UNIVERSITY OF CALIFORNIA, SAN DIEGO

Fractional Recall: Understanding Forgetting of Mathematics Learning from Computer Assisted Instruction Data and Implications for Personalized Learning in Low Resource Contexts

A Dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive Science

by

Richard Tibbles

Committee in charge:

Professor Terry Jernigan, Chair
Professor Jeffrey Elman
Professor James Hollan
Professor David Kirsh
Professor Scott Klemmer
Professor Harold Pashler

2017

Copyright
Richard Tibbles, 2017
All rights reserved.

The Dissertation of Richard Tibbles is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

DEDICATION

To Stephanie.

For our love that bore this trial and more, and continues to grow.

EPIGRAPH

The presence of those seeking the truth is infinitely to be preferred to the presence of those who think they've found it.

Terry Pratchett (*Monstrous Regiment*) —

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables xiv

Acknowledgements xviii

Vita xxi

Abstract of the Dissertation xxii

Chapter 1 Introduction 1
1.1 Computer Assisted Instruction 1
1.1.1 Fall of MOOCs, Rise of PCOCs 2
1.2 Personalizing Learning 2
1.3 Kolibri: Computer Assisted Instruction in Low Resource Contexts 7
1.3.1 The Digital Divide 7
1.4 Designing Learning for Everyone 10
1.5 Dynamics of Learning Retention Over Time 12
1.5.1 The Testing Effect 12
1.5.2 Forgetting Over Time 39
1.5.3 Spacing, Massing, and Optimizing Review Schedules 78
1.5.4 Interleaving against Blocking Instruction 82

Chapter 2 Mathematics Learning and Computer Assisted Instruction . . . 86
2.1 Mathematics Learning 86
2.2 Current State of Computer Assisted Instruction for Mathematics 86
2.3 Primary Data Source: Khan Academy 87
2.3.1 User Stories 89
2.4 Secondary Data Source 96
2.4.1 KA Lite 96

Chapter 3 Forgetting in Declarative Learning 99
3.1 Introduction 99
3.2 Data Selection 100
3.3 Single Fact Data 102
3.4 Retention Curves 108
3.5 Interference Effects 115
3.5.1 Summary 119

Chapter 4 Forgetting in Mathematics Learning 121
4.1 Introduction 121
4.2 Mathematics Exercises in Khan Academy 123
4.3 Data Selection 124
4.4 Item Level Analysis 125
4.4.1 Item Level Data for Adding and Subtracting Fractions with Like Denominators Word Problems . . . 125
4.4.2 Retention Curves 130
4.4.3 Interference Effects 133
4.4.4 Replication in KA Lite Data 136
4.5 Exercise Level Analysis 141
4.5.1 Exercise Level Data for Adding and Subtracting Fractions with Like Denominators Word Problems . . . 141
4.5.2 Retention Curves 143
4.5.3 Signs of Consolidation 149
4.5.4 Replication in KA Lite Data 150
4.6 Summary 161

Chapter 5 Spacing in Mathematics Learning 162
5.1 All Time Intervals 163
5.2 Long Intervals 166
5.3 Summary 166

Chapter 6 The Impact of Forgetting of Prerequisite Knowledge on Subsequent Learning 169
6.1 Introduction 169
6.2 Prerequisite Skills in Khan Academy 170
6.3 Impact of Forgetting of Prerequisite Skill Learning 170

Chapter 7 Implications for Design for Computer Assisted Instruction Systems in Low Resource Contexts 173
7.1 Introduction 173
7.2 Design Constraints for Kolibri 174
7.3 Personalized Review 175
7.3.1 High Data Content Libraries 175
7.3.2 Low Data Content Libraries 176

7.4 Learning Analytics 178
7.4.1 Introduction to Learning Analytics 178
7.4.2 Learner Analytics in Kolibri 181

Appendix A Supplementary Tables and Figures 186
A.1 Declarative Fact Learning 186
A.2 Mathematics Learning 186

Bibliography 192

List of Figures

1.1 Response Curves for Exponential and Power models of retention . 56

1.2 Fit of episodic memory density over time calculated from Crovitz and Schiffman (1974) 62

1.3 Spaced vs Massed Learning regimes 79

2.1 Infographic of the Khan Academy Mastery model in 2014 (taken from Faus (2014)) 90

2.2 One example question from the Khan Academy 'Recognize fractions 1' exercise 97

3.1 Histogram of user/question engagements in declarative fact learning data 102

3.2 Distribution of declarative fact learning test-retest pairs across questions 104

3.3 Distribution of declarative fact learning test-retest pairs across exercises 105

3.4 Distribution of declarative fact learning test-retest pairs across users 106

3.5 Distribution of declarative fact learning test-retest time intervals 107

3.6 Scatter plots of aggregated declarative fact learning data points at bin sizes from 100 to 1000 109
3.7 Declarative fact learning candidate function fits for a bin size of 100 112
3.8 Declarative fact learning candidate function fits for a bin size of 100, subsampled to give a ratio of 3:2 short:long interval test-retest data points 114
3.9 Declarative fact learning candidate function mean R^2, on a per question basis. 116
3.10 Declarative fact learning candidate function fits for a bin size of 100, on a per question basis for fits with R^2 in excess of 0.5. . . . 117

4.1 Histogram of user/question engagements in Adding and Subtracting Fractions with Like Denominators Word Problems data. 127
4.2 Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems test-retest pairs across questions . . . 128
4.3 Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems test-retest pairs across users 129
4.4 Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems test-retest time intervals 130
4.5 Scatter plots of aggregated Adding and Subtracting Fractions with Like Denominators Word Problems data points at bin sizes from 100 to 1000 131
4.6 Adding and Subtracting Fractions with Like Denominators Word Problems candidate function fits for a bin size of 100 133

4.7 Adding and Subtracting Fractions with Like Denominators Word Problems candidate function mean R^2, on a per question basis. . 134
4.8 Adding and Subtracting Fractions with Like Denominators Word Problems candidate function fits for a bin size of 50, on a per question basis for fits with R^2 in excess of 0.5. 135
4.9 Distribution of KA Lite test-retest pairs across users 137
4.10 Histogram of user/question engagements in KA Lite data. 138
4.11 Distribution of KA Lite test-retest time intervals 139
4.12 Scatter plots of aggregated KA Lite data points at bin sizes from 100 to 1000 140
4.13 Histogram of user/exercise engagements in Adding and Subtracting Fractions with Like Denominators Word Problems data. 142
4.14 Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems exercise test-retest time intervals . . 144
4.15 Scatter plots of aggregated Adding and Subtracting Fractions with Like Denominators Word Problems exercise data points at bin sizes from 100 to 1000 145
4.16 Adding and Subtracting Fractions with Like Denominators Word Problems exercise candidate function fits for a bin size of 100 . . 146
4.17 Adding and Subtracting Fractions with Like Denominators Word Problems exercise candidate function fits for short time intervals, for a bin size of 100 148

4.18 Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for initially correct test-retest trials for Maths Exercise 151
4.19 Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all trials before mastery has been achieved for Maths Exercise 152
4.20 Histogram of user/exercise engagements in KA Lite data. 153
4.21 Distribution of KA Lite exercise test-retest time intervals 154
4.22 Scatter plots of aggregated KA Lite exercise data points at bin sizes from 100 to 1000 155
4.23 KA Lite exercise candidate function fits for a bin size of 100 . . . 156
4.24 KA Lite exercise candidate function fits for short time intervals, for a bin size of 100 157
4.25 Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for initially correct test-retest trials for KA Lite 159
4.26 Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all trials before mastery has been achieved for KA Lite 160

6.1 Scatter plots of aggregated Adding and Subtracting Fractions with Like Denominators Word Problems exercise data points at bin sizes from 100 to 1000 171

6.2 Adding and Subtracting Fractions with Like Denominators Word Problems candidate function fits for a bin size of 100 172

7.1 KA Lite Coach Reports Per Item Detail View 181

7.2 Kolibri Coach Reports Learner Detail View 182

A.1 Declarative fact learning candidate function proportion of varianced explained compared to best fitting model, on a per question basis. 188

A.2 Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all test-retest trials 189

A.3 Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all trials 190

A.4 Histogram of most commonly used exercises in KA Lite data. . . 191

List of Tables

1.1 Table of candidate functions to fit empirical forgetting data 52

1.2 Levels of Mathematic Rehearsal for Different Levels of Class Taken, extract from Table 2 of Bahrick and Hall (1991) 65

3.1 R^2 values for different bin sizes for aggregating pure declarative data across all users and questions 111

3.2 R^2 values for different ratios of short to long data, aggregated at a bin size of 100, for pure declarative data across all users and questions 113

3.3 R^2 values for different bin sizes for aggregating pure declarative data across all users and questions, fitted only against a holdout data set using interference measures as regressor variables. 118

3.4 R^2 values for different bin sizes for aggregating pure declarative data across all users and questions, fitted against a holdout data set using log-time and interference measures as regressor variables. . . 119

3.5 R^2 values for different bin sizes for aggregating pure declarative data across all users and questions, fitted against a holdout data set using only log-time as a regressor variable. 119

