doi:10.1007/978-0-387-95922-1_4

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics*. Academic Press.

Bradshaw, L. & Templin, J. (2013). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403–425. doi:10.1007/s11336-013-9350-4

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L. J., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition*, *149*, 40–55. doi:10.1016/j.cognition.2016.01.004

Brinkhuis, M. J. S. (2014). *Tracking educational progress* (Doctoral dissertation, University of Amsterdam). Retrieved from http://hdl.handle.net/11245/1.433219

Brinkhuis, M. J. S., Bakker, M., & Maris, G. (2015). Filtering data for detecting differential development. *Journal of Educational Measurement*, *52*(3), 319–338. doi:10.1111/jedm.12078

Brinkhuis, M. J. S. & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems* (Measurement and Research Department Reports No. 09-01). Cito. Arnhem. Retrieved from https://www.researchgate.net/publication/242357963

Brinkhuis, M. J. S. & Maris, G. (2010). *Adaptive estimation: How to hit a moving target* (Measurement and Research Department Reports No. 10-01). Cito. Arnhem. Retrieved from http : / / www . cito . nl / onderzoek % 20en % 20wetenschap / achtergrondinformatie / publicaties/measurement_reports

Brinkhuis, M. J. S., Savi, 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. doi:10.17605/osf.io/g4z85

Brown, J. S. & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, *2*(2), 155–192. doi:10.1207/s15516709cog0202_4

Brown, J. S. & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, *4*(4), 379–426. doi:10.1207/s15516709cog0404_3

Brown, R. E. (2016). Hebb and Cattell: The genesis of the theory of fluid and crystallized intelligence. *Frontiers in Human Neuroscience*, *10*. doi:10.3389/fnhum.2016.00606

Buwalda, T., Borst, J., van der Maas, H. L. J., & Taatgen, N. (2016). Explaining mistakes in single digit multiplication: A cognitive model. In D. Reitter & F. E. Ritter (Eds.), *Proceedings of the 14th international conference on cognitive modeling*: University Park, PA: Penn State. Retrieved from http://acs.ist.psu.edu/iccm2016/proceedings/buwalda2016iccm.pdf

Carroll, J. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. doi:10.1037/h0046743

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Elsevier Science.

Ceci, S. J., Barnett, S. M., & Kanaya, T. (2003). Developing childhood proclivities into adult competencies: The overlooked multiplier effect. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 70–92). Cambridge University Press. doi:10.1017/cbo9780511615801.005

Chafkin, M. (2013, November 14). Udacity's sebastian thrun, godfather of free online education, changes course. Retrieved July 11, 2014, from http://www.fastcompany.com/3021473/udacity-sebastian-thrun-uphill-climb

Chen, M.-R. & Wei, C.-Z. (2005). A new urn model. *Journal of Applied Probability*, *42*(04), 964–976. doi:10.1017/s0021900200001030

Chen, N., Glazier, J. A., Izaguirre, J. A., & Alber, M. S. (2007). A parallel implementation of the cellular potts model for simulation of cell-based morphogenesis. *Computer Physics Communications*, *176*(11-12), 670–681. doi:10.1016/j.cpc.2007.03.007

Cheng, H. F., Yu, B., Park, Y. H., & Zhu, H. (2017, April 2). Projectlens: Supporting project-based collaborative learning on moocs. Retrieved April 2, 2017, from http://www-users.cs.umn.edu/~bowen/Publications/lsw134-chengA.pdf

Christian, B. (2012, April 25). The a/b test: Inside the technology that's changing the rules of business. Retrieved July 8, 2014, from https://www.wired.com/2012/04/ff_abtesting

Cioletti, L. & Vila, R. (2015). Graphical representations for Ising and Potts models in general external fields. *Journal of Statistical Physics*, *162*(1), 81–122. doi:10.1007/s10955-015-1396-5

Content Experiments. (n.d.). Retrieved January 19, 2018, from https://edx-open-learning-xml.readthedocs.io/en/latest/content-experiments/index.html

Conway, A. R. A. & Kovacs, K. (2015). New and emerging models of human intelligence. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(5), 419–426. doi:10.1002/wcs.1356

Coomans, F., Hofman, A. D., Brinkhuis, M., van der Maas, H. L. J., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy - response time data. *PLOS ONE*, *11*(5), 1–19. doi:10.1371/journal.pone.0155149

Coughlan, S. (2014, June 12). School appoints 'head of research'. Retrieved July 8, 2014, from http://www.bbc.com/news/education-27803949

Coursera. (2014a, May 17). Government of trinidad and tobago works with coursera to boost education and improve career skills nationwide. Retrieved July 11, 2014, from http://blog.coursera.org/post/87056905797/government-of-trinidad-and-tobago-works-with-coursera

Coursera. (2014b, June 17). Singapore government builds training program around johns hopkins university data science specialization. Retrieved July 11, 2014, from http://blog.coursera.org/post/89063045827/singapore-government-builds-training-program-around

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi:10.1007/bf02310555

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. doi:10.1037/h0043943

Dalege, J., Borsboom, D., van Harreveld, F., Waldorp, L. J., & van der Maas, H. L. J. (2017). Network structure explains the impact of attitudes on voting decisions. *Scientific Reports*, *7*(1). doi:10.1038/s41598-017-05048-y

Deaton, A. & Cartwright, N. (2016). *Understanding and misunderstanding randomized controlled trials*. National Bureau of Economic Research. National Bureau of Economic Research. doi:10.3386/w22595

Dehaene, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, *284*(5416), 970–974. doi:10.1126/science.284.5416.970

Dickens, W. T. (2007). *What is g?* Retrieved from https://www.brookings.edu/research/what-is-g/

Dickens, W. T. & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*(2), 346–369. doi:10.1037/0033-295x.108.2.346

Dickens, W. T. & Flynn, J. R. (2002). The IQ paradox is still resolved: Reply to Loehlin (2002) and Rowe and Rodgers (2002). *Psychological Review*, *109*(4), 764–771. doi:10.1037/0033-295x.109.4.764

Duijn, P. A. C., Kashirin, V., & Sloot, P. M. A. (2014). The relative ineffectiveness of criminal network disruption. *Scientific Reports*, *4*(1). doi:10.1038/srep04238

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, *14*(1), 4–58. doi:10.1177/1529100612453266

Ebbinghaus, H. (1913). *Memory; a contribution to experimental psychology*. Teachers college, Columbia university. Retrieved from https://archive.org/details/memorycontributioebbiuoft

Eggenberger, F. & Pólya, G. (1923). Über die statistik verketteter vorgänge. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, *3*(4), 279–289. doi:10.1002/zamm.19230030407

Elo, A. E. (1978). *The rating of chess players, past and present*. London: B. T. Batsford, Ltd.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4). doi:10.18637/jss.v048.i04

Epskamp, S., Maris, G. K. J., Waldorp, L. J., & Borsboom, D. (2016, September 9). Network psychometrics. arXiv: 1609.02818v1 [stat.ME]

Ferguson, G. A. (1954). On learning and human ability. *Canadian Journal of Psychology/Revue Anadienne de Psychologie*, *8*(2), 95–112. doi:10.1037/h0083598

Fill, J. & Huber, M. (2000). The randomness recycler: A new technique for perfect sampling. In *Proceedings 41st annual symposium on foundations of computer science*. IEEE Comput. Soc. doi:10.1109/sfcs.2000.892138

Fodor, J. A. (1983). *The modularity of mind*. MIT PR.

Fontenla, J., Perez, R., & Caeiro, M. (2011). Using ims basic lti to integrate games in lmss. In *2011 ieee global engineering education conference (educon)*. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/educon.2011.5773152

Fortuin, C. & Kasteleyn, P. (1972). On the random-cluster model. *Physica*, *57*(4), 536–564. doi:10.1016/0031-8914(72)90045-6

Fox, P. (2014, March 3). A/b testing curriculum: To sneak peek or not? Retrieved July 8, 2014, from http://cs-blog.khanacademy.org/2014/03/ab-testing-curriculum-to-sneak-peek-or.html

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. doi:10.1126/science.1255484

Freire, M., del Blanco, A., & Fernandez-Manjon, B. (2014). Serious games as edx mooc activities. In *2014 ieee global engineering education conference (educon)*. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/educon.2014.6826198

Freund, J., Brandmaier, A. M., Lewejohann, L., Kirste, I., Kritzler, M., Kruger, A., ... Kempermann, G. (2013). Emergence of individuality in genetically identical mice. *Science*, *340*(6133), 756–759. doi:10.1126/science.1235294

Friedman, B. (1949). A simple urn model. *Communications on Pure and Applied Mathematics*, *2*(1), 59–70. doi:10.1002/cpa.3160020103

Gamage, D., Fernando, S., & Perera, I. (2016). To mooc or not to mooc, that is the problem. In *Advances in educational technologies and instructional design* (pp. 131–148). IGI Global. doi:10.4018/978-1-5225-0466-5.ch007

Ginder, S. & Stearns, C. (2014, June 2). Enrollment in distance education courses, by state: Fall 2012. Retrieved July 8, 2014, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014023

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. doi:10.1111/1467-9876.00159

Glickman, M. E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, *28*(6), 673–689. doi:10.1080/02664760120059219

Goel, V. (2014, July 2). After uproar, european regulators question facebook on psychological testing. Retrieved September 10, 2014, from http://bits.blogs.nytimes.com/2014/07/02/facebooks-secret-manipulation-of-user-emotions-under-british-inquiry/

Graner, F. & Glazier, J. A. (1992). Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical Review Letters*, *69*(13), 2013–2016. doi:10.1103/physrevlett.69.2013

Greenberg, B., Medlock, L., & Stephens, D. (2011, December 6). Lessons learned from a blended learning pilot. Retrieved June 25, 2014, from http://www.blendmylearning.com/2011/12/06/white-paper/

Greenwood, M. & Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, *83*(2), 255. doi:10.2307/2341080

Grieneisen, V. A., Xu, J., Marée, A. F. M., Hogeweg, P., & Scheres, B. (2007). Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature*, *449*(7165), 1008–1013. doi:10.1038/nature06215

Grimmett, G. (2006). *The random-cluster model*. Springer Berlin Heidelberg. doi:10.1007/978-3-540-32891-9

Groeneveld, C. M. (2014). Implementation of an adaptive training and tracking game in statistics teaching. In *Computer assisted assessment. research into e-assessment* (pp. 53–58). Springer International Publishing. doi:10.1007/978-3-319-08657-6_5

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement*, *48*(1), 1–4. doi:10.1177/0013164488048001O2

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301–321. doi:10.1111/j.1745-3984.1989.tb00336.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Measurement Methods for the Social Sciences. Newbury Park, CA: Sage.