4.1 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions 132
4.2 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set using log-time and exercise interference as a regressor variable. 134
4.3 R^2 values for different bin sizes for aggregating KA Lite data across all users and questions 141
4.4 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users by exercise 144
4.5 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users by exercise for short time intervals 147
4.6 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users by exercise for long time intervals 147
4.7 R^2 values for different bin sizes for aggregating KA Lite data across all users by exercise 156
4.8 R^2 values for different bin sizes for aggregating KA Lite data across all users by exercise for short time intervals 157
4.9 R^2 values for different bin sizes for aggregating KA Lite data across all users by exercise for long time intervals 158

4.10 R^2 values for different bin sizes for aggregating KA Lite data across all users (excluding those in a particular structured school environment) by exercise 161

5.1 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set. . . 165

5.2 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set using fixed interval. 165

5.3 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set using expanding interval. 165

5.4 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions for retention intervals greater than 24 hours, fitted against a holdout data set. 166

5.5 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions for retention intervals greater than 24 hours, fitted against a holdout data set using fixed interval. 167

5.6 R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions for retention intervals greater than 24 hours, fitted against a holdout data set using expanding interval. . 167

6.1 R^2 values for different bin sizes for aggregating Adding And Subtracting Fractions With Like Denominators Word Problems Prerequisites data across all users 172

A.1 R^2 values for different candidate fit functions, aggregated at a bin size of 100, for pure declarative data across all users fitted per question 187

ACKNOWLEDGEMENTS

Acknowledgements are due to many. Firstly, the Cognitive Science department, and particularly my adviser, Terry Jernigan, for accepting me into the program, and then giving me the latitude and freedom to pursue my interests - while still steering me clear of the rocks when I paid too much heed to the siren's call. Many thanks to the Center for Human Development, for giving me an academic home in my time at UC San Diego, and for the many people there that I worked alongside.

Thanks are also due to the Temporal Dynamics of Learning Center, for giving me many opportunities to broaden my perspective on research in Learning Sciences, allowing me to see the full breadth and depth of the field in which I found myself immersed (and frequently ignorant of many aspects). Further, the funding they provided through their small grants programme laid the foundations for the preliminary work that would eventually become this thesis.

Additionally, I would like to thank Khan Academy and all its users for the data that I used for much of this dissertation - it is an invaluable trove of learning data, of unusual size and variety, which I am sure will provide many more insights in the future.

Jamie Alexandre has been integral to my time in graduate school, whom I met (along with Jeremy Karnowski) before I had even officially begun the programme. With them I relearned to program, this time in Python, and rediscovered a part of myself that I had let lie fallow for many years. As counter points to myself, I could ask for no better than Jamie and Jeremy, where I am laconic and reserved, they are enthusiastic and outgoing. With both, I found a shared love and passion for education, and I was very glad to have the opportunity with Jamie and many others to found

Learning Equality that became so important to my time in graduate school, and is now fundamental to my time beyond it.

Andy Alexander, Conor Frye, Marybel Robledo, and Megan Bardolph, as my remaining PhD cohort have provided much support (in wildly varying forms) and my time would have been much less rich without them. The collegiate nature of the department has meant that I have also been fortunate to have the friendship and support of many others, including Adam Mekrut, Melissa Troyer, Laura Shelly, Matt Schalles, and many more.

Learning Equality has given me a home and a purpose as I wrestled with all the usual existential angst of a PhD, and all the people involved, past and present, have given me continual hope and inspiration that determined and good people working together can help make a small contribution for the betterment of our world. Elizabeth Vu became a friend through our shared commitment to using educational technology to improve education, and her friendship and support has been invaluable. Ben Cippolini showed me how much it was possible for one person to work, and I have taken his example as both an inspiration and a warning. Dylan Barth continually surprised me with his wisdom and determination, and then always by how young he really was as well. It has been a pleasure building Learning Equality with all of my colleagues there, and I look forward to the bright future that it holds for all of us, and the world.

My family have shaped me in ways that I can never understand, but see in myself at every turn. My late mother, Beth, without whose influence I would never have strived so hard academically, my father, Tony, whose work ethic and persistence I hope that through my emulation I do even a sliver of justice to. My brother, Alex, who weaned me off physical fights in favour of verbal argumentation, a pattern that

would eventually transmute into our shared interest in Philosophy, is a person that I have always consciously tried not to emulate, but have, through my actions, done anything but. My sister-in-law, Bianca, whom I have known nearly since I first started higher education, and whose depth of thought, empathy, and concern for others has been a support I could hardly be without.

Last, and by no means least, Stephanie, whom I have known but three years, but with whom I have a depth of shared experience that almost feels like a life time. Without her support I could never have finished this journey, and the love and care that she has shown me and Bella, my dear cat.

VITA

2004	M. A. Physics and Philosophy, Brasenose College, University of Oxford
2006	PGCE Secondary Science: Physics, Institute of Education, University of London
2009	M. A. Philosophy, Birkbeck College, University of London
2017	Ph.D. Cognitive Science, University of California, San Diego

PUBLICATIONS

Tibbles, R. (2015). Exploring the Impact of Spacing in Mathematics Learning through Data Mining. In Proceedings of the 8th International Conference on Educational Data Mining. Madrid, Spain: Educational Data Mining Society.

Tibbles, R., & Alexandre, J. (2014). KA Lite: Implementation and Research in the Offline Learning Revolution. In Proceedings of MOOCs4D: Potential at the Bottom of the Pyramid. University of Pennsylvania, Philadelphia, USA.

TEACHING EXPERIENCE

Winter, Fall 2016	Instructor, Hands on Computing: Introduction to Robotics, Cognitive Science Department, UC San Diego.
2015 - 2016	Senior Teaching Assistant, Cognitive Science Department, UC San Diego.
Summer 2013	Facilitator for faculty workshop on Online and Blended Learning, UC San Diego.
2012 - 2015	Learning Management System Developer/Administrator: Cognitive Neuroscience, UC San Diego.

PROFESSIONAL EXPERIENCE

2013 - Present	Learning Lead and Software Developer - Learning Equality
2009 - 2011	Instructional Strategies Specialist - IDECORP
2006 - 2009	Teacher of Science (Physics) - Fortismere School

ABSTRACT OF THE DISSERTATION

Fractional Recall: Understanding Forgetting of Mathematics Learning from Computer Assisted Instruction Data and Implications for Personalized Learning in Low Resource Contexts

by

Richard Tibbles

Doctor of Philosophy in Cognitive Science

University of California, San Diego, 2017

Professor Terry Jernigan, Chair

Recent and historical work in forgetting and spacing of learning across long time scales suggests that which has been learned is easily forgotten. In addition, other strands of work suggest that forcing recall through testing, and temporal spacing of testing episodes can ameliorate the effect of forgetting over time. In order to be able to prevent this forgetting from happening in educational contexts, it is useful to understand how this forgetting and prevention of forgetting occurs outside of the laboratory. Further, few studies have examined the time course of forgetting and

the impact of the ameliorative repeated testing and temporal spacing effects in the domain of Mathematics, an area of school learning in which students frequently fail to progress, and on which depends success for a wide range of other fields. Also, learning of later Mathematics skills is frequently highly dependent on having maintained (i.e. not forgotten) previously learned Mathematics skills.

Data from two computer assisted instruction systems, Khan Academy and KA Lite, are analysed in order to better understand the time course of forgetting. Declarative fact learning data from Khan Academy are used as an initial conceptual replication of previous studies on forgetting. The difference between the learning of declarative facts and Mathematics skills is compared and contrasted, showing that individual Mathematics questions behave similarly to fact learning, but that the learning of Mathematics skills may potentially have differing short and long term trajectories. However, comparisons between the Khan Academy and KA Lite data show the strong influence factors external to the computer assisted instruction systems have on the observed patterns of learning and forgetting that occur within the system.

The implications of these analyses are then discussed, specifically in the context of Kolibri, a computer assisted instruction system designed to provide personalized learning in contexts without Internet access. Implications for scheduled spaced retrieval practice, teacher dashboards and analytics, and future data collection needs for iterative design of the learning experience in the platform are discussed.

Chapter 1

Introduction

1.1 Computer Assisted Instruction

The promise for computers to revolutionize the way that humans learn has been long recognized, with early experiments in computer assisted instruction to teach introductory reading (Atkinson & Hansen, 1966), computer programming (Barr, Beard, & Atkinson, 1975), and foreign languages (Atkinson, 1975). In spite of this early promise, continuing discussion of the potential impact of computer assisted instruction in the classroom (Taylor, 1980), and meta-analyses showing positive impacts of computer assisted instruction (Kulik & Kulik, 1991; Fletcher-Flinn & Gravatt, 1995), it has not been until the advent of widespread adoption of the Internet, and the Internet as a primary delivery platform for computer assisted instruction, that popular and widespread use of computer assisted instruction has become prevalent.