Hebb, D. O. (1949). *The organization of behavior*. John Wiley & Sons Inc. Retrieved from https://archive.org/details/organizationofbeoohebbrich

Heckman, J. J. (1981). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in labor markets* (pp. 91–140). University of Chicago Press. Retrieved from https://www.nber.org/chapters/c8909

Heffernan, N. T. & Heffernan, C. L. (2014). The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. doi:10.1007/s40593-014-0024-x

Henrick, G. (2012). Moodle as a the central hub of learning with tools plugged in-learning tool interoperability. Retrieved April 2, 2017, from http://research.moodle.net/id/eprint/39

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*(10), e0136449. doi:10.1371/journal.pone.0136449

Hofman, A. D., Visser, I., Jansen, B., Marsman, M., & van der Maas, H. L. J. (2017). Fast and slow strategies in multiplication. doi:10.17605/osf.io/aw3qq

Hoppe, F. M. (1984). Pólya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*, *20*(1), 91–94. doi:10.1007/bf00275863

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. doi:10.1037/h0023816

Hu, D. (2011, November 2). How khan academy is using machine learning to assess student mastery. Retrieved July 8, 2014, from http://david-hu.com/2011/11/02/how-khan-academy-is-using-machine-learning-to-assess-student-mastery.html

ICDE. (2013, December 18). An international outlook on distance education: Global online higher education report, globaloher. Retrieved August 18, 2014, from http://www.icde.org/filestore/News/2013_July-Dec/GlobalOHERinanutshell.pdf

Inhelder, B. & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence* (A. Parsons & S. Milgram, Trans.). Basic Books, New York, NY.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, *31*(1), 253–258. doi:10.1007/bf02980577

Jaggars, S. S., Edgecombe, N., & Stacey, G. W. (2013, April). What we know about online course outcomes. Retrieved July 11, 2014, from http://ccrc.tc.columbia.edu/publications/what-we-know-online-course-outcomes.html

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. doi:10.1016/j.lindif.2016.08.027

Jansen, B. R. J., Louwerse, J., Straatemeier, M., der Ven, S. H. G. V., Klinkenberg, S., & der Maas, H. L. J. V. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, *24*, 190–197. doi:10.1016/j.lindif.2012.12.014

Jensen, A. R. (2002). Psychometric g: Definition and substantiation. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* Mahwah, NJ: Lawrence Erlbaum Associates.

Johnson, H. & Mejia, M. C. (2014, May). Online learning and student outcomes in community colleges. Retrieved July 11, 2014, from http://www.ppic.org/main/publication.asp?i=1096

Kac, M. (1968). Statistical physics: Phase transitions and superfluidity, vol. 1, brandeis university summer institute in theoretical physics. In M. Chrétien, E. Gross, & S. Deser (Eds.), (Chap. Mathematical Mechanisms of Phase Transitions, pp. 241–305). New York: Gordon and Brache Science Publishers.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31. doi:10.1207/s15326985ep3801_4

Kan, K.-J., Ploeger, A., Raijmakers, M. E. J., Dolan, C. V., & van der Maas, H. L. J. (2010). Nonlinear epigenetic variance: Review and simulations. *Developmental Science*, *13*(1), 11–27. doi:10.1111/j.1467-7687.2009.00858.x

Karpicke, J. D. & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. doi:10.1126/science.1152408

Kena, G., Aud, S., Johnson, F., Wang, X., Zhang, J., Rathbun, A., ... Rosario, V. (2014, May 28). The condition of education 2014. Retrieved July 8, 2014, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014083

Kizilcec, R. F. & Brooks, C. (2017). Diverse big data and randomized field experiments in moocs. In C. Lang, G. Siemens, A. F. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 211–222). Society for Learning Analytics Research (SoLAR). doi:10.18608/hla17.018

Klinkenberg, S. (2014). High speed high stakes scoring rule. In *Computer assisted assessment. research into e-assessment* (pp. 114–126). Springer International Publishing. doi:10.1007/978-3-319-08657-6_11

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

Koedinger, K. R., Kim, J., Jia, J. Z., McLaughlin, E. A., & Bier, N. L. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a mooc. In *Proceedings of the second (2015) acm conference on learning @ scale - l@s 15*. Association for Computing Machinery (ACM). doi:10.1145/2724660.2724681

Kohavi, R., Deng, A., Longbotham, R., & Xu, Y. (2014). Seven rules of thumb for web site experimenters. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining - kdd 14*. Association for Computing Machinery (ACM). doi:10.1145/2623330.2623341

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2008). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, *18*(1), 140–181. doi:10.1007/s10618-008-0114-1

Kolowich, S. (2014, January 16). Exactly how many students take online courses? Retrieved June 5, 2014, from http://chronicle.com/blogs/wiredcampus/exactly-how-many-students-take-online-courses/49455

Kovacs, K. & Conway, A. R. A. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151–177. doi:10.1080/1047840x.2016.1153946

Kroeze, R., van der Veen, D. C., Servaas, M. N., Bastiaansen, J. A., Voshaar, R. C. O. V., Borsboom, D., ... Riese, H. (2017). Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study. *Journal for Person-Oriented Research*, *3*(1), 1–11. doi:10.17505/jpor.2017.01

Kruis, J. & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, *6*(1). doi:10.1038/srep34175

Kuhn, D. (1995). Microgenetic study of change: What has it told us? *Psychological Science*, *6*(3), 133–139. doi:10.1111/j.1467-9280.1995.tb00322.x

Kuo, B.-C., Chen, C.-H., & de la Torre, J. (2017). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement*, *42*(3), 179–191. doi:10.1177/0146621617722791

Kuo, B.-C., Chen, C.-H., Yang, C.-W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology*, *36*(6), 1115–1133. doi:10.1080/01443410.2016.1166176

Lack, K. A. (2013, March 21). Current status of research on online learning in postsecondary education. Retrieved September 18, 2014, from http://www.sr.ithaka.org/sites/default/files/reports/ithaka-sr-online-learning-postsecondary-education-may2012.pdf

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. doi:10.1126/science.1248506

Lee, J. J., ad Robbee Wedow, Okbay, A., Kong, E., Maghzian, O., Zacher, M., ... Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*. doi:10.1038/s41588-018-0147-3

Lewin, T. (2012, July 17). Consortium of colleges takes online education to new level. Retrieved July 15, 2014, from http://www.nytimes.com/2012/07/17/education/consortium-of-colleges-takes-online-education-to-new-level.html

Li, J. F. & Lowengrub, J. (2014). The effects of cell compressibility, motility and contact inhibition on the growth of tumor cell clusters using the cellular potts model. *Journal of Theoretical Biology*, *343*, 79–91. doi:10.1016/j.jtbi.2013.10.008

Long, P. & Siemens, G. (2011, September 12). Penetrating the fog: Analytics in learning and education. Retrieved July 8, 2014, from https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company, Inc. Retrieved from https://archive.org/details/in.ernet.dli.2015.139135

Lu, L. & Liu, C. (2014). Separation strategies for three pitfalls in a/b testing. Retrieved July 8, 2014, from http://www.ueo-workshop.com/wp-content/uploads/2014/04/Separation-strategies-for-three-pitfalls-in-AB-testing_withacknowledgments.pdf

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Dover Publications. doi:10.1037/14396-000

Mahmoud, H. (2008). *Polya urn models*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Makel, M. C. & Plucker, J. A. (2014). Facts are more important than novelty. *Educational Researcher*, *43*(6), 304–316. doi:10.3102/0013189x14545513

Maris, G. & van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. doi:10.1007/s11336-012-9288-y

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., ... Maris, G. K. J. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research*, *53*(1), 15–35. doi:10.1080/00273171.2017.1379379

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports*, *5*(1). doi:10.1038/srep09050

McGrew, K. & Flanagan, D. (1998). *Intelligence test desk reference (itdr): The gf-gc cross-battery assessment*. Pearson Education.

Means, B., Bakia, M., & Murphy, R. (2014). *Learning online: What research tells us about whether, when and how*. Routledge.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. Retrieved from http://www.jstor.org/stable/1723414

Mishra, P. & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017–1054. doi:10.1111/j.1467-9620.2006.00684.x

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. doi:10.1207/s15366359mea0204_1

Molenaar, P. C. M., Boomsma, D. I., & Dolan, C. V. (1993). A third source of developmental differences. *Behavior Genetics*, *23*(6), 519–524. doi:10.1007/bf01068142

Molnar, M. (2013, August 30). Ipad-centered 'steve jobs schools' open in the netherlands. Retrieved June 25, 2014, from http://blogs.edweek.org/edweek/marketplacek12/2013/08/ipad-centered_steve_jobs_schools_open_in_the_netherlands.html

Muller, D. A., Bewes, J., Sharma, M. D., & Reimann, P. (2007). Saying the wrong thing: Improving learning with multimedia by including misconceptions. *Journal of Computer Assisted Learning*, *24*(2), 144–155. doi:10.1111/j.1365-2729.2007.00248.x

Muller, D. A., Sharma, M. D., Eklund, J., & Reimann, P. (2007). Conceptual change through vicarious learning in an authentic physics setting. *Instructional Science*, *35*(6), 519–533. doi:10.1007/s11251-007-9017-6

Murphy, R., Gallagher, L., Krumm, A., Mislevy, J., & Hafter, A. (2014, March 7). Research on the use of khan academy in schools: Research brief. Retrieved July 8, 2014, from https:

//www.sri.com/work/publications/research-use-khan-academy-schools-research-brief

Murray, T. & Arroyo, I. (2002). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In *Intelligent tutoring systems* (pp. 749–758). Springer Berlin Heidelberg. doi:10.1007/3-540-47987-2_75

NCES. (2012, July 10). Ipeds changes for 2012-13 relating to distance education. Retrieved July 8, 2014, from http://nces.ed.gov/ipeds/news_room/ThisWeekInIPEDS.asp?TWID=75

Ng, A. (2014, May 16). A personal message from co-founder andrew ng. Retrieved July 10, 2014, from http://blog.coursera.org/post/85921942887/a-personal-message-from-co-founder-andrew-ng

Ngambi, D. & Bozalek, V. (2015). Editorial: Massive open online courses (moocs): Disrupting teaching and learning practices in higher education. *British Journal of Educational Technology*, *46*(3), 451–454. doi:10.1111/bjet.12281

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, *67*(2), 130–159. doi:10.1037/a0026699

Nižnan, J., Pelánek, R., & Řihák, J. (2015, July 26–29). Student models for prior knowledge estimation. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechnizkiy, A. Merceron, P. Mitros, ...M. Desmarais (Eds.), *Proceedings of the 8th international conference on educational data mining* (pp. 109–116). International Educational Data Mining Society. The 8th international conference on educational data mining. Madrid, Spain. Retrieved from http://www.educationaldatamining.org/EDM2015/proceedings/full109-116.pdf

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1–15. doi:10.1037/0033-295x.88.1.1

Norman, G. (2003). Rct = results confounded and trivial: The perils of grand educational experiments. *Medical Education*, *37*(7), 582–584. doi:10.1046/j.1365-2923.2003.01586.x

Novet, J. (2013, November 19). Coursera improves online education using – you guessed it – data. Retrieved August 11, 2014, from http://venturebeat.com/2013/11/19/coursera-improves-online-education-using-you-guessed-it-data/

Nyamsuren, E. & Taatgen, N. A. (2013). Set as an instance of a real-world visual-cognitive task. *Cognitive Science*, *37*(1), 146–175. doi:10.1111/cogs.12001

Olson, D. R. (2004). The triumph of hope over experience in the search for "what works": A response to Slavin. *Educational Researcher*, *33*(1), 24–26. doi:10.3102/0013189x033001024

Partchev, I. & Boeck, P. D. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32. doi:10.1016/j.intell.2011.11.002

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793. doi:10.1177/0956797615571017

Pelánek, R. (2014, July 4–7). Application of time decay functions and the elo system in student modeling. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th international conference on educational data mining* (pp. 21–27). International Educational Data Mining Society. The 7th international conference on educational data mining. London, UK. Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/21_EDM-2014-Full.pdf

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. doi:10.1007/s11257-016-9185-7

Perc, M. (2014). The Matthew effect in empirical data. *Journal of The Royal Society Interface*, *11*(98), 20140378–20140378. doi:10.1098/rsif.2014.0378

Piech, C., Huang, J., Nguyen, A., Phulsuksombati, M., Sahami, M., & Guibas, L. (2015). Learning program embeddings to propagate feedback on student code. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 1093–1102). Proceedings of Machine Learning Research. Lille, France: PMLR. Retrieved from http://proceedings.mlr.press/v37/piech15.html

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, *31*(6), 459–470. doi:10.1016/s0883-0355(99)00015-4

Plomin, R. & Deary, I. J. (2014). Genetics and intelligence differences: Five special findings. *Molecular Psychiatry*, *20*(1), 98–108. doi:10.1038/mp.2014.105

Poesse, R. & Wiles, S. (2016). Qualtrics lti bridge tool. Zenodo. doi:10.5281/zenodo.48436

Protopapas, A., Parrila, R., & Simos, P. G. (2014). In search of Matthew effects in reading. *Journal of Learning Disabilities*, *49*(5), 499–514. doi:10.1177/0022219414559974

Prusinkiewicz, P. & Runions, A. (2012). Computational models of plant development and form. *New Phytologist*, *193*(3), 549–569. doi:10.1111/j.1469-8137.2011.04009.x

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Rafferty, A. N., Ying, H., & Williams, J. J. (2018). Bandit assignment for educational experiments: Benefits to students versus statistical power. In *Lecture notes in computer science* (pp. 286–290). Springer International Publishing. doi:10.1007/978-3-319-93846-2_53

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Expanded edition, 1980. Chicago: The University of Chicago Press. Copenhagen: The Danish Institute of Educational Research.

Reber, R., Brun, M., & Mitterndorfer, K. (2008). The use of heuristics in intuitive mathematical judgment. *Psychonomic Bulletin & Review*, *15*(6), 1174–1178. doi:10.3758/pbr.15.6.1174

Reckase, M. (2009). *Multidimensional item response theory*. Springer New York. doi:10.1007/978-0-387-89976-3

Reich, J. (2015). Rebooting mooc research. *Science*, *347*(6217), 34–35. doi:10.1126/science.1261627

Rivera, S., Reiss, A., Eckert, M., & Menon, V. (2005). Developmental changes in mental arithmetic: Evidence for increased functional specialization in the left inferior parietal cortex. *Cerebral Cortex*, *15*(11), 1779–1790. doi:10.1093/cercor/bhi055

Rowan, D. (2013, August 27). Online education is redefining learning itself, says khan academy founder. Retrieved August 11, 2014, from http://www.wired.co.uk/magazine/archive/2013/08/start/reboot-the-teacher

RStudio Team. (2015). *Rstudio: Integrated development environment for r*. RStudio, Inc. Boston, MA. Retrieved from http://www.rstudio.com/

Ruzgis, P. (1994). Thurstone, l. l. (1887–1955). In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence*. Macmillan.

Salzmann, C., Gillet, D., & Piguet, Y. (2016). Mools for moocs: A first edx scalable implementation. In *2016 13th international conference on remote engineering and virtual instrumentation (rev)*. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/rev.2016.7444473

Sauce, B. & Matzel, L. D. (2018). The paradox of intelligence: Heritability and malleability coexist in hidden gene-environment interplay. *Psychological Bulletin*, *144*(1), 26–47. doi:10.1037/bul0000131

Savi, A. O., Ruijs, N. M., Maris, G. K. J., & van der Maas, H. L. J. (2018). Delaying access to a problem-skipping option increases effortful practice: Application of an a/b test in large-scale online learning. *Computers & Education*, *119*, 84–94. doi:10.1016/j.compedu.2017.12.008

Savi, A. O., van der Maas, H. L. J., & Maris, G. K. J. (2015). Navigating massive open online courses. *Science*, *347*(6225), 958–958. doi:10.1126/science.347.6225.958

Savi, A. O., Williams, J. J., Maris, G. K. J., & van der Maas, H. L. J. (2017, February 27). The role of a/b tests in the study of large-scale online learning. doi:10.17605/OSF.IO/83JSG

Schalke-Mandoux, D. S. (2016). *The impact of educational attainment and grade retention on the development of intelligence* (Doctoral dissertation). Retrieved from http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000102722

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84–95. doi:10.1016/j.intell.2016.08.003

Segal, A., David, Y. B., Williams, J. J., Gal, K., & Shalom, Y. (2018, April 14). Combining difficulty ranking with multi-armed bandits to sequence educational content. arXiv: http://arxiv.org/abs/1804.05212v1 [cs.AI]

Severance, C., Hanss, T., & Hardin, J. (2010). Ims learning tools interoperability: Enabling a mash-up approach to teaching and learning tools. *Technology, Instruction, Cognition and Learning*, *7*(3-4), 245–262.

Severance, C., Hardin, J., & Whyte, A. (2008). The coming functionality mash-up in personal learning environments. *Interactive Learning Environments*, *16*(1), 47–62. doi:10.1080/10494820701772694

Shah, D. (2016). By the numbers: Moocs in 2016. Retrieved April 2, 2017, from https://www.class-central.com/report/mooc-stats-2016/

Shaki, S. & Fischer, M. H. (2017). Competing biases in mental arithmetic: When division is more and multiplication is less. *Frontiers in Human Neuroscience*, *11*. doi:10.3389/fnhum.2017.00037

Shaywitz, B. A., Holford, T. R., Holahan, J. M., Fletcher, J. M., Stuebing, K. K., Francis, D. J., & Shaywitz, S. E. (1995). A Matthew effect for IQ but not for reading: Results from a longitudinal study. *Reading Research Quarterly*, *30*(4), 894. doi:10.2307/748203

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520. doi:10.1016/0010-0285(76)90016-5

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science*, *3*(1), 1–5. doi:10.1111/1467-8721.ep10769817

Siegler, R. S. & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, *46*(6), 606–620. doi:10 . 1037 / 0003 - 066x.46.6.606

Siemens, G. (2015). The role of moocs in the future of education. In *Moocs and open education around the world* (pp. viii–xvii). Routledge.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*(6), 467–482. Retrieved from http://www.jstor.org/stable/985254

Simonite, T. (2013, June 5). As data floods in, massive open online courses evolve. Retrieved June 25, 2014, from http://www.technologyreview.com/news/515396/as-data-floods-in-massive-open-online-courses-evolve/

Sims, Z. (2014, April 23). We're learning too: A new codecademy. Retrieved June 5, 2014, from http://www.codecademy.com/blog/136-we-re-learning-too-a-new-codecademy

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. doi:10.3102/0013189x031007015

Sonas, J. (2015). Chessmetrics formulas: Chessmetrics rating as a "weighted and padded simultaneous performance rating". Retrieved March 13, 2018, from http://www.chessmetrics.com/cm/CM2/Formulas.asp

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. Retrieved from http://www.jstor.org/stable/1412107

Spearman, C. (1927). *The abilities of man: Their nature and assessment*. Macmillan and Company, Limited. Retrieved from https://archive.org/details/abilitiesofman031969mbp

Spunt, R. P. & Adolphs, R. (2017). A new look at domain specificity: Insights from social neuroscience. *Nature Reviews Neuroscience*, *18*(9), 559–567. doi:10.1038/nrn.2017.76

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. Retrieved from http://www.jstor.org/stable/747612

Stevenson, S. (2014). How do you say addictive in spanish? Retrieved July 8, 2014, from http://www.slate.com/articles/technology/technology/2014/01/duolingo_the_free_language_learning_app_that_s_addictive_and_fun.html

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

Tabery, J. (2007). Biometric and developmental gene–environment interactions: Looking back, moving forward. *Development and Psychopathology*, *19*(04). doi:10 . 1017 / s0954579407000478

Takeuchi, L. M. & Vaala, S. (2014, January 26). Joan ganz cooney center - level up learning: A national survey on teaching with digital games. Retrieved October 21, 2014, from http://www.joanganzcooneycenter.org/publication/level-up-learning-a-national-survey-on-teaching-with-digital-games/

Taraghi, B., Frey, M., Saranti, A., Ebner, M., Müller, V., & Großmann, A. (2015). Determining the causing factors of errors for multiplication problems. In *Communications in computer and information science* (pp. 27–38). Springer International Publishing. doi:10.1007/978-3-319-22017-8_3

Taraghi, B., Saranti, A., Legenstein, R., & Ebner, M. (2016). Bayesian modelling of student misconceptions in the one-digit multiplication with probabilistic programming. In *Proceedings of the sixth international conference on learning analytics & knowledge - lak 16*. ACM Press. doi:10.1145/2883851.2883895

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. doi:10.1111/j.1745-3984.1983.tb00212.x

Templin, J. L. & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. doi:10.1037/1082-989x.11.3.287

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology, 1904-1920*, *8*(3), 271–281. doi:10.1111/j.2044-8295.1916.tb00133.x

Thomson, G. H. (1951). *The factorial analysis of human ability*. University Of London Press Limited. Retrieved from https://archive.org/details/factorialanalysi032965mbp

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. Teachers College Bureau of Publications. doi:10.1037/11240-000

Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.

Tryon, R. C. (1935). A theory of psychological components–an alternative to "mathematical factors." *Psychological Review*, *42*(5), 425–454. doi:10.1037/h0058874

Ungerleider, N. (2014). How duolingo uses a/b testing to understand the way you learn. Retrieved July 8, 2014, from https://www.fastcompany.com/3029531/how-duolingo-uses-a-b-testing-to-understand-the-way-you-learn

van den Bergh, M., Schmittmann, V. D., Hofman, A. 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. doi:10.2466/23.25.pms.121c26x6

van der Linden, W. J. & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Springer Netherlands. doi:10.1007/0-306-47531-6

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. doi:10.1037/0033-295x.113.4.842

van der Maas, H. L. J., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence*, *5*(2), 16. doi:10.3390/jintelligence5020016

van der Ven, S. H. G., Klaiber, J. D., & van der Maas, H. L. J. (2016). Four and twenty blackbirds: How transcoding ability mediates the relationship between visuospatial working memory and math in a language with inversion. *Educational Psychology*, 1–24. doi:10.1080/01443410.2016.1150421

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. doi:10.1016/j.lindif.2015.08.013

van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., Morris, R. G. M., & Fernández, G. (2014). Building on prior knowledge: Schema-dependent encoding processes relate to academic performance. *Journal of Cognitive Neuroscience*, *26*(10), 2250–2261. doi:10.1162/jocn_a_00630

VanLehn, K. (1986). Arithmetic procedures are induced from examples. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics*: (pp. 133–179). Hillsdale, NJ: Lawrence Erlbaum Associates. Retrieved from http://psycnet.apa.org/record/1986-98511-006

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. doi:10.1080/00461520.2011.611369

Veldkamp, B. P., Matteucci, M., & Eggen, T. J. H. M. (2011). Computerized adaptive testing in computer assisted learning? In *Communications in computer and information science* (pp. 28–39). Springer Berlin Heidelberg. doi:10.1007/978-3-642-20074-8_3

Vernau, K. & Hauptmann, M. (2014, April). Corporate learning goes digital: How companies can benefit from online education. Retrieved July 8, 2014, from http://www.rolandberger.com/media/pdf/Roland_Berger_TAB_Corporate_Learning_E_20140602.pdf

Vos, J., Evers, J. B., Buck-Sorlin, G. H., Andrieu, B., Chelle, M., & de Visser, P. H. B. (2009). Functional–structural plant modelling: A new versatile tool in crop science. *Journal of Experimental Botany*, *61*(8), 2101–2115. doi:10.1093/jxb/erp345

Vygotskii, L. S. (1978, July 1). *Mind in society*. Harvard University Press.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer*. NJ, US: Lawrence Erlbaum Associates, Inc.