1.1.1 Fall of MOOCs, Rise of PCOCs

The rise to prominence of the MOOC (Massively Open Online Course) providers, Coursera, EdX, Udacity, and Udemy (Koller, 2012), as well as sites like Khan Academy, and Assistments (Turner, Macasek, Nuzzo-Jones, Heffernan, & Koedinger, 2005) that provide lower barrier access to course material (with not all content being initially hidden by a paywall) in a common model for the web has allowed many more people to access computer assisted instruction, increasing the scale of its impact. Coursera, EdX, and Udacity claim 24 million registered users ("Udacity, Coursera and edX Now Claim Over 24 Million Students (EdSurge News)," 2015), with Khan Academy having 10 million unique users per month (Hirasaki, 2014). Unfortunately, recent trends in MOOCs, as identified in Shah (2016), are increasingly towards producing content that is behind a pay wall, and classes that are taken on an 'on-demand' basis, meaning that the MOOC providers are now simply Personal Closed Online Courses (PCOC) providers. However, this presents possibly the greatest promise in Computer Assisted Instruction - personalization.

1.2 Personalizing Learning

One of the most important features of computer assisted instruction, and one of the most frequently cited motivations, is the possibility of personalizing learning for a student with computer assisted instruction, in the same way that an individual human tutor might personalize instruction. Given that paying for a well trained personal tutor for every child is beyond the means of most education systems, attempting to replicate this interaction and hence solve Bloom (1984)'s '2 sigma' problem, is a frequently

cited goal of work in computer assisted instruction and intelligent computerized tutors. While Bloom's '2 sigma' problem might not be solveable with this human intensive approach, as a meta-analysis of subsequent studies of human tutoring have shown an effect size of closer to '0.8 sigma' for human tutoring (VanLehn, 2011), the same meta-analysis found a similar effect size of Intelligent Tutoring Software that broke problems down into steps. However, the development of models of performance in such domains, and the breaking down of every problem into steps is labour intensive and requires considerable human input (Cen, Koedinger, & Junker, 2006). As such, some learning providers, such as Khan Academy, have focused on producing a larger range of content that allow for the input of an answer, and then associated feedback - this kind of 'answer based'(VanLehn, 2011) instruction seems to have a more modest effect size of about 0.3 (VanLehn, 2011) (further, the 'step based' instruction is focused on more procedural content, like Mathematical problem solving and technical skills, so is not applicable to all kinds of learning).

As such, in order to provide a much larger level of improvement in learning, while also being able to scale the breadth and extent of content, it is necessary to understand how individual difference variability can impact student learning, whether that is variability in cognitive states, emotional or motivational states, current level of knowledge, or history of interaction with material. This, by no means exhaustive, list shows the wide range of factors that need to be investigated in order to more effectively personalize learning across a range of domains. The area that seems most promising, and that has received a great deal of attention, is in estimating the learner's current level of knowledge, in order to present a learner with activities that present a 'desirable difficulty' (Bjork, 1994; Bjork & Bjork, 2011) for the learner. While

there are many ways that Bjork and Bjork (2011) detail of creating these desirable difficulties for a learner, such as changing the context of study, interleaving study with different items (not blocking study, see subsection 1.5.4), and retrieval, rather than exposure through restudying, the piece of advice that most strongly urges tracking the user's knowledge state over time is this: 'when some skill or knowledge is maximally accessible from memory, little or no learning results from additional instruction or practice' (Bjork & Bjork, 2011), therefore in order to properly time the presentation of learning opportunities to students, it is important to know the current state of a learner's knowledge. This may be accomplished by tracking a learner's interactions with material (whether passively watching, or, more informatively, engaging with interactive exercises that are automatically graded), but also by considering the pattern of interaction and how recently learners have engaged with material. Further, these interactions that assess a learner's knowledge also act as learning events in themselves, due to the widely documented 'the testing effect' (subsection 1.5.1), and some computerized instruction does this quite well, with automated quizzes that require recall, although many still require simple recognition, which reduces the desirable difficulties in recall.

Bjork (1994)'s 'desirable difficulties' can also be seen as the obverse pedagogical consideration to Vygotsky (1978)'s 'Zone of Proximal Development'. Vygotsky's consideration of the different performances that a learner can produce when confronted with tasks with varying levels of scaffolding (which in his examples are generally through a dialogic interaction with the questioner, but could equally be through the intervention of some sort of automated tutor, or by decreasing the level of difficulty of the presentation of the question) correlate with the factors that Bjork outlines that

can vary the difficulty of a task - if a task is made easier, then that may spell the difference between a learner being able to successfully complete the task, and failing to do so. Thus the change in difficulty has acted as scaffolding if it has helped the learner to successfully complete the task, and hence reinforce previous learning. The complement of Bjork's work to Vygotsky, therefore, comes from a recognition that introducing too much scaffolding (for that particular learner, given their knowledge state, and the current context) will not produce the most efficient results in learning - making successful retrieval effortless will allow the student to enjoy success (which may have other motivational benefits), but will not produce benefits for future learning, while getting the difficulty level 'just right'(Southey, 1837) will place learners in the 'Zone of Proximal Development', allowing them to perform a task, or respond to a question, that they otherwise would not be able to do, but also in a way that is sufficiently effortful that it will promote learning, and allow the task to be completed with less scaffolding in future.

In addition to making connections to developmental psychology in Vygotsky, 'desirable difficulties' can also be connected to the dopamine reward prediction error hypothesis (for a full discussion of its historical development and empirical support see Glimcher (2011)). Glimcher (2011) details the evolution of this hypothesis from the beginnings of the application of mathematical models of reinforcement learning by Bush and Mosteller (1951) to explain Pavlov's experiments, on to the development of the Rescorla and Wagner (1972) model, and finally the development of the Temporal Difference model by Sutton and Barto (1998). The Temporal Difference model shifted the focus of these models of learning reward from learning a model that captured previous reward values, to a model whose goal was to predict rationally expected

future reward (i.e. the multiplicative product of the likelihood of the reward, and the size of the reward). Learning occurs then, when there is a difference between the rationally expected reward predicted by the model and that received. If we consider a learner faced with two different ways of answering the same question (which they have already encountered) - one that makes it very easy (by having a multiple choice question where the correct answer has been seen by the learner before), and one that makes it more difficult (by forcing the learner to recall the answer and record it) - in the first, easy way, the very low uncertainty will yield a prediction of reward weighted by this low uncertainty (reward is very likely, and so the rational expectation of the reward is roughly equal to the reward itself), in the second, more difficult way, the higher level of uncertainty associated with a more demanding task means that the rational expectation of reward is tempered by the uncertainty of getting the right answer in this more demanding task, thus the rational expectation is lower. If the difficulty is 'desirable', and the learner still gets the question right in this latter instance, then there will be considerable reward prediction error, as success will be achieved. Further, if we consider a third situation where the difficulty of the question is too high (such as the question is obliquely written, or relies heavily on background knowledge not available to the student) the rationally expected reward will be close to zero; while, in the unlikely case that the student got the question correct, this would produce a large prediction error, the more likely outcome is that the student gets the question wrong, there is no prediction error, and no learning occurs. It seems then, that under the reward prediction error hypothesis, 'desirable difficulties' amount to the adding of sufficient uncertainty to produce a prediction error, while also still allowing the rewarded outcome to occur sufficiently regularly for efficient learning.

1.3 Kolibri: Computer Assisted Instruction in Low Resource Contexts

1.3.1 The Digital Divide

The above trend in the move from MOOCs towards PCOCs is particularly suprising given the recent attestations by Coursera (Zhenghao et al., 2015) in the Havard Business Review, that the people most benefiting from Online Courses provided by Coursera are actually those in the developing world with the claim that 'Those with low socioeconomic status from non-OECD countries are most likely to report benefits'(Zhenghao et al., 2015), whom it would seem are least able to take advantage of online course material, and most likely to have more difficulty to afford the paid certifications that more courses are being paywalled behind. However, while the survey data itself is not available, a closer look at a presentation (Coursera, 2015) containing some of the survey results reveals a strong response bias in the survey data on which these reports are based. Out of 780000 Coursera users who had completed a Coursera course at the date the survey was sent out, only 51954 responded to the request for data - no attempt to counterbalance for response bias is evident in any of the analyses. In addition 83% of respondents had a bachelor's degree or higher, compared to 73% of typical learners in the MITx and Harvardx courses from 2012 to 2016 (Chuang & Ho, 2016). In addition, rates of reporting of 'some tangible benefit' were at most 42%, for people from non-OECD countries already with a Bachelor's degree. Further, while financial aid is available to take Coursera courses ("Apply for Financial Aid," 2017), it is an additional barrier, and hardly the utopian dream that originally launched the MOOC movement of open access for all (Koller, 2012).

As this partial flowering of the promise of computer assisted instruction for all attenuates, barriers to the remaining open materials are still problematic, not least because of the digital divide. Alexandre (2014) describes the stark contrast between those already with access to both education and Internet connectivity - able to access, but in much less need of, the online learning revolution that the MOOC providers and Khan Academy are facilitating. One first attempt to bridge this digital divide and bring the benefits of computer assisted instruction to those without reliable Internet connectivity is KA Lite, a platform that brings access to Khan Academy videos and interactive exercises, through a web type experience (Alexandre, 2014; Tibbles & Alexandre, 2014). In addition to access to content, KA Lite provides progress tracking, coach reporting tools, content recommendation, support for multiple languages, and simple gamification to encourage student participation. This initial foray into bringing 'online learning' to the 60% of the world who lack Internet access, and the many more for whom access is too slow or unreliable to effectively use online learning tools, has brought Khan Academy learning materials to an estimated 3.4 million learners. KA Lite is developed and maintained by Learning Equality, a non-profit located on the University of California, San Diego campus, and of which I am a co-founder and core team member.