Wang, C. & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. doi:10.1111/bmsp.12054

Watson, J., Murin, A., Vashaw, L., Butch, G., & Rapp, C. (2013). Keeping pace with k-12 online and blended learning. Retrieved August 18, 2014, from http://kpk12.com/cms/wp-content/uploads/EEG_KP2013-lr.pdf

Wauters, K., Desmet, P., & den Noortgate, W. V. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562. doi:10.1111/j.1365-2729.2010.00368.x

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. doi:10.1016/0001-6918(77)90012-9

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Williams, J. J., Li, N., Kim, J., Whitehill, J., Maldonado, S., Pechenizkiy, M., ... Heffernan, N. (2014). The mooclet framework: Improving online education through experimentation and personalization of modules. *SSRN Electronic Journal*. doi:10.2139/ssrn.2523265

Williams, J. J., Paunesku, D., Haley, B., & Sohl-Dickstein, J. (2013). Measurably increasing motivation in moocs. In *Aied 2013 workshops proceedings* (p. 55). Retrieved from https://people.csail.mit.edu/zp/moocshop2013/paper_22.pdf

Xiong, X. & Beck, J. E. (2014). A study of exploring different schedules of spacing and retrieval interval on mathematics skills in its environment. In *Intelligent tutoring systems* (pp. 504–509). Springer International Publishing. doi:10.1007/978-3-319-07221-0_63

A

Supplement to Chapter 6

Slips

Using the model outlined in the Methods section, and the method outlined in Equation 6.6 to quantify the likelihood of a particular misconception, it is irrelevant whether the model includes a 'slip' misconception that is connected to every error. We will show that the probability that misconception i and a slip is present, given that misconception i is the only additional misconception, is the same as the probability that misconception i is present, given only one misconception is present, in the model without slips. Let μ_s be the external field parameter associated with the slip and $a_{sj} = 1$ for all $j = 1, \ldots, m$ to indicate the slip is connected to all

errors.

$$p_{is} = p(c_i = 1, c_s = 1 | \sum_{i=1}^{n} c_i = 1, c_s = 1, \mathbf{e} = 1, \mu, \beta) \tag{A.1}$$

$$= \frac{\exp\left(\beta\mu_i + \beta\mu_s + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{ij} + 2\beta\sum_{j=1}^{m}a_{sj}\right)}{\sum_{k=1}^{n}\exp\left(\beta\mu_k + \beta\mu_s + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{kj} + 2\beta\sum_{j=1}^{m}a_{sj}\right)} \tag{A.2}$$

$$= \frac{\exp\left(\beta\mu_i + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{ij}\right)\exp\left(\beta\mu_s + 2\beta\sum_{j=1}^{m}a_{sj}\right)}{\sum_{k=1}^{n}\exp\left(\beta\mu_k + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{kj}\right)\exp\left(\beta\mu_s + 2\beta\sum_{j=1}^{m}a_{sj}\right)} \tag{A.3}$$

$$= \frac{\exp\left(\beta\mu_i + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{ij}\right)}{\sum_{k=1}^{n}\exp\left(\beta\mu_k + \beta\mu_s + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{kj}\right)} \tag{A.4}$$

$$= p(c_i = 1 | \sum_{i=1}^{n} c_i = 1, \mathbf{e} = \mathbf{1}, \mu, \beta) \tag{A.5}$$

$$= p_i \tag{A.6}$$

B

Supplement to Chapter 7

The marginal distribution

The distribution of the node states $p(x)$, i.e., the probability of observing a particular configuration of skills, knowledge or abilities, is an integral part of our formal theory. In fact, each of the phenomena that we aim to explain with the proposed model constitutes a particular pattern of observations from this marginal distribution. It is therefore highly convenient that the marginal distribution $p(x)$ of the model for the node states x and their relations ω is of a known form (e.g., Fortuin & Kasteleyn, 1972; Grimmett, 2006). We next show that the marginal distribution $p(x)$ of the formal model is the Ising network model (Ising, 1925), i.e., is of the form

$$\sum_{\omega \in \Omega} p(x, \omega) = p(x) = \frac{1}{Z_I} \exp\left(\sum_{1 \leq i < j \leq p} \sigma_{ij} x_i x_j + \sum_{i=1}^{p} \mu_i x_i \right),$$

where μ_i denotes the main effect of node i and σ_{ij} denotes the interaction between nodes i and j. In the general case, nodes i and j tend to be in the same state when $\sigma_{ij} > 0$, and tend to be in different states when $\sigma_{ij} < 0$. In the absence of the influence of other variables in the network, node i tends to $+1$ when $\mu_i > 0$ and tends to -1 when $\mu_i < 0$. Here, it is assumed that the σ_{ij}

and μ_i are all positive.

Without loss of generality we consider the case where the edge set E consists of all $p(p-1)/2$ possible edges between p nodes, with a unique probability ϑ_{ij} is associated to each pair of nodes i and j, and a unique external field μ_i associated to each node i. The Fortuin-Kasteleyn model that is consistent with this specification is,

$$p(x,\omega) = \frac{1}{Z_F} \prod_{1 \le i < j \le p} \left\{ \vartheta_{ij}\, \delta_{(\omega_{ij},1)}\, \delta_{(x_i,x_j)} + (1-\vartheta_{ij}) \delta_{(\omega_{ij},0)} \right\} \times \prod_{i=1}^{p} \exp\left(\mu_i \left[\delta_{(x_i,1)} - \delta_{(x_i,-1)}\right]\right).$$

The first step in expressing its marginal distribution $p(x)$ is to sum out the edge states

$$p(x) = \sum_{\omega \in \Omega} p(x,\omega) = \frac{1}{Z_F} \prod_{1 \le i < j \le p} \left\{ \vartheta_{ij}\, \delta_{(x_i,x_j)} + 1 - \vartheta_{ij} \right\} \prod_{i=1}^{p} \exp\left(\mu_i \left[\delta_{(x_i,1)} - \delta_{(x_i,-1)}\right]\right).$$

The second step is to rewrite this expression using $\vartheta_{ij} = 1 - \exp(-2\sigma_{ij})$, making use of the assumption that $\sigma_{ij} > 0$. Observe that for a pair of nodes i and j this boils down to,

$$\vartheta_{ij}\, \delta_{(x_i,x_j)} + 1 - \vartheta_{ij} \Rightarrow \left\{1 - \exp\left(-2\sigma_{ij}\right)\right\} \delta_{(x_i,x_j)} + \exp\left(-2\sigma_{ij}\right) = \exp\left(2\sigma_{ij}[\delta_{(x_i,x_j)} - 1]\right).$$

The third step is to make use of the relations $\delta_{(x_i,x_j)} = \frac{1}{2}(1 + x_i x_j)$ and $\delta_{(x_i,1)} - \delta_{(x_i,-1)} = x_i$, which leads us to the following expression:

$$\begin{aligned} p(x) &= \frac{1}{Z_F} \prod_{1 \le i < j \le p} \exp\left(\sigma_{ij}\left[x_i x_j - \frac{1}{2}\right]\right) \prod_{i=1}^{p} \exp\left(\mu_i x_i\right) \\ &= \frac{1}{Z_F} \exp\left(\sum_{1 \le i < j \le p} \sigma_{ij} x_i x_j - \frac{1}{2} \sum_{1 \le i < j \le p} \sigma_{ij} + \sum_{i=1}^{p} \mu_i x_i \right). \end{aligned}$$

Based on the above it is now trivial to see that

$$Z_F \exp\left(\frac{1}{2} \sum_{1 \le i < j \le p} \sigma_{ij} \right) = Z_I = \sum_{x} \exp\left(\sum_{1 \le i < j \le p} \sigma_{ij} x_i x_j + \sum_{i=1}^{p} \mu_i x_i \right),$$

where the sum on the right hand side is taken over all 2^p possible realizations of x. That is, we

have found the marginal distribution

$$p(x) = \frac{1}{Z_I} \exp\left(\sum_{1 \leq i < j \leq p} \sigma_{ij} x_i x_j + \sum_{i=1}^{p} \mu_i x_i \right),$$

which we set out to do.

The positive manifold and the Rasch model

The starting point of the formal model is a simplest non-trivial Fortuin-Kasteleyn model, which assumes a single parameter ϑ for all pairs of nodes i and j. The marginal distribution $p(x)$ that is associated to this simplest case is known as the Curie-Weiss model (Kac, 1968),

$$p(x) = \frac{1}{Z_C} \exp\left(\sum_{1 \leq i < j \leq p} \sigma x_i x_j + \sum_{i=1}^{p} \mu_i x_i \right) = \frac{1}{Z_C} \exp\left(\sigma \left(\sum_{i=1}^{p} x_i \right)^2 + \sum_{i=1}^{p} \mu_i x_i \right),$$

which is consistent with a fully connected network with all node pairs having the same interaction strength σ.

We have shown elsewhere (e.g., Marsman et al., 2018) that the Curie-Weiss network model is an analytic characterization of a marginal Rasch model,

$$p(x) = \int_{\mathbb{R}} p(x \mid \vartheta) f(\vartheta) \, \mathrm{d}\vartheta,$$

where $p(x \mid \vartheta)$ denotes the Rasch model using an ability parameter ϑ, and $f(\vartheta)$ denotes the latent variables' distribution, a mixture of normal distributions. In this characterization, the μ_i are item-easiness parameters and $2\sqrt{\sigma}$ is an overall discrimination index (or standard deviation of the latent variable). Importantly, the Rasch model is well-known for its ability to generate data that are consistent with the positive manifold.

The hierarchical structure and multidimensional IRT

The idea is to group nodes into communities and to introduce one probability ϑ_W for connecting nodes within a community and one probability ϑ_B—with $\vartheta_B < \vartheta_W$—to connect nodes from different communities. The two community specific probabilities ϑ_W and ϑ_B in our Fortuin-Kasteleyn model translate to community specific interaction terms $\sigma_W = -\frac{1}{2} \ln(1 -$

ϑ_W) and $\sigma_B = -\frac{1}{2}\ln(1-\vartheta_B)$ in the associated Ising model.

One way to represent the structure this imposes in the Ising model is to consider it in the form

$$p(x) = \frac{1}{Z_I} = \exp\left(\frac{1}{2}x^\mathsf{T}\Sigma x + x^\mathsf{T}\mu\right),$$

where $\Sigma = [\sigma_{ij}]$ is a $p \times p$ connectivity matrix that consists of two parts:

$$\Sigma = \sigma_B \mathrm{I}_p + (\sigma_W - \sigma_B)\mathrm{C} = \sigma_B \mathrm{I}_p + \sigma_C \mathrm{C},$$

where I_p is the $p \times p$ matrix of ones, $\sigma_C = \sigma_W - \sigma_B > 0$, and where $\mathrm{C} = [c_{ij}]$ is a $p \times p$ matrix with entry $c_{ij} = 1$ when nodes i and j belong to the same community and $c_{ij} = 0$ otherwise. When the nodes are ordered w.r.t. their communities we find that C—and thus also Σ—is a block-diagonal matrix.