Through iterative development and user feedback, Learning Equality has identified a set of critical needs that current solutions do not yet meet. Chief amongst these is a need to be able to curate and align content to match local curricular standards and objectives, in order to support maximally effective implementation, especially in schools. In addition, there is strong demand to be able to bring in new sources of content beyond Khan Academy, from other online content repositories, as

well as content that is created locally to meet needs that are specific to the cultural and curricular context. To meet these needs, and more, Learning Equality is currently developing a next-generation offline educational content platform, Kolibri.

Kolibri includes both online and offline components: an online content repository and curation interface, through which content creators can create custom content "channels" that can then be synced down and installed into the offline Kolibri application, which will be able to run on a wide variety of hardware (Windows, Linux, Android, Chromebook, and OS X devices). Usage data is stored locally in the application, but can then be synced - via any combination of low-bandwidth connections, peer-to-peer connections with other devices running Kolibri, or portable storage media over the 'sneakernet' - back to a central cloud server for aggregation and reporting. Syncing of new content and software updates can also happen via any of these three methods, enabling seamless offline and low-bandwidth sharing, extending the limits of the Internet to reach those on the far side of the Last Mile.

As Kolibri is focused on enabling access to high-quality educational opportunities for those around the world who currently have the least access and are the most socio-economically disadvantaged. Kolibri is built and continues to be developed for the "lowest common denominator" for resources, meaning it adheres to the following technical constraints:

- Be able to run completely offline, leveraging Internet opportunistically as available.
- Run on a diverse range of low-cost and legacy hardware, to reduce cost barriers
- Be simple to install for non-technical users, reducing barriers to grassroots adoption

- Support a diverse set of languages, to enable access in the local language of instruction

In addition, in order to provide not just access to online resources, but to allow for the impactful personalized instruction that is afforded by computer assisted instruction, Kolibri will make the following features available:

- Content recommendation, utilizing a large library of content, will give teachers and students the flexibility to engage at their own pace with content, engaging with content at the right level of difficulty, at the right time, in order to improve the efficiency of learning, and allowing students to learn more. As the content available on Kolibri grows, this will allow for greater exploration of material, as well as more targeted intervention by teachers and content recommendation to improve student outcomes.

- Coach reports, by providing tools to support blended learning practices in the classroom, teachers will be able to quickly diagnose student performance and intervene with struggling or bored students, while students are provided with rapid, repeated feedback through engaging with the interactive exercises.

1.4 Designing Learning for Everyone

Kolibri allows for the creation of packages of instructional content (channels) that will be used across a huge array of cultural, linguistic, and national contexts - embracing a huge diversity of learners, and content. In addition, the platform will predominantly be used in entirely offline settings, with highly intermittent or no access to the Internet. This means that one of the major advantages of the 'web scale'

computer assisted instruction - the use of machine learning methods on large data sets - is not available immediately to the Kolibri software.

When learning happens in MOOCs/PCOCs, Khan Academy, and other large scale, web based instructional platforms - the data is aggregated and accumulated at a central point, making it amenable to analysis. This allows the use of the large data sets that are accrued to be analyzed to create better models of student learning, and even for course providers to actively collect data through rapid testing and deployment of different alternatives in instruction (A/B testing). These large data sets are then amenable to the application of machine learning methods, which, due to the scale of the data being collected, allow the observation of regularities that would not be observable on a smaller scale. This scale of instruction allows for systematic and rigorous enhancement of instruction through data analysis.

In contrast to the engineering of online learning platforms, therefore, Kolibri faces far more challenging design constraints - users (learners and teachers) who are frequently unfamiliar with computer technology, a wide range of cultural and linguistic contexts, and a proliferation of small pockets of data that are the only locally available resource to adaptively personalize the learning experience for students.

In order to design effective computer assisted instruction in these low resource contexts, it is necessary to make the most of data analysis on larger data sets for optimizing the learning process when available, but also providing the capability to fall back to more local data sets and constraints when it is not.

1.5 Dynamics of Learning Retention Over Time

Learning efficiently is one of the main drivers of personalized instruction. By ensuring that students engage with material only for as long as they need to in order to master it, intelligent instruction can push students further in less time, allowing outcomes to be improved more rapidly, and also to reduce the risk of boredom and loss of motivation. In addition, retention over longer time scales is important to the goals of Education as a whole. While the old adage "Education is what is left once what is learned has been forgotten" is oft quoted, in many Educational contexts, and in particular Mathematics, the necessity of prerequisite knowledge for learning higher order material means that such forgetting is far less desirable.

1.5.1 The Testing Effect

What Is A Test?

In the sphere of the testing effect, a test is anything that prompts some attempt at recall of the item that is being tested. Different tests will induce recall with different amounts of difficult. At the most difficult end of the spectrum is a free recall test, where a student is asked to remember everything they can about a particular topic - whether it is words memorized on a word list, the facts contained in some paragraphs they read, or the contents of a video that they watched.

Cued recall tests provide a prompt for each item to be recalled, for a particular fact learned from reading some paragraphs of material, this might be a cueing question. In a paired associates task (such as learning the mapping from English to Spanish vocabulary words), the cue would be one half of the pair that has been learned (such

as the English word to prompt recall of the Spanish word).

The easiest recall difficulty is found in recognition tasks - where the participant is either presented with multiple options and asked to pick which one they recall as being the right choice (such as in a multiple choice question in education settings), or the participant is shown an item and asked if it is one of the set that they were asked to learn. The latter is often used in word list memorization, where participants are presented with a sequence of learned and non-learned words and asked to identify if they were words that they had previously been asked to learn.

Early Work

The 'testing effect' has been described extensively in the literature, going back, seemingly, as far as 1909. Abott (1909) first described experiments comparing simply engaging in memorization through study, with engaging in recall of the material to be studied. Abott uses the word 'Einprägung' to describe memorization through study, a word that literally translates as 'imprinting'("einprägen - Wiktionary"), but which Abott describes as having a more nuanced meaning of 'the process of attending to sensory materials with a view to remembering it'(Abott, 1909), in the subsequent literature on the testing effect this has been designated as 'studying'.

This comparison gives the result that broadly describes the testing effect, engaging with recall, rather than simply Einprägung, results in better performance on subsequent recall tests of the material. Unlike much subsequent investigation of the testing effect, however, the recall that participants engaged in was undirected and unprompted on individual items - participants were simply asked to rehearse the previously studied materials in whatever way seemed best to them. This may

explain the additional finding of Abott that longer delays between initial study and subsequent recall resulted in an attenuation of the testing effect four hours after all learning was completed (attempting an unprompted retrieval, a difficult task, would be made even harder by delaying that would allow forgetting to occur, meaning that recall is unlikely to be successful, and hence less helpful).

Subsequent research into the testing effect followed similar paradigms, with Gates (1917) requiring 3rd, 4th, 5th, and 6th graders to memorize brief biographies and lists of nonsense words. Engaging in recitation of the material to practice recall produced much better retention than simply studying the material. In addition, Raffel (1934) split participants into multiple groups, with those groups that engaged in recall of a list of words, one day subsequent to initial exposure, had improved performance on subsequent recall, and those that engaged in many consecutive days of recall maintained performance over time. So, repeated rehearsal practice maintained performance in subsequent sessions. Similar performance on word list free recall was found by Darley and Murdock (1971), with the added corollary that while it improved subsequent recall of the word lists, no improvement was found on recognition performance.

In Spitzer's 1939 paper, however, we see the first shift away from the use of rehearsal and recitation as the vehicle for forcing recall, with 3605 6th graders being asked to read passages and then being prompted with multiple choice questions about the passages as a means of promoting later recall. To measure the effect on recall over time, students were asked disjoint sets of questions about the passages at various intervals, with the initial test being delayed up to 63 days depending on the group. Engaging in recall tests immediately after studying the materials resulted in strong

improvements in subsequent test performance, compared with no test, or taking the test at a delay. Engaging in recall tests at a delay, rather than immediately after study, produced subsequent performance that was consistent with how the 6th graders performed on the delayed tests, implying that testing has the effect of making items that are successfully recalled more robustly encoded, but providing no help for items that were not successfully recalled at the time of the test.