Observe that we can decompose the quadratic form $x^\mathsf{T}\Sigma x$ as follows

$$x^\mathsf{T}\Sigma x = x^\mathsf{T}\left(\sigma_B \mathrm{I}_p + \sigma_C \mathrm{C}\right)x = \sigma_B\left(\sum_{i=1}^{p} x_i\right)^2 + \sum_{c=1}^{n}\sigma_C\left(\sum_{i\in \mathrm{V}_c} x_i\right)^2,$$

where the second term on the right consists of a sum over nodes i in a community c, for communities $c = 1, \ldots, n$. Using ideas that are similar to the ideas that were used to characterize the latent variable expression of the Curie-Weiss model (e.g., Marsman et al., 2018), we can express the Ising model,

$$p(x) = \frac{1}{Z_i}\exp\left(\frac{1}{2}\sigma_B\left(\sum_{i=1}^{p} x_i\right)^2 + \frac{1}{2}\sum_{c=1}^{n}\sigma_C\left(\sum_{i\in \mathrm{V}_c} x_i\right)^2 + \sum_{i=1}^{p}\mu_i x_i\right),$$

as the marginal distribution of an $n+1$ dimensional two-parameter logistic model (Reckase, 2009) with a latent variable associated to each of the $n+1$ terms in the quadratic form above. For $n = 2$ communities the matrix of discriminations—factor loadings—of the three-dimensional

IRT model is of the form

$$\Lambda = \begin{pmatrix} \sqrt{2\sigma_B} & \sqrt{2\sigma_C} & 0 \\ \vdots & \vdots & \vdots \\ \sqrt{2\sigma_B} & \sqrt{2\sigma_C} & 0 \\ \sqrt{2\sigma_B} & 0 & \sqrt{2\sigma_C} \\ \vdots & \vdots & \vdots \\ \sqrt{2\sigma_B} & 0 & \sqrt{2\sigma_C} \end{pmatrix},$$

which reflects one single overarching dimension, and n community specific dimensions, i.e., a hierarchical structure.

The external field

In the Ising network model the external field is a main effect that is associated to nodes; node i tends to $+1$ when $\mu_i > 0$ and to -1 when $\mu_i < 0$. In the IRT model the external field is an item easiness parameter; item i is correctly answered $(+1)$ more often when $\mu_i > 0$ and answered incorrectly more often when $\mu_i < 0$.[1] In the Fortuin-Kasteleyn model the external field has a similar effect; node i tends to $+1$ when $\mu_i > 0$ and to -1 when $\mu_i < 0$. This influence of the external field, however, is a population effect in the Fortuin-Kasteleyn model, as for individual networks the external field acts on the clusters and not on the individual nodes.

The effect of the external field for an individual network in the Fortuin-Kasteleyn model is most easily revealed through the conditional distribution $p(x \mid \omega)$, which factor into the conditional distributions

$$p(x \mid \omega) = \prod_{k=1}^{\kappa(\omega)} p(x_i \mid i \in K_k(\omega)),$$

where $\kappa(\omega)$ denotes the number of open clusters in a network, and $K_k(\omega)$ the set of nodes that are in cluster k, $k = 1, \ldots, \kappa(\omega)$. The probability that the node states in cluster k are equal to

[1] The population average in the latent variable expression of the Ising model is fixed at zero in each dimension.

+1 is (Cioletti & Vila, 2015)

$$p(x_i = +1 \mid i \in K_k(\omega)) = \frac{\exp\left(\sum_{i \in K_k(\omega)} \mu_i\right)}{\exp\left(\sum_{i \in K_k(\omega)} \mu_i\right) + \exp\left(-\sum_{i \in K_k(\omega)} \mu_i\right)}.$$

Observe that this probability depends on the sum of the external fields of all nodes in the cluster. In the model there is a positive external field μ that applies to all nodes equally. In this case, the probability that the node states in cluster k are equal to $+1$ is

$$p(x_i = +1 \mid i \in K_k(\omega)) = \frac{\exp\left(|K_k(\omega)|\mu\right)}{\exp\left(|K_k(\omega)|\mu\right) + \exp\left(-|K_k(\omega)|\mu\right)},$$

where $|K_k(\omega)|$ is the size of cluster k. With $\mu > 0$ this probability is strictly larger than 0.5 for each cluster, and it is an increasing function of cluster size. The original symmetric case is obtained when $\mu = 0$, which reveals the uniform assignment of the values $+1$ and -1 across clusters.

The external field also affects the correlation between any two nodes x_i and x_j, as it alters the probability that the nodes are in the same state. The following expression from Cioletti and Vila (2015, p. 93) confirms this,

$$p(x_i = x_j) = \frac{1}{2} p(i \leftrightarrow j) + \frac{1}{2} \mathbb{E}\left(\mathcal{I}(i \not\leftrightarrow j) \times \tanh\left(|K_t(\omega)|\mu\right) \times \tanh\left(|K_u(\omega)|\mu\right)\right),$$

where $\mathcal{I}(i \not\leftrightarrow j)$ denotes an indicator function that is equal to one whenever nodes i and j are not connected, and $K_t(\omega)$ and $K_u = K_u(\omega)$ are two disjoint connected components, with $x_i \in K_t(\omega)$ and $x_j \in K_u(\omega)$. However, $\tanh\left(|K_k(\omega)|\mu\right)$ is positive for every cluster k, as $\mu > 0$. This implies that the term on the right hand side of the expression for $p(x_i = x_j)$ is positive, and thus that $p(x_i = x_j) > 0.5$. In other words, a positive manifold.

Acknowledgements

The research presented in this book would not have been possible without many magnificent collaborators. Here the original publications are listed, along with the various contributors.

Chapter 1 The quoted letter: Savi, A. O., van der Maas, H. L. J., & Maris, G. K. J. (2015). Navigating massive open online courses. *Science, 347*, 958. doi:10.1126/science.347.6225.958

Chapter 2 Brinkhuis, M. J. S., Savi, A. O., Hofman, A. D., Coomans, F., van der Maas, H. L. J., & Maris, G. K. J. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics, 5*, 29-46. doi:10.18608/jla.2018.52.3

Matthieu Brinkhuis and Alexander Savi contributed equally to this work, and Frederik Coomans and Abe Hofman contributed equally to this work. We sincerely thank Oefenweb for making available the data.

Chapter 3 Savi, A. O., Williams, J. J., Maris, G. K. J., & van der Maas, H. L. J. (2017, February 27). The role of A/B tests in the study of large-scale online learning. doi:10.31219/osf.io/83jsg

The article was written in 2014, in order to study the state of online learning at the time, and to propose a direction for future research.

Chapter 4 Savi, A. O., Ruijs, N. M., Maris, G. K. J., & van der Maas, H. L. J. (2018). Delaying access to a problem-skipping option increases effortful practice: Application of an A/B test in large-scale online learning. *Computers & Education, 119*, 84-94. doi:10.1016/j.compedu.2017.12.008

We sincerely thank Oefenweb for giving us the opportunity to run the experiment and making available the data.

Chapter 5 Savi, A. O. (2017, July 13). Customization in online learning: A tool to support MOOC teachers with questionnaires, adaptive lessons, and experimental comparisons. doi:10.31219/osf.io/q53jx

I sincerely thank Rens Poesse for rewriting and extending the LTI protocol and including extensive documentation, and Annemarie Zand Scholten for providing an opportunity to test the the protocol.

Chapter 6 Savi, A. O., Deonovic, B. E., Bolsinova, M., van der Maas, H. L. J., & Maris, G. K. J. (2018, July 31). Automated diagnosis of misconceptions in single digit multiplication. doi:10.31234/osf.io/53muj

We sincerely thank Oefenweb for making available the data.

Chapter 7 Savi, A. O., Marsman, M., van der Maas, H. L. J., & Maris, G. K. J. (2018, July 26). The wiring of intelligence. doi:10.31234/osf.io/32wr8

We sincerely thank Abe Hofman and Frederik Coomans for their valuable contributions during the early stages of this research, and Lourens Waldorp for the invaluable consults.

Nederlandse samenvatting

Onderwijs kun je zien als een lange sequentie van interventies, in een zelforganiserend en ontwikkelend systeem. Wat ik daarmee bedoel wordt duidelijk aan de hand van een voorbeeld uit tuinieren.

Als je ooit hebt geprobeerd om je eigen groenten te kweken, dan weet je dat het een delicate activiteit kan zijn. Je weet ook dat ondanks dat een deel van de verzorging eenmalig is, zoals het vinden van een plek met de juiste temperatuur en de gewenste hoeveelheid zonlicht, het grootste deel structureel is, zoals het verzorgen van voeding en bewatering. Afhankelijk van je vaardigheden, de eisen van de plant en de geschiktheid van zijn omgeving, zal hij bloeien of verwelken.

Nu vraag je je af, waarom zou ik me druk maken over de verzorging van planten als het onderwerp onderwijs is? De reden is simpel. Plantenzorg kan simpelweg worden beschouwd als een lange sequentie van interventies—water geven, voedingsstoffen toevoegen, opnieuw water geven, bladluis verwijderen, enzovoort. Dit is niet anders in het onderwijs. Natuurlijk zijn de interventies die onderwijs vormen veelal complexer dan de heldere ingrepen in de plant-analogie. Maar onderwijs kan ook worden gezien als een lange sequentie van interventies—motiveren, oefeningen geven, instrueren, enzovoort—deze keer gericht op de cognitieve (of sociale, emotionele of affectieve) groei van de student, in plaats van de fysieke groei van een plant.

Bovendien zijn de doelstellingen van tuinieren en onderwijs vergelijkbaar. Doorgaans streeft een tuinman naar het creëren van de optimale omstandigheden voor zijn of haar planten om te bloeien met minimale structurele inspanningen, zodat idealiter de tuin in toenemende mate zelfondersteunend wordt. Met veel verschillende soorten planten is dit geen triviale taak. Ook docenten streven ernaar de individuele studenten de optimale voorwaarden te bieden voor zelfondersteunend leren. Eveneens een zeer uitdagende taak.

De plant-analogie en de abstracte, sequentiële, interventieconceptualisatie van onderwijs (hierna de *educatieve sequentie* genoemd) helpen ons de hoofdthema's in dit boek te onderscheiden. Hieronder introduceer ik deze hoofdthema's, verduidelijk ik de educatieve sequentie en

gebruik ik deze educatieve sequentie uiteindelijk als een rode draad om de hoofdstukken die volgen te introduceren.

Idiografie

Het hoofdthema van dit boek begint waar de plantanalogie stopt. Eeuwen van veredeling hebben gewassen gecreëerd met allerlei gewenste eigenschappen. Als resultaat daarvan profiteert elke individuele plant, bijvoorbeeld in een veld vol met tulpen, op dezelfde manier van dezelfde behandeling. Denk nu, in plaats van aan een veld vol tulpen, aan een klaslokaal vol studenten. Deze studenten kunnen heel verschillend profiteren van de educatieve sequentie. Factoren als gezinssituatie, gezondheid, welvaart en buitenschoolse activiteiten, kunnen allemaal bijdragen aan een enorme variabiliteit waarin een educatieve interventie het beste bij een student past, en op welk moment die interventie het best kan worden toegepast.

Deze heterogeniteit vraagt om een idiografische benadering van de wetenschap van het onderwijs. Idiografie wordt gedefinieerd als de studie van het individu, en idiografische wetenschap wordt vaak gecontrasteerd met nomothetische wetenschap, de formulering van universele wetten. In de wetenschappelijke psychologie legt Molenaar (2004) uit dat de idiografische wetenschap "de specifieke studie van het individu [...] terugbrengt, voorafgaand aan het samenbrengen met andere individuen. Elke persoon wordt aanvankelijk opgevat als een mogelijk uniek systeem van interactieve dynamische processen, waarvan de ontvouwing aanleiding geeft tot een individueel levenspad in een hoogdimensionale psychologische ruimte."

In onderwijs wordt de idiografische benadering bijvoorbeeld gerechtvaardigd door het feit dat individueel onderwijs superieure leerresultaten oplevert ten opzichte van traditioneel klassikaal onderwijs. Hieruit volgt het idee dat de educatieve sequentie moet worden toegespitst op het individu. Een karikaturale beschrijving van traditioneel onderwijs aan de andere kant, kan bestaan uit frontale instructie en lineaire lesmethoden, waarbij exact het tegenovergestelde bewerkstelligd wordt—opeenvolgingen van interventies worden gecreëerd die sterk vergelijkbaar zijn voor elk individu. De achterkant van dit boek illustreert dergelijke ongewenste sequenties, waarbij de letters interventies voorstellen, de kleuren verschillende soorten interventies vertegenwoordigen en de rijen individuele reeksen van interventies vertegenwoordigen.

Je zou kunnen beweren dat zo'n karikatuur nauwelijks bestaat, maar een vrij recente onderwijsontwikkeling, die van Massive Open Online Courses (MOOC's), komt heel dichtbij. Hoewel er veel te zeggen valt voor het leren op grote schaal, de uitdaging om MOOC's af te stemmen op de behoeften van de individuele student is een serieus probleem, en een probleem

dat we—Savi e.a. (2015)—aan de orde stellen in Science, in reactie op een actuele en constructieve discussie van MOOC-onderzoek door Reich (2015):

> We zijn het volledig eens met J. Reich dat onderzoek naar de effectiviteit van Massive Open Online Courses (MOOC's) zich moet concentreren op leren in plaats van alleen maar klikken ("Rebooting MOOC research," Education Forum, 2 januari, p. 34). Onze grootste uitdaging zal zijn uit te zoeken wat het meest geschikt is voor een individuele student op een gegeven moment.
>
> Idealiter zou een MOOC moeten werken als de GPS-navigatie in je auto. Je vertelt het waar je heen wilt, het zoekt uit waar je bent en het leidt je over de meest optimale route. Als we de analogie aanhouden dan zijn de huidige MOOC's net alsof alle GPS-navigatie elke autorijder instrueert om op maandagochtend om 9.15 uur rechts af te slaan.
>
> Als we onderwijs en oefening niet kunnen aanpassen aan de individuele leerling, zullen MOOC's nooit meer zijn dan een digitale vorm van lesgeven in de klas. Om de leerervaring te personaliseren, hebben we eerst een gedetailleerde beschrijving nodig van wat een student wel en niet kan. Dergelijke informatie kan worden bepaald door traditionele tests of door krachtigere methoden, zoals de oefengebaseerde volgsystemen die al bestaan in andere domeinen van online onderwijs (Klinkenberg e.a., 2011). De A/B-testen die worden besproken in het Education Forum, bieden ons een ideale methodologie om te beginnen met het plaatsen van wegen op de educatieve kaart. Zodra we informatie hebben verzameld over verschillende omstandigheden, kunnen we de optimale route van elke student in kaart brengen.

Als je de dimensie van educatieve sequenties neemt, is het ene uiterste gevuld met sequenties die identiek zijn voor elke student (geïllustreerd op de achterkant van dit boek), terwijl aan de andere kant alle sequenties perfect zijn afgestemd op het individu (geïllustreerd op de voorkant van dit boek). Alle educatieve programma's liggen ergens tussen deze twee uitersten, en in dit boek onderzocht ik methoden die kunnen helpen bij het verbeteren van de afstemming van onderwijs.

In vivo laboratorium

Het tweede thema van dit boek kan worden onderbouwd door een zeer verwante ongelijkheid tussen planten en mensen. Terwijl planten niet alleen onder extreem gecontroleerde omstandig-

heden kunnen worden bestudeerd, zoals in kassen, kunnen ze bovendien ook worden veredeld en zelfs genetisch gemanipuleerd. Overweeg deze benadering nu met mensen. De meesten van ons zijn het erover eens dat we dan al snel tegen serieuze ethische beperkingen aanlopen. Het is een van de redenen dat er verfijnde plantmodellen bestaan (bijvoorbeeld Vos e.a., 2009) die fenomenen als groeisnelheid en vertakkingen kunnen verklaren, terwijl men in de Psychologie forse uitdagingen heeft bij het in kaart brengen van de enorme variabiliteit onder mensen.

Zacht uitgedrukt zijn mensen een zeer lastig onderwerp om te bestuderen, en zo geldt dat dus ook voor studenten. Cognitief-psychologen zijn succesvol in het toepassen van de experimentele methode, wat de ontdekking heeft vergemakkelijkt van vele belangrijke effecten op het gebied van leren (bijvoorbeeld Dunlosky e.a., 2013; Karpicke & Roediger, 2008). Verder hebben onderwijspsychologen gerandomiseerde onderzoek geadopteerd—wat in klinisch onderzoek als een 'gouden standaard' worden beschouwd—om het effect van grote onderwijs-interventies te bestuderen. In onderwijsonderzoek worden dergelijke studies echter geconfronteerd met verschillende kritieken, waaronder een nogal ernstige: het is vrijwel onmogelijk om dergelijk onderzoek dubbelblind uit te voeren.

Een specifieke vorm van veldexperimenten biedt een elegante oplossing voor veel van de problemen die zich voordoen in onderwijsonderzoek. In dit boek maken we dankbaar gebruik van de opkomst van grootschalige online leeromgevingen. Deze omgevingen zijn fascinerend, niet alleen omdat randomisatie en een dubbelblinde uitvoering over het algemeen gemakkelijk kunnen worden bereikt, maar bovendien omdat ze ons een blik gunnen in een natuurlijke en authentieke leercontext. Online leeromgevingen zijn vanzelfsprekend niet hetzelfde als een daadwerkelijke klaslokaal, maar kunnen wel worden gezien als een model voor menselijk leren, dat wordt gevormd door zowel de online ervaringen als de ervaringen in het klaslokaal.

Bovendien creëren online leeromgevingen op grote schaal data die moeilijk of zelfs onmogelijk te bereiken zijn in traditioneel onderwijs. In Hoofdstuk 2 laten we zien dat deze omgevingen mogelijk niet alleen het werk van studenten inzichtelijk maken, door het type, het aantal en de volgorde van de oefeningen te verzamelen, maar bovendien ook de soorten fouten, reactietijden, oefening-moeilijkheden en student-vaardigheden kunnen verzamelen. Indrukwekkend genoeg gebeurt dat live, met weinig inspanning, op een enorme schaal, en zijn we bijgevolg steeds beter in staat om de zogenaamde *microgenetica* van natuurlijk leren te vangen. Analoog aan de kassen van biologen, bieden online leeromgevingen onderwijspsychologen de middelen voor een systematisch onderzoek naar het natuurlijke leren van de mens.

Leren en vaardigheid

Tot slot betreft het derde thema van dit boek een schijnbare dichotomie in de wetenschappelijke psychologie. Welbekend is Cronbach (1957)—toenmalig president van de American Psychological Association—die zich richtte op "de scheiding van de disciplines": de observatie dat experimentele en correlationele psychologie in relatief isolement bestaan. Hier verwijst experimentele psychologie naar de pogingen om variatie binnen personen te verklaren, terwijl correlationele psychologie verwijst naar de pogingen om variatie tussen personen te verklaren. De gelijkenis tussen dit probleem en de nomothetische en idiografische benaderingen van de wetenschap valt hier op.

Slechts een paar jaar voorafgaand aan de presidentiële toespraak van Cronbach, besprak Ferguson (1954) dezelfde kwestie met betrekking tot intelligentie. In een artikel dat leest als een roman, stelt hij een enkel conceptueel raamwerk voor dat het menselijk leren en de menselijke vaardigheid moet overbruggen. In zijn woorden, "zij die zich bezighouden met de beschrijving en classificatie van de vaardigheid van de mens hebben meestal een individuele verschilbenadering aangenomen. Ze hebben nauwelijks aandacht besteed aan problemen van leren. De experimentalisten, verdiept in de studie van het leren, hebben om verschillende theoretische en praktische redenen weinig belangstelling getoond voor individuele verschillen. Ze lijken zich er niet van bewust dat ook zij studenten zijn van de vaardigheden van de mens."

Hoewel Klinkenberg e.a. (2011) zich er mogelijk niet van bewust waren, komt hun computeradaptieve oefenomgeving voor rekenen—dat wordt bestudeerd in Hoofdstuk 2, 4 en 6—dicht bij het verenigen van de twee disciplines. Aan de ene kant gaat hun adaptieve algoritme, dat studenten koppelt aan oefeningen, om met belangrijke leerprincipes, zoals de zone van naaste ontwikkeling en *scaffolding*. Aan de andere kant biedt precies hetzelfde algoritme vaardigheidsschattingen die de ontwikkeling van elke student volgen.

In Hoofdstuk 7 lossen we de belofte in om de twee disciplines te verenigen, vanuit een sterk theoretisch gezichtspunt. We stellen een formeel raamwerk voor dat fundamentele fenomenen in menselijke vaardigheid verklaart, en dat de cruciale brug naar het menselijk leren biedt. Onze idiografische theorie maakt het niet alleen mogelijk om theorieën over leren te evalueren, maar ook om het effect van educatieve interventies te bestuderen, en geeft de educatieve sequentie dus haar gewenste plek in de studie naar menselijke vaardigheidheden.

De besproken thema's bieden een goede context voor het begrijpen van de verschillende hoofdstukken. Daarnaast vat ik in het kort elk van de hoofdstukken samen en leg ik uit hoe ze met elkaar zijn verbonden.

Overzicht

In Hoofdstuk 2 bespreken we Rekentuin, de eerdergenoemde computer-adaptieve oefenomgeving voor rekenonderwijs. Rekentuin past de educatieve sequentie aan door leerlingen te koppelen aan rekenoefeningen, op basis van real-time vaardigheids- en moeilijkheidsschattingen. We bespreken de methoden die worden gebruikt om het systeem in realtime te sturen en analyseren, door middel van een scoreregel voor nauwkeurigheid en reactietijd, en het adaptieve algoritme dat studenten koppelt aan oefeningen. In het bijzonder bespreken we verschillende uitdagingen die bij deze aanpak aan de orde komen. We onderzoeken de mate waarin het voor de schattingen gebruikte model aansluit bij de manier waarop leerlingen oefenen, en achterhalen verschillende vormen van misfit, zoals schendingen van de veronderstelde unidimensionaliteit. Bij de synthese van onze ervaringen suggereren we dat *learning analytics* actief moet helpen bij het nastreven van de gewenste educatieve doelstelling, die kan worden bereikt door de cyclus van analytics en interventies te verkorten.

In Hoofdstuk 3 bespreken we de staat van online leren, en stellen we een methodologie voor ter vergroting van de leerwinst van online leeromgevingen. We bespreken het succes van het aantrekken van grote aantallen leerlingen wereldwijd en het falen om de beloftes van online leren waar te maken. Belangrijk is dat we een weg vooruit voorstellen: grootschalige online gerandomiseerde, gecontroleerde veldexperimenten. We stellen dat online leeromgevingen gezien kunnen worden als in vivo laboratoria, wat uiteindelijk ten goede komt van de student. We geven een kort overzicht van het gebied van online leren in 2014. We voeren aan dat bij dergelijke tests rekening moet worden gehouden met de typische aard van (online) leren, en dat het gebruik van kennis uit de verschillende leerwetenschappen moet worden aangemoedigd om interventies te identificeren die een beter leerproces beloven. Tot slot hebben we zowel de beperkingen als de beloften van deze zogenaamde A/B-tests geïdentificeerd en laten we zien hoe deze uiteindelijk kunnen bijdragen aan leren dat is toegesneden op elke individuele leerling.