The interaction of the testing effect with forgetting over time (which ultimately led to the 'spacing effect' see subsection 1.5.3) is clearly characterized by Whitten and Bjork (1977), showing that increasing the temporal spacing of intervening tests can increase final test performance. Bahrick (1979) continued this investigation of the temporal interactions in the testing effect from two different perspectives, one, a longitudinal method that is similar to those outlined above, with successive sessions of relearning English-Spanish word pair associations, and another, cross-sectional method that looks at data from students who have previously graduated from a college, to see how their knowledge of campus geography (in particular the spatial ordering of streets on campus) has attenuated over time - and further, how return visits to their alma mater have impacted that recall. In the first experiment of Bahrick, participants repeatedly learned word pair associations at different intersession intervals (1, 7, and 30 days), and for the longer intervals where forgetting was more likely, large increases in performance were found at each successive session. In the cross-sectional study, length of time away from campus predicted decreased performance on recall, with more frequent visits back to campus, and longer duration of those visits, both predicting higher performance on recall.

Further investigation of the testing effect examined the kind of tests that were

most efficacious for promoting later recall. Mandler and Rabinowitz (1981) found that while intervening recognition tests improve performance on free recall of word lists, this was only due to what appears to be a generalized facilitation of recall - which then also increases free recall of categorically related distractors. Thus, using recognition tests will improve recollection for the tested items, but will also increase mistaken recall of related items. In a paired associates training paradigm, Cuddy and Jacoby (1982) found that presenting the word pairs in a cued recall test after initial exposure improved final performance over simply restudying, but only when the intervening cued recall tests were different in form to the original presentation, e.g. the participant's initial testing exposure was to complete the missing letters from "LAWYER C–RT" and then subsequently fill the missing letters in "LAWYER COU–". Sequences like this improved final free recall of the paired associates, compared to a repeated exposure to the same cued recall test.

In a similar vein to Whitten and Bjork (1977) and Bahrick (1979), the work of Glover (1989) found that temporally separated free recall tests promote later recall of word lists (and that more tests facilitate more). When comparing different kinds of intervening test, Glover (1989) found that free recall intervening tests promoted better recall on final tests of any kind (free recall, cued recall, and recognition) compared to intervening cued recall, and recognition tests. This last result is particularly interesting, as it dismisses a suspicion that the effect of some kind of recall test (as compared to continued study) is simply due to the training to perform in the same way that the final test is administered, whereas here, an intervening free recall test promotes better performance than an intervening recognition test on a final recognition test.

Glover and Krug (1990) further investigated the nature of the testing effect,

in order to clarify if the improved performance from testing is due to something inherent about testing itself, or simply that a single instance of testing promoted more processing time than a single instance of study, to investigate this, participants were split into groups and underwent either recall or rehearsal of word lists, where rehearsal was being exposed to previously learned words among distractors, while retrieval was actively responding when a previously learned word was recognized among distractors - in the rehearsal condition, previously learned words were presented multiple times to increase 'processing time' for the study condition - nevertheless, final test free recall was still superior for the test condition. Carrier and Pashler (1992) continued this line of investigation, examining paired associates learning, rather than word list recall, contrasting trials with a pure study condition, and a study followed by test condition - the total amount of presentation of the word pair is higher in the pure study condition, as the second presentation in the test condition only shows the first half of the word pair for half of the trial, whereas it is present for the entire duration in the study only condition. Experiments showed the same result as Glover and Krug (1990) did for word list recall, that in spite of greater presentation time, the study only condition was still outperformed by the test condition.

Kuo and Hirshman (1996) argued that Carrier and Pashler (1992) failed to completely separate the impact of testing and study, to counter this Kuo and Hirshman investigated different sequences of three item word list engagement, each sequence starting with a study trial, and test trials being free recall. In this way, unlike Carrier and Pashler's procedure, it is possible to create a sequence in which participants are tested, but with no additional opportunities for studying items again. Performance in the pure testing sequence was significantly higher than performance in the pure

study condition, showing that representation after testing does not explain the testing effect, and that it is rather associated with the act of retrieval during the free recall trial. Kuo and Hirshman used a ten minute retention interval, after repeated study and/or test trials (following the initial study, four of a mixture of study and test trials); conversely, to give a more ecologically valid context, Roediger and Karpicke (2006a) asked participants to study short passages of Psychology texts, and then to either be tested or restudy (as the kind of repeated testing or restudy seen in Kuo and Hirshman (1996) is unlikely given time constraints, and learners are more likely to be engaging with more extended text materials than word lists). In this context, the twice repeated study condition produced much higher retention at a 5 minute delayed test, but the testing condition produced much higher retention for two day and one week delays.

In the last decade, interest in the testing effect has seemingly proliferated, both from a theoretical, and an applied stance, with three reviews or meta-analyses of different aspects of the testing effect (Pashler, Rohrer, Cepeda, & Carpenter, 2007; Carpenter, 2012; Rowland, 2014), and three articles targeted at audiences of educators urging the use of the testing effect for practical impact in the classroom (Roediger & Karpicke, 2006b; Rohrer & Pashler, 2010; Rawson & Dunlosky, 2012). In addition, there has been substantial interest in the range of applicability of the testing effect, in other domains, and in educational contexts, its impact on transfer and related material, and how complex the materials tested can be. Further, the exact sequencing of tests and studying has been investigated, and a range of theoretical proposals for the mechanisms underlying the testing effect have been proposed.

Other Domains and Ecological Validity

One area of research that has received particular attention in the last decade is the investigation of extensions of the testing effect into other learning domains, and into more ecologically valid studies in formal education contexts. The latter is of particular importance, given the strong prescriptions in favour of testing offered to educators by Roediger and Karpicke; Rohrer and Pashler; Rawson and Dunlosky. The former is also of interest if we are to consider how the testing effect may have an impact in ecologically valid learning contexts, where learning is rarely restricted to the relatively constrained paradigms represented in the studies discussed so far (memorization of word lists, recall of factual material from texts, and word pair associations).

Both Carpenter and Pashler (2007) and Rohrer, Taylor, and Sholar (2010) found evidence to support the testing effect in visuo-spatial learning, specifically being able to recall locations on maps for various features, showing that the testing effect is evidenced in learning tasks that may be of importance in tasks far removed from what we might ordinarily consider rote memorization. Yamashita (2016) showed a similar visuo-spatial memory testing effect using the Rey–Osterrieth complex figure test (Osterrieth, 1944), where participants are required to reproduce a complex figure from memory following a relatively brief presentation (with no instruction to maintain memory of it for later recall). Some participants are also asked to reproduce the diagram at a short delay, giving a natural testing effect within the administration of the task - Yamashita followed up with these patients, and showed evidence of a testing effect at one year delay for those that were in the delayed test condition. Meanwhile, Johnson and Mayer (2009) found evidence for the testing effect on delayed recall in a

multimedia presentation on lightning, while Pierce and Hawthorne (2016) extended the traditional verbal word list to an auditory presentation, and found the testing effect pertaining in an auditory modality as well.

In a similar vein to previous studies that involved reading texts, Wiklund-Hörnqvist, Jonsson, and Nyberg (2014) studied the impact of testing as opposed to rereading texts on conceptual knowledge of introductory Psychology concepts. Unfortunately, due to the nature of their materials, it is not clear that this work shows the extension of the testing effect to conceptual learning, as the materials learned in the testing active condition were merely similarly worded sentences to the final test question. It seems therefore, that what was being learned was merely rote learning of definitions. From the perspective of an introductory Psychology course, however, this may well be ecologically valid, even if it fails to show the testing effect operating for broader conceptual learning.

Ecologically valid tests of the testing effect have been conducted in a range of subjects ranging from Cognitive Neuroscience (McDaniel, Anderson, Derbish, & Morrisette, 2007) and Psychology (Wiklund-Hörnqvist et al., 2014; Trumbo, Leiting, McDaniel, & Hodge, 2016) at the college level, 8th Grade History (Carpenter, Pashler, & Cepeda, 2009), Middle School Science (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011, 2013), and 6th Grade Social Studies (Roediger, Agarwal, McDaniel, & McDermott, 2011). All of these ecologically valid tests were primarily concerned with the direct impact of testing on factual recall (improving recall for related items and transfer tasks is discussed below in section 1.5.1), broadly showing the robustness of the testing effect in formal education contexts, across a reasonably wide age range of participants, from Middle School through to Undergraduates.

Transfer

Unfortunately for educational applications of the testing effect, the institution of widespread high stakes testing (in the US in particular) has caused a general backlash against testing as a whole in Education circles (Volante, 2004). The generally expressed concern is that overuse of testing will tend to focus on rote memorization of facts that are tested, rather than developing skills and broader conceptual knowledge. It is important, therefore, both for wider uptake of testing as an educational tool, and also for the broader purposes of education beyond rote memorization of facts, to show whether the testing effect can facilitate transfer, and do more than simply improve recall for the set of tested facts (this is particularly important, as there will always be a limited pool of questions that are asked as part of testing, and so biases in the way the questions are asked, may bias the facts that are learned).