In Hoofdstuk 4 gebruiken we de voorgestelde benadering—online gerandomiseerde experimenten—om Rekentuin te optimaliseren. Het experiment is bedoeld om een onvoorziene mogelijkheid om met minimale inspanning te kunnen oefenen te elimineren. Sommige kinderen hebben de neiging oefeningen over te slaan die een te grote inspanning vergen en beperken zich tot de oefeningen die ze spontaan kunnen oplossen. Onze interventie vertraagde de mogelijkheid om een oefening over te slaan, en bevorderde daarmee de inspanning die kinderen moeten leveren. De resultaten tonen een toename van de inspanning, zonder dat dit ten koste gaat van het plezier. We kunnen niet bevestigen of de extra inspanning ook een positieve invloed op de

leerwinst van de kinderen heeft. Belangrijk is dat we tot slot het *holdout*-principe introduceren en illustreren: een methode om de betrouwbaarheid van verkennend onderzoek te vergroten.

In Hoofdstuk 5 betoog ik dat online leeromgevingen veel kansen bieden, op voorwaarde dat docenten over voldoende mogelijkheden beschikken. Op maat gemaakte educatieve sequenties en evidence-based verbeteringen op basis van experimentele vergelijkingen vereisen veelzijdige leeromgevingen. Helaas is dit niet altijd het geval, omdat de mogelijkheid tot maatwerk vaak beperkt is tot de functionaliteit van de gekozen software. Ik bespreek *Learning Tools Interoperability* als een belangrijke uitzondering, omdat het docenten in staat stelt om een online leeromgeving uit te breiden met externe software, om zo het pedagogisch bereik te vergroten. In dit hoofdstuk introduceer ik een softwareprotocol dat *Qualtrics*—populaire software voor het maken en verspreiden van vragenlijsten—inzet om de standaard functionaliteit van een online leeromgeving uit te breiden met niet alleen aanvullende onderwijselementen, maar ook de mogelijkheid om onderwijselementen te personaliseren, en de mogelijkheid om studenten willekeurig toe te wijzen aan verschillende onderwijs-condities ten behoeve van gerandomiseerd onderzoek. Op deze manier kunnen de middelen voor adaptiviteit en experimentele vergelijkingen die door Qualtrics worden aangeboden—zij het rudimentair—nu door zowel leraren als onderwijsonderzoekers worden gebruikt.

In Hoofdstuk 6 introduceren we een model dat tot doel heeft misvattingen van een student te identificeren op basis van de fouten die hij of zij maakt. Waar experimenten cruciaal zijn om te bepalen welke interventie het beste werkt in de educatieve sequentie, zijn goede vaardigheidsmetingen cruciaal bij het bepalen wanneer een bepaalde interventie moet worden gebruikt. Waar Rekentuin de vaardigheidsschatting van een persoon gebruikt om de timing van een oefening te bepalen, verkennen we in dit hoofdstuk een diagnostische benadering. Cognitieve diagnose is gericht op het begrijpen van de tekortkomingen van een individu, en hier introduceren we een intuïtief model om de misvattingen te identificeren die de fouten van een student veroorzaken. We passen de methode toe op enkelcijferige vermenigvuldiging; een domein dat zeer geschikt is voor onze methode, dat goed is bestudeerd, en ons bovendien in staat heeft gesteld meer dan 25.000 fouten van 335 studenten te analyseren. De resultaten laten zien dat het model beter presteert dan een willekeurige selectie uit de mogelijke oorzaken van een waargenomen fout en dat het de mogelijke oorzaak van de volgende geobserveerde fout van een persoon tot meer dan 75% van de gevallen correct voorspelt. We bespreken hoe de methode kan dienen als diagnostische benadering van personalisatie.

Tot slot introduceren we in Hoofdstuk 7 een nieuwe idiografische benadering voor de ontwikkeling van intelligentie, met grote theoretische implicaties. In de afgelopen eeuw zijn ver-

schillende formele modellen voor intelligentie voorgesteld, waaronder het dominante *g*-factor-model, de herontdekte *sampling*-theorie en het recente *multiplier-effect*-model en *mutualisme*-model. Wij stellen een nieuw idiografisch model voor, waarin we intelligentie conceptualiseren als evoluerende netwerken. Het statische model, een uitbreiding van het Fortuin-Kasteleyn-model, biedt een zuinige verklaring voor de *positive manifold* en de hiërarchische factorstructuur van intelligentie. Bovendien laten we zien hoe het model het Matthew-effect in ontwikkelingsstadia verklaart, en stellen we een methode voor om groeidynamica te bestuderen. Belangrijk is dat dit hoofdstuk een theoretisch kader biedt voor het begrijpen van aanpassingen aan de educatieve sequentie en de invloed ervan op de cognitieve ontwikkeling.

Dankwoord

Dit proefschrift schreef ik niet alleen. Om te beginnen verraden de co-auteurs van de verschillende hoofdstukken al diverse samenwerkingen. Bovendien is een promotie meer dan het boekje. Ik heb het getroffen wetenschappelijk op te mogen groeien aan het sterke instituut Psychologie van de UvA—in het bijzonder aan de fenomenale afdeling Psychologische Methodenleer—en bijdragen te mogen leveren aan onderwijs en de bredere academische gemeenschap. Ik ben iedereen die ik de afgelopen jaren daarbij heb leren kennen dankbaar voor deze leerzame en te gekke periode.

Gunter en Han, ik had me geen betere promotoren kunnen wensen. Niet alleen als individuen, maar ook als team. Jullie tellen bij elkaar op; jullie middelen elkaar niet uit. Ik hoop nog vaak producten van jullie samenwerking te zien en blijf er natuurlijk graag bij betrokken.

Gunter, wij waren denk ik niet de meest voor de hand liggende match (kuchmathematischestatistiek), maar toch schonk je mij het vertrouwen met een plekje op je beurs. Voor mij bleek je de ultieme latente factor. Ongrijpbaar, maar met een enorme invloed op mijn denken, dat zich duidelijk manifesteert door het gehele proefschrift. In gesprekken die we hadden beheerste je de kunst om de eenvoud van onmogelijk complexe problemen bloot te leggen. Je leerde me denken in modellen, patronen te herkennen door de gehele wetenschap, zowel in tijd als ruimte, en introduceerde me in fundamentele theorievorming. Voor dit alles ben ik je ontzettend dankbaar.

Han, tja, waar moet ik beginnen. Ik heb een grote waardering voor je. Ik bewonder je creativiteit, je pragmatisme, je rebellie. Mede dankzij jou is PML de eclectische en vooruitstrevende afdeling die het is, en mede dankzij jou heb ik er mogen opgroeien. Je was geen latente factor maar een hub in een netwerk: je betrokkenheid bij alle verschillende facetten van de academische gemeenschap is inspirerend. En je rijkdom aan ideeën, snel inzicht in—en advies bij—ieder mogelijk probleem, het vele divergeren maar tijdig convergeren en bovenal het gegeven vertrouwen, hebben enorm bijgedragen aan dit proefschrift. Het spontaan lasergamen, de potjes tafeltennis, het dansen op de feestjes, en de minivakantie in Finland, zal ik bovendien ook allemaal niet snel vergeten.

Abe, soms weet je vanaf het eerste moment dat je iemand ontmoet dat je die persoon graag mag. Je loopt een paar stappen op me voor en ik kon je spoor dus mooi volgen. Je bent relaxed, weet goed wat je belangrijk vindt in het leven, handelt daar naar en bent bovendien veel te bescheiden voor je theta. Ik genoot van onze gesprekken; over onderwijs, onderzoek en alles daarbuiten. Samen het statistiekvak ontwerpen en geven (een punt aftrek als studenten de p-waarde niet konden uitleggen, en dan nog ging het soms mis...), de onderwijsdagen, tafeltennissen (met eigen rubbers), matig voetballen, Finland, de spreekspleet tussen onze beeldschermen, de onderwijsdagen, ITGWO[2] (bijna ITGWO[3]), Hamburg, en, zo, voort. Het was allemaal te gek!

Sacha, kamergenoot van het eerste uur en methodologiewinkelmaatje, ik vind het tof om te zien hoe je je plek hebt gevonden. Mede dankzij jou ontgroeit de netwerkbenadering in de Psychologie langzaam de rebellenfase en wordt het een volwassen vakgebied. Joost en Alexandra, fijn dat ik met jullie de kamer mocht delen gedurende de tweede helft van mijn promotie. De monday-morning-meetings, onze gezamenlijke plant van 80 euro die het hooguit 8 dagen heeft gered, en natuurlijk onze NYC loft; ik denk er met veel plezier aan terug.

Robert en Mats, de Coffee Company crew en eigenlijk gewoon officieuze kamergenoten. Dit boekje had er niet gelegen zonder onze vele vrijdagen in verschillende koffietentjes, met CC Java in het bijzonder. Het bracht afwisseling met de UvA, fatsoenlijke koffie en goed kunnen werken zonder afgeleid te worden. We bespraken ons onderzoek en wisselden ervaringen uit vanuit drie verschillende universiteiten. Maar veel belangrijker dan dat alles: de vele avonden basketbal en de fijne gesprekken over de leven. Bedankt jongens. En tot vrijdag!

Frederik, Matthieu en Marjan, we kwamen elkaar tegen op het Creative Industries project. Het heeft veel verschillende projecten opgeleverd, en een erg mooi gezamenlijk stuk! Timo en Peter, geweldig dat jullie zijdelings vanuit Cito ook bij enkele projecten betrokken waren. Brenda, Annemarie, Nienke, Maria en Maarten, ik ben blij dat ik met jullie op verschillende andere projecten heb mogen samenwerken. Brenda, ik raakte betrokken bij je onderzoek en vond het erg prettig met je samenwerken. En we hebben het file-drawer-probleem verslagen! Annemarie, te gek dat we in je cursus konden experimenteren en fijn om met je samen te werken. Ik ben wel gaan twijfelen of spaced practice ook in onderzoek effectief is ;-). Nienke, ik heb je leren kennen als harde werker en als kritische geest, tof om met jou de eerste Oefenweb A/B test te publiceren. En mooi dat we vanuit de inspectie alweer aan het volgende project begonnen zijn! Maria, je vaardigheden komen uitstekend van pas op het misconcepties project; erg leuk dat je op de valreep aanhaakte. En Maarten, ik ben ontzettend blij met onze samenwerking op het meest ambitieuze project uit dit proefschrift. Het duurde even voordat we elkaar vonden, maar

het heeft iets moois opgeleverd. Ik waardeer je enorm, zowel in de samenwerking als daarbuiten.

Joseph and Benjamin, I really enjoyed our collaborations. Joseph, being possibly the first in Europe to work on A/B tests in online learning, it was great to discover your pioneering work in the US. Although the world did not seem to be ready for our story back then, I have since seen several papers that tell it, and I am happy that it has finally started to take off. I enjoyed our discussions and your visit, and still enjoy the great papers you keep writing. Benjamin, I am fortunate to have had the opportunity to meet you and very much enjoyed the Iowa visit. I hope our misconception collaboration is not our last. Also, you are probably my most international colleague; Iowa, New York, Amsterdam, where will we meet next? Last but certainly not least thanks to everyone at ACT Next. You were the most welcoming and gifted group I could have hoped for!

Oefenweb, en in het bijzonder Marthe, Mischa, Marin en Mark, zonder jullie was dit proefschrift half zo dik. De toegevoegde waarde van de verbintenis van Oefenweb met de wetenschap is evident en staat veelvuldig in dit proefschrift beschreven. Met de verschillende uitdagingen die er zijn, zoals de AVG en concurrenten die vooral investeren in gelikte marketing, is het bewonderenswaardig dat jullie trouw blijven aan jullie wetenschappelijke wortels. Ik ben er van overtuigd dat de verschillende samenwerkingen bijdragen hebben geleverd aan kennis over leren, en hoop dat het de vele leerlingen in binnen- en inmiddels ook buitenland bovendien een steeds betere leerervaring oplevert.

Onderwijs De docenten van vandaag leggen de basis voor de wetenschappers van morgen. Het was te gek om daar de afgelopen jaren aan bij te mogen dragen. Mariska en Denny, dank voor jullie inspirerende lessen. Jullie waren de beste voorbeelden die ik me kon wensen en de eersten aan wie ik denk als ik op zoek ben naar een goed idee.

Max, ik had niet durven hopen dat ik samen met mijn favoriete honours-docent een vak zou gaan ontwikkelen. Dank voor het vertrouwen, alles wat ik van je heb mogen leren en afkijken, de leuke lessen en de fijne gesprekken. En nu worden we ook nog eens IIS collega's! Joost, Jolanda en Vera, mede dankzij jullie was het vak zo'n succes, heel leuk om het met jullie vorm te geven.

Rifka, bedankt voor het begeleiden van het BKO traject en de gezellige praatjes in de pantry. En tot slot natuurlijk dank aan alle studenten, die me inspireerden of het leven zuur maakten, zonder jullie is er natuurlijk geen reet aan.

Gemeenschap Waar de docenten de basis leggen voor morgen, is de academische gemeenschap de bemesting van vandaag. Collega's van PML, de hardst werkende én meest sociale

afdeling van het instituut, jullie waren te gek. De PML feestjes zal ik niet snel vergeten. Lisa, Lisanne, Marie, Riet, Jonas, Ravi, Lotte, Nihayra, Joost, ik ben blij dat ik samen met jullie de promotieperiode heb doorlopen en ben dankbaar voor het fijne contact dat we hebben en hadden. Collega's van IOPS, en in het bijzonder van MTO Tilburg, het was tof jullie regelmatig te treffen en te zien wat er in de rest van de Lage Landen gebeurt. De avonden en Airbnb's waren onovertroffen (ik noem een Wellnesshuis in Enschede).

Agneta, je weet volgens mij als geen ander dat de academische gemeenschap actief vormgegeven en ondersteund moet worden. De ruimte die je daarvoor biedt draagt denk ik bij aan het succes van het instituut. Marco, bedankt voor de goede raad en inspirerende gesprekken ten aanzien van promovendi en beleid. En Daniela, ik vond het niet alleen heel leuk om ons ideetje voor een dialoog over het data-opslag-protocol uit te voeren, maar was ook ontzettend blij met je als collega-docent en later als collega-promovendus en collega-Psaiko.

Jonas, toen we tijdens de koffie in het Bakhuys fantaseerden over een kruisbestuivingsfestival konden we nog niet bevroeden dat we het een jaar later (en ongeveer 365 potjes tafeltennis verder) zouden organiseren. Ik zal niet snel vergeten hoe er de dagen na het festival door verschillende collega's de eenheid van het instituut werd benadrukt. Ook de instituutsbrede tafeltennisladder hebben we voor elkaar gekregen, leuk! Lisa en Lisanne, zo leuk dat jullie ook nog bij de festivalorganisatie aanhaakten.

Alle Psaiko's, bedankt voor het geweldige werk dat jullie verzetten, dat niet door iedereen wordt gezien en waar je proefschrift ook niet per se beter van wordt. Lisa, je weet hoe blij ik was dat je het stokje van voorzitter wilde overnemen. Dat was het moment dat ik zeker wist dat Psaiko in goede handen was.

Thomas, ons ludieke idee voor het Spinaziecentrum heeft ook voor bemesting gezorgd. In ieder geval in vrij letterlijke zin. Ik had me geen betere compagnon kunnen voorstellen. Gaaf dat dankzij het enthousiasme van veel collega's het initiatief verder is gegroeid en hopelijk zal leiden tot de kruisbestuiving die we vanaf het begin voor ogen hadden.

Vrienden & familie Lieve vrienden, wat is het toch fijn om na een dag in abstractie te hebben geleefd een borrel te kunnen drinken, een wandeling te kunnen maken, een 6c te kunnen klimmen, een potje basketbal te kunnen spelen, Mia op te kunnen zoeken, een matig gesprek te kunnen voeren, te kunnen EenNullen, een feest, concert, of voorstelling te kunnen bezoeken, of lekker te kunnen eten. En wat is het toch heerlijk om dat met júllie gedaan te kunnen hebben. Overdag geven jullie me moed, 's avonds en in het weekend geluk.

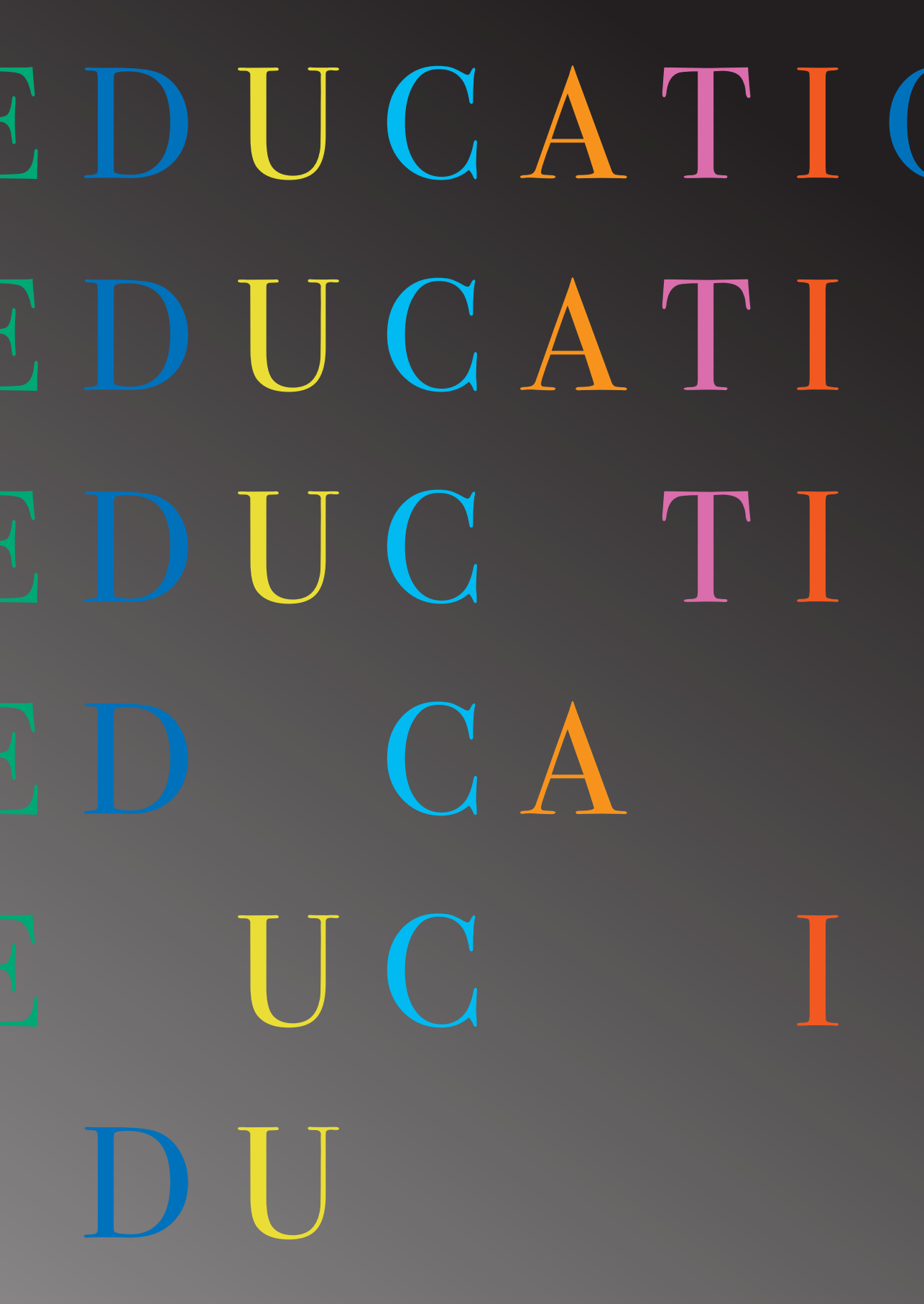
Lieve pa en ma, jullie hebben me altijd de ruimte gegeven me te ontwikkelen in de richting

die bij me past, en jullie hebben de bodem gelegd waarop ik een open en kritische geest kon ontwikkelen. Daarvoor ben ik jullie ontzettend dankbaar! Het is de reden dat ik dit proefschrift aan jullie opdraag.

Lieve Clarent, als broer heb je de wereld altijd alvast een beetje voor me verkend. Het voelde verbonden om de afgelopen jaren tegelijkertijd iets op te bouwen: ik ben trots op hoe je op eigen kracht en eigen wijze Hempje hebt ontwikkeld en bewonder hoe je, samen met Nina, Mia kennis laat maken met de wereld. Je bent me dierbaar.

Lieve Nina, je kent mij met name uit de periode als promovendus, een pak waarvan de broek direct als gegoten zat, maar waarvan het jasje nog op maat gesneden moest worden. Je bent op belangrijke momenten een enorme steun geweest. En je doet het werk waar ik slechts over theoretiseer: je intervenieert in jonge levens en verlegt hun pad naar een betere toekomst. Je inspireert me, boven alles als mens. Ik ben gek op je en hou van je.

Notes

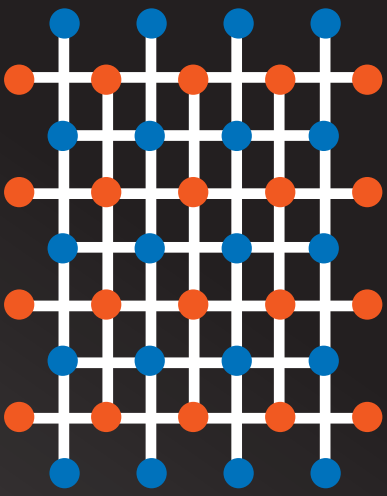

EDUCATIO
EDUCATI
EDUC TI
ED CA
E UC I
DU

Need a break?

On both flaps you'll find some completely unrelated effortful practice, in the form of graph games. Each of these games requires you to find an opponent, and none of them can end in a draw. Chapter 7 shows you why graphs, or networks, are not only fun, but can also be powerful abstract representations of pretty much everything. So it's not entirely unrelated after all…

Bridgit

Find an opponent. Pick one color each (blue or red). Take turns in coloring the link between two adjacent nodes of your color. Edges may not cross. You win as soon as you create a chain from top to bottom (blue player) or left to right (red player).



By David Gale.

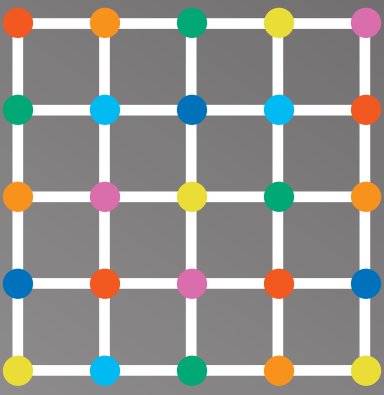
Sim

Find an opponent. Pick one color each. Take turns in coloring the link between any two adjacent nodes. You lose as soon as you create a triangle of your color.

By Gustavus J. Simmons.

Crosscram

Find an opponent. Pick one color each. Take turns in coloring the link between any two adjacent nodes. The first player may only create vertical links, the second only horizontal links. Nodes can only have one link. The first player unable to make a move loses.



By Göran Andersson.