In a series of experiments Chan has attempted to study the impact of the testing effect on recall for related but untested items. Of particular concern is that testing only on a subset of items that it is desirable for a learner to later recall may result in retrieval induced forgetting (Anderson, Bjork, & Bjork, 2000) on the untested items. Contrary to the predictions of retrieval induced forgetting, however, Chan, McDermott, and Roediger III (2006) found instead retrieval induced facilitation of untested items when participants read passages and were tested on some of the facts contained therein. Further, it seemed that there was more facilitation of untested items, when there were long reaction times (suggesting 'effortful' retrieval) on the untested items in the final test, and in the tested items on the initial test. Chan (2009) further showed that presenting text material in an incoherent way (i.e. as sequential jumbled sentences, rather than one coherent body of text) and using a short delay

between the retrieval practice and the final test produced retrieval induced forgetting, as such for the testing effect to have the desired impact on untested material, the original material needs to be initially presented in an integrated way. These results were further replicated by Chan (2010), with recall tested at either 20 minutes, 24 hours, or 7 days - notably, there was an uptick in performance at the 24 hour mark, compared to the 20 minute test, indicating that the impact of the initial test on both the tested and untested items may not be fully realized until after some small delay.

Wissman, Rawson, and Pyc (2011) showed in a series of experiments that testing did not just facilitate material presented alongside tested material, but also that testing on passages of text facilitated later recall on subsequently presented related text passages. This suggests that the facilitation was seen not only the for original material, but also for related matieral not presented simultaneously to the tested material. Work on the testing effect and deductive inferences by Tran, Rohrer, and Pashler (2014), Eglington and Kang (2016) reinforces this notion, where Tran et al. found a string of null results for the testing effect on deductive inferences, seeming to indicate that being tested on facts had no impact on subsequent ability to make inferences based on them. By contrast, Eglington and Kang found that presenting the related facts synchronously and then subsequently testing them, as opposed to serial presentation, showed a testing effect in subsequent deductive inference questions.

The foregoing would seem to allay some of the possible concerns about negative effects of testing in educational contexts, where, if tests and materials are presented in a way that promotes associations and integration of material, retrieval induced facilitation, rather than forgetting, seems to be the more likely outcome. However, educators are not usually interested solely in rote recall of knowledge, but an ability

to recall the right information and use it in new ways, either in very different contexts from where it was originally learned, through analogical reasoning, or by integrating a range of learned information to reach new conclusions. In their multimedia testing effect experiment Johnson and Mayer (2009) found that improved performance in delayed transfer tasks (i.e. applying their learned understanding of lightning to questions that required a deeper conceptual understanding of lightning) was only found when participants engaged with other transfer questions during the testing phase - participants who only engaged with recall questions performed similarly to participants who only engaged in restudy and not testing. In contrast, Rohrer et al. (2010) found in a map learning task, where participants learned the location of cities in a road network, performance on an integrative transfer task, where participants had to state the cities that a particular city to city route would pass through, was improved in the testing condition over simply restudying.

Butler (2010) found evidence for a testing effect when reading passages, in a simple recall test, in tests of inference within the same knowledge domain as studied, and by analogical reasoning to other unstudied knowledge domains (i.e. from the mechanics of bat flight to aerodynamic engineering). In addition, Butler found, similarly to Johnson and Mayer, that testing on such an analogical reasoning type transfer question improved performance on other analogical reasoning questions. Similar results were found by McDaniel et al. (2013) in their work in Middle School Science classrooms, where, while performance was most improved for definition questions, students performed as well on differently worded questions, and had improved performance on application questions related to items that they had been tested on rather than simply studied. Finally, as with Johnson and Mayer (2009), Butler (2010), the most

improvement was found on application questions when they were tested on application questions.

A concern highly relevant to much testing of factual material is highlighted by Pan, Pashler, Potter, and Rickard (2015, 2016)'s work on triple paired associates (e.g. 'DOG CAT OX'), and their more ecologically valid cognate, the multi-part fact. Many facts that are to be learned in the classroom involve multiple parts, such as 'Winston Churchill was Prime Minister of Great Britain in World War II' - however, multiple choice testing of such facts may well emphasize only one element of that fact for retrieval, such as the question: 'Who was Prime Minister of Great Britain in World War II?' Pan et al. (2015) show first with triple paired associates that testing on one word from the triple paired associates will promote cued recall for that word, but not for any other pairings from the triple. Similarly, Pan et al. (2016) showed that testing for recall of one fact from a multi-part fact will result in a testing effect for that particular missing part of the fact, but not for the other parts. As such, poorly created tests, while promoting recall for one particular part of the fact, will not be promoting general performance. As such, tests, particularly tests that rely on a 'fill in the blank' mechanism will have to test every permutation of missing facts in order to guarantee a positive impact of a testing effect on the whole multi-part fact. Fortunately, however, Pan et al. found no evidence for retrieval induced forgetting of other parts of the fact, so the potential for negative (rather than neutral) impacts of testing are limited.

Complex Materials

As much interest in the testing effect is related to its potential for causing changes in the way that we structure instruction, understanding the extent to which complex learning can be reinforced by the testing effect is of considerable interest. An experiment by Darabi, Nelson, and Palanki (2007) with Chemical Engineering students found evidence for a testing effect compared to following worked examples in solving problems in a computerized simulation of a chemical processing plant. This is in contrast to a later study by Gog et al. (2015) which found no evidence of a testing effect when compared to worked examples for Educational Studies students solving Electrical Engineering problems. While the latter null result suggests that there may be theoretical limits to the impact of the testing effect (i.e. if the material is not fully processed or understood during study, then being tested on it may produce no additional benefit), the former result uses a much more ecologically valid population, and so even for complex materials it would seem to be appropriate to employ the testing effect to improve learning.

Coherent text passages have also been proposed as being too complex to benefit from the testing effect. Jonge, Tabbers, and Rikers (2015) found no effect of testing as opposed to restudy of coherent text passages, until they split the text passages into isolated sentences and scrambled their order. However, as noted by Rawson (2015), this single null result is in stark contrast to a wealth of previously discussed positive results, such as Roediger and Karpicke (2006b), Chan et al. (2006), Chan (2009), Chan (2010), McDaniel et al. (2007), Wiklund-Hörnqvist et al. (2014), Trumbo et al. (2016), and even as far back as Spitzer (1939).

In spite of this, Leahy, Hanham, and Sweller (2015) (in experimental work)

and Gog and Sweller (2015) (in a review of the literature) have proposed that the high 'elemental interactivity' of complex materials either renders the testing effect null or reverses it (Leahy et al., 2015). Leahy et al. conducted an experiment where primary school students were instructed on how to read a bus timetable (a task that they claim to be high in 'elemental interactivity', as the bus timetable has multiple elements, each of which need to be understood and the connections between each properly manipulated by the participant in order to successfully complete the task). The procedure required to successfully read the bus timetable is sequential, and failure at any step in the process will result in a failure to complete the task. Students were presented either with eight repeated worked examples, or alternating worked examples and practice problems. The participants who received only worked examples out performed the group who engaged in practice problems. Similarly to Jonge et al., 2015, Leahy et al. (2015) use learner group that lacks basic proficiency with the task being instructed, and find no positive impact of testing.

While much heat is generated by Leahy et al. (2015)'s 'reverse testing effect' on an immediate test, this is consistent with the vast majority of the literature, where immediate tests tend to show no different or sometimes reduced performance in testing versus study. A null result at a delay in Leahy et al. (2015)'s experiment is therefore left to do the heavy lifting of this claim, but, as noted by Karpicke and Aue (2015), the concept of 'elemental interactivity' is poorly defined, not quantified, and not even varied in their experiment. As such, it is impossible to draw the conclusion that the null result was due to the increased 'elemental interactivity' that they claimed was in their bus timetable learning task.

Further, Karpicke and Aue observe that the subjectivity of this 'elemental

interactivity' skews the literature review that Gog and Sweller undertake to prove their thesis. With no objective measure to determine 'high elemental interactivity', only the few experiments that support their thesis outlined in this section, and Tran et al. (2014)'s null results for deductive inferences are accepted as unequivocally of 'high elemental interactivity' - however, Karpicke and Aue show that by several quantitative measures (word count, reading ease, grade level, or referential cohesion - the extent to which ideas are referenced across multiple sentences in a text) there is no apparent distinction between the textual materials that are rated as having 'high elemental interactivity' that show null results and others that show positive results, such as Roediger and Karpicke (2006b), Johnson and Mayer (2009), McDaniel, Howard, and Einstein (2009), Hinze and Wiley (2011), Karpicke and Blunt (2011).

Finally, Karpicke and Aue note that by any reasonable interpretation of 'elemental interactivity' the categorization of recall tasks into low, medium, and high by Gog and Sweller is not incoherent, but backwards. Tasks where the demanded recall requires integration of multiple facts and ideas, such as conceptual short answer questions (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Blunt & Karpicke, 2014; Weinstein, McDermott, & Roediger, 2010; Kang, McDermott, & III, 2007) are rated as low in elemental interactivity, in spite of the presumably high cognitive load required by integration across multiple ideas to respond to questions across a range of topics from the Voyager space probe, to the KGB, to Enzymes. In addition, as Karpicke and Aue note, Blunt and Karpicke required students also to engage in creating concept maps of the material they had studied as their recall test - requiring connections to be made among all the facts and information that the students had learned, this would seem to be a quintessentially 'high elemental interactivity' recall

activity, but Gog and Sweller rated it as 'low/medium'.

To summarize, in spite of the concerns raised in the literature, and as suggested by the reviews written by Rawson and Karpicke and Aue, it seems that the testing effect does pertain even in the learning of complex materials, be they problem solving techniques, or complex passages. What is apparent, however, is the pertaining of the testing effect may depend on the proper presentation of material, as shown by Eglington and Kang (2016)'s replication and extension of Tran et al. (2014)'s work on deductive inferences. Further, the contrast shown by the null result for complex Engineering problem solving tasks for Educational Studies students (Gog et al., 2015), but not for Chemical Engineering students (Darabi et al., 2007), suggests that for testing to be effective material has to be completely comprehensible by the student in initial study.

Study and Test Ordering and Question Design

Studies of the testing effect rely on an initial study phase, where participants are exposed to the material to be learned, followed by either restudy or testing to see the impact of testing on later recall. Instinctively, many learners will rely on restudying as a means of rehearsing material, rather than retesting - it is interesting to discover, therefore, what particular schedule of studying and testing is optimal for later recall. In an attempt to answer this question, Karpicke and Roediger (2007) tested different sequences of testing and studying - either studying followed by alternating testing and studying, by repeated study then testing, or repeated testing with no additional study. Importantly, the testing phases consisted simply of the participants recalling the memorized word lists, with no feedback. Superior performance was found

for the alternating testing and studying, mirroring the results seen for testing with feedback by Carrier and Pashler (1992).

The findings of Kang et al. (2007) support these conclusions about the efficacy of repeated exposure to material after testing, in a comparison of multiple choice tests against short answer tests, in an experiment where no corrective feedback was given, the multiple choice intervening test condition produced better retention on a mixed multiple choice and short answer retention test compared to the short answer intervening test condition. However, when corrective feedback was supplied, the short answer condition (presumably the more effortful retrieval of the two) produced greater retention in a delayed test. It is reasonable to infer that the multiple choice condition gave a weak opportunity for restudy, whereas the corrective feedback gave a much stronger opportunity. In a similar vein, Carpenter, Pashler, Wixted, and Vul (2008) used corrective feedback in a comparison of testing against restudy. Contrary to Roediger and Karpicke (2006a) (and others) they found improved performance at 5 minute delay for testing against studying (where previous studies had found the converse) - presumably due to the restudy opportunity afforded by corrective feedback.

Supporting the conclusion of multiple choice questions giving a slight restudy benefit, Bishara and Lanzo (2015) investigated the impact of using 'all of the above' type multiple choice questions, without feedback, in intervening tests. When 'all of the above' was the correct response, i.e. all of the responses shown to the participant were in effect a restudy opportunity, participants performed better on the retention test, regardless of whether the questions asked were standard multiple choice or cued recall - no such effect was found when the 'all of the above' option on the intervening test was incorrect, i.e. some of the material presented for the effective restudy opportunity was

incorrect. This was true, for both delayed and immediate recall tests, mirroring the results of Carpenter et al. (2008) but without the use of feedback to give additional restudy, however, as with Carrier and Pashler (1992) matching timed exposure to the options but without an 'all of the above' prompting question also led to improved retention, indicating that the act of recall was instrumental in producing the effect, not just the restudy opportunity.

In their 2011 study in Middle School Science classrooms, McDaniel et al. used different test sequences to attempt to determine when was the most efficacious timing of a test for subsequent recall. All tests were conducted without feedback to the students. Three different intervening tests were used in exhaustive combination, a prelesson quiz, where students were tested on the material after reading a textbook chapter, but before a formal lesson, a postlesson quiz, where students were tested immediately after the formal lesson, and a review quiz, where students were tested twenty four hours before a unit test. The different sequences made by combinations of these different tests were mixed within subjects on class by class and per item basis so that the impact of each testing sequence could be estimated. Prelesson quizzes seemed to have no effect on their tested items in the unit test, with postlesson quizzes performing better, and review quizzes producing the best retention (unsurprisingly given the much shorter delay from review test to unit test), these results were replicated on end of semester and end of year tests, with a small increase in retention from having postlesson quizzes preceding the review quizzes compared to the review quizzes alone. Interestingly, in light of the apparent null effect of the prelesson quizzes (which seem to replicate the procedure used in many studies already discussed of studying text passages with subsequent testing), student participation in the postlesson and

review quizzes was incentivized as part of their class grade, whereas the prelesson quiz was not. As all the quizzes were administered via multiple choice clicker response, the lack of a motivational incentive in the prelesson quiz condition could have resulted in a much higher guessing rate. In spite of this, and the significantly lower average score on the prelesson quizzes McDaniel et al. insist that no significant motivational differential existed, and that this is merely indicative of the inadequacy of prelesson quizzing in an ecologically valid context. Alternatively, it could be indicative of the inadequacy of the textbook material for conveying the information to the students in the first place.

To further examine the mechanism for the impact of the testing effect McDaniel, Bugg, Liu, and Brick (2015) conducted experiments on undergraduate participants using material from an introductory Psychology text. In the first experiment, after initial study of passages, a sequence of either three tests, three restudy sessions, or test/restudy/test were administered, where the tests gave no feedback to the participants. The restudy sequence significantly underperformed the other two sequences in a retention test five days later, while the other two sequences were indistinguishable.

In the follow up experiment, tests gave feedback about correctness, but did not indicate the correct answer - this information was also available during any subsequent restudy session. Additionally, in the repeated study condition, the material in the questions themselves was made available during the second session. In the repeated test condition, the correct answer was given to participants alongside correct/incorrect feedback in the second and third testing sessions. In the test/restudy/test condition participants were asked to highlight the passages to show information relevant to the answers to the questions.

In this set of circumstances, the test, study, test sequence outperformed both the repeated testing sequence and the pure study sequence, which had indistinguishable performance, implying that in the context of complex text passages, the testing effect is simply a consequence of highlighting the important material to be learned. However, this contrasts with the previously discussed findings of Chan (2009), Chan (2010) where testing after reading passages improved retention of even untested material. In addition, when McDaniel et al. collapsed the results across the two experiments, the repeated test condition still significantly outperformed the repeated study condition. It seems more parsimonious to conclude that, parallel to Karpicke and Roediger (2007), Carrier and Pashler (1992), Kang et al. (2007), Carpenter et al. (2008), the targeted and extensive feedback of the test/restudy/test condition is an important component of any educational application of the testing effect, while questions must be properly designed to prevent the kind of piecemeal learning found by Pan et al. (2016) in the learning of multipart facts.

In a comparison of free recall and cued recall questions following the reading of texts, Smith, Blunt, Whiffen, and Karpicke (2016) found no difference in the testing effect benefit - participants did just as well compared to restudy in either condition, however, free recall was reported to be more interesting and enjoyable than the cued recall condition. Conversely, in a paired associates task Halamish and Bjork (2011) found not only that the testing effect is moderated by the difficulty of the final question format (i.e. a more pronounced effect will be found in free recall tests, followed by cued recall, then multiple choice), but that a cued recall intervening test produced a maximal testing effect with a free recall final test.

In order to assess more precisely when tests should be administered, Weinstein,

Nunes, and Karpicke (2016) conducted a lab study, an online study, and a classroom study examining the impact of interleaving tests with study or blocking at the end of study of material on APA style. In all three studies, interleaving produced better mean performance on the intervening tests, but produced no differences in retention at delays of one week (for the lab and online procedures) or nineteen days (for the classroom study). Similarly, Abel and Roediger (2016) found that blocking or interleaving testing and restudy in the learning of English-Swahili word pairs had no differential impact on retention either at a short or long delays or at a short delay with or without intervening interference activity.

Finally, Rawson and Dunlosky (2011) carried out an exhaustive series of experiments in an attempt to delineate heuristics for the application of the testing effect in educational contexts, in these three experiments, they varied the criterion for initial learning (ranging from one correct recall of an item to four correct), and number of subsequent intervening test sessions (where the participant was only required to demonstrate correct recall once on an item). Across the three experiments, increasing the initial criterion resulted in improved recall and increased efficiency of relearning at the subsequent intervening test, however, beyond that test, the impact of the initial criterion was attenuated, and as more intervening tests were taken by participants its impact disappeared. By assessing the total number of trials spent on learning, however, Rawson and Dunlosky (2011) propose the heuristic prescription of learning to a criterion of three correct recalls, with three subsequent intervening tests at intervals of weeks, giving durable performance either at a one month delayed test, analogous to examination within a US Higher Education system, or at a four month delayed test, analogous to reusing learned information in a subsequent course. Further

implications for the time course of repeated testing are discussed in subsection 1.5.2 and subsection 1.5.3.

In summary, testing seems to be best conducted in a way that will promote effortful, but successful recall, while at the same time increasing learner exposure to correct material. As such, the repeated use of multiple choice questions with incorrect distractors and little opportunity to engage with the correct material (as in McDaniel et al. (2015)'s repeated test condition) may cause little or no benefit for testing when compared with restudy. The exact sequencing of testing with study of other materials seems to have no long term effect on retention, so quizzes can be administered when convenient after initial study. Finally, it seems it may be beneficial for overall efficiency and durability of learning (assuming a goal of longer term retention) to engage in a repeated testing procedure with feedback until the correct answer is produced three times (although, this may vary depending on the kind of questions being answered).

Theory and Mechanisms

In a recent meta-analysis Rowland (2014) discerned a general trend in studies of the testing effect that those that used a free recall paradigm as the intervening test, produced a larger testing effect than those that used recognition tests. The implication drawn, therefore, is that some form of effortful processing is required in order to elicit the testing effect. This is further supported by the results of a paired associates task by Carpenter (2009), where a strong testing effect was found for paired associates that had only a weak rather than strong natural association. Pyc and Rawson (2009) systematically manipulated the difficulty of retrieval during the testing phase, with short and long interstimulus intervals between initial presentation and testing - in

line with the retrieval effort hypothesis, longer interstimulus intervals, which would produce more difficult recall, led to greater post test retention. Endres and Renkl (2015) found similar results with improved performance of free recall over cued recall, and cued recall over restudy when studying texts - interestingly, the impact of testing was entirely moderated by self-reports of mental effort.

Carpenter and Pyc and Rawson (2010) both claim that their results are indicative of a semantic elaboration effect (Carpenter, 2009) or an improvement in the encoding of semantic mediators between the cue and the learned target (Pyc & Rawson, 2010) - however, this semantic mediator account, whereby the elaborative retrieval that occurs during testing strengthens connections between mediators and the target, while useful for word based paired associates tasks, does not seem satisfactory for non-semantic encoding tasks, such as visuo-spatial learning (Carpenter & Pashler, 2007; Rohrer et al., 2010; Yamashita, 2016), or complex tasks, for which Rawson (2015) argues forcefully in favour of the existence of a testing effect. Further, Pyc and Rawson, by explaining the testing effect by the better learning of semantic mediators that connect the cue with the target, are not explaining the testing effect, but how semantic paired associate learning is achieved - it leaves aside the mechanism (beyond a requirement for effortful retrieval) by which engaging in effortful recall, rather than simply recall or restudy, produces a stronger connection in memory between the mediator and the target.

Halamish and Bjork (2011), Kornell, Bjork, and Garcia (2011) propose a model that treats all events from study to free recall as a continuum of recall difficulties. The retrieval strength for a particular memory is modeled as a probability distribution, where the probability of recall is area of the distribution greater than the difficulty

of the recall event. When successful recall occurs (which is almost guaranteed under study conditions) then the strength of a memory trace is enhanced, in proportion to the difficulty of the recall event, this is represented by an increase in the mean of the distribution, so that, as a whole, recall in the future is more likely. As such, free recall, when successful, will tend to produce larger gains in later retention, while the opposite end of the spectrum, study will produce much smaller gains. Finally, a difficult final test will tend to benefit more from a testing effect, as more items will have their distributions shifted to produce relatively high levels of recall, compared to restudy. For an easy final test, however, such as a recognition based multiple choice test, the testing effect may well be moderated as the gains from restudy will be sufficient to produce a high probability of recall for most items.

This clearly contrasts with a retrieval practice account of the testing effect, whereby engaging in the intervening test (as opposed to restudy) gives practice of precisely the kind of retrieval that will occur in the subsequent recall test - Halamish and Bjork show this to be the case by producing the largest testing effect gains with a cued recall intervening test, but with a free recall final test. The most important prediction of this model, however, is that retrieval success during intervening study should entirely moderate any subsequent testing effect, if no feedback is given subsequent to testing. This is supported by work on word list memorization by Rowland and DeLosh, 2015, where even on short intervals a testing effect compared to restudy is only shown when looking at items that had successful retrieval in the intervening test. Kornell, Klein, and Rawson (2015) further examined the impact of success by showing there was no clear differential between a successful attempt and a failed attempt with feedback, implying that the retrieval attempt caused some potentiation of the memory trace

that was subsequently enhanced by the presentation of the paired associate completion - the same potentiation occurs in the success condition, but the item is brought to mind by recall rather than external stimulus.

Finally, while there is clear evidence that retrieval practice enhances the strength of retrieval, Sutterer and Awh (2015) showed that retrieval practice failed to improve the precision of colour recall - when selecting the appropriate colour paired with a distinct shape from a continuous colour wheel, distributions of responses indistinguishable from guessing were reduced due to retrieval practice, but the variance of distributions around the correct answer were indistinguishable between test and restudy conditions.

In summary, current theoretical accounts of the testing effect (predominantly focused on paired associates paradigms) suggest that effortful retrieval enhances potentiation for retrieval of an item, and once the item is available (either from recall or external presentation) the presence of that item enhances the retrieval strength of the memory for that item. However, the memory itself does not appear to be enhanced in terms of its precision.

Summary and Implications for Mathematics Learning

The testing effect seems to pertain across a wide range of domains, from paired associates for meaningful word pairs, all the way up to complex electrical engineering problem solving. One area that seems to have been neglected so far in the literature on the testing effect, however, is Mathematics, which presumably would have more in common with the complex electrical engineering problem solving than the paired associates tasks.

Testing seems to be most effective when paired with corrective feedback, although the simple act of being tested may still produce gains. In the case of Mathematics, when, beyond the learning of basic Arithmetic facts, the feedback given is not to be memorized for later use, but rather used as a way of reinforcing procedures or conceptual understanding, it is interesting to see how much of an impact the use of feedback has on integrating these ideas. While the work of Wiklund-Hörnqvist et al. (2014) implies that Psychology concepts can be reinforced by testing and corrective feedback, there was very little similar to the frequent learning of calculation procedures that is seen in Mathematics learning.

As noted by Rohrer, Dedrick, and Burgess (2014), this application of procedure means that Mathematics can frequently be a two step process. Firstly, there is the recall of the correct procedure, then the application of that procedure to the problem at hand. In many instances, the recall of the correct procedure may not be difficult (due to exercise sequences that are blocked by procedure), if this kind of exercise sequencing is used, then, presumably, no testing effect would apply to the recall stage, only to the application stage, where (as in the worked examples of Darabi et al. (2007)) the problem solving procedures are applied and may well show better results in subsequent attempts. However, if the students were then tested in a mixed exercise paradigm, then due to the lack of retrieval practice for the correct procedure, students may well perform significantly worse.

1.5.2 Forgetting Over Time

The Nature of Forgetting

Early studies of forgetting (Radosavljevich, 1907; Ebbinghaus, 1913; Finkenbinder, 1913; Strong, 1913; Luh, 1922; Burtt & Dobell, 1925) all establish the basic phenomenon of forgetting - over time items learned will either be recalled less frequently (Luh, 1922), recognized less frequently (Strong, 1913), or will take more exposures to relearn (Ebbinghaus, 1913). Additionally, this decay will happen in a way that is non-linearly related to the amount of time that has passed. For Ebbinghaus this relationship was logarithmic, but for others, variability was found between subjects (Luh, 1922). This work on declarative learning was extended to motor task proficiency (Ammons et al., 1958), but no change in mean reported auditory stimulus intensity or frequency was found by King (1963b, 1963a) - unfortunately, these latter results failed to report the variance of performance over time, which might have been a better index of decaying performance over time.

Interestingly, Ebbinghaus noted a slight recovery at the twenty four hour mark, but dismissed it as a discrepancy that fell within error in the measurement - however, a recent replication of his original study (Murre & Dros, 2015) not only replicated Ebbinghaus's original results, but also this recovery at the twenty four hour time point, at an extreme not accountable by random variation. This implies that, at least for measurements of savings, the form of the curve is not guaranteed to be monotonically decreasing. One explanation for this is the effect of sleep in the intervening twenty four hours, Jenkins and Dallenbach (1924) made an early report of evidence of the stabilization of forgetting curves following sleep, with fairly constant performance across the hours of sleep (participants were awoken from sleep after between one and

eight hours of sleep), compared to rapid forgetting of syllables they memorized at the start of the day.

As summarized in the meta-analysis by Rubin and Wenzel (1996), the forgetting phenomenon has been demonstrated in both recognition and recall memory, across a range of paradigms and task demands, with short term recognition (on the order of minutes to hours) being demonstrated (Wickelgren, 1968; Wickelgren, 1972; Begg & Wickelgren, 1974; Wickelgren, 1974, 1975a), long term recognition (on the order of hours to days and beyond) (Strong, 1913; Luh, 1922; Burtt & Dobell, 1925; Spitzer, 1939; Wickelgren, 1972; Fajnsztejn-Pollack, 1973; Wickelgren, 1975b; Gehring, Toglia, & Kimble, 1976; Glasnapp, Poggio, & Ory, 1978; Squire, 1989; Fioravanti & Di Cesare, 1992), short term recall (Peterson & Peterson, 1959; Murdock, 1961; Waugh & Norman, 1965; Bregman, 1968; Turvey & Weeks, 1975; Wixted & Ebbesen, 1991), and long term recall (Bean, 1912; Luh, 1922; Burtt & Dobell, 1925; Krueger, 1929; Minke & Stalling, 1970; Nelson, Shimamura, & Leonesio, 1980; Runquist, 1983; Fioravanti & Di Cesare, 1992; Conway et al., 1993). In spite of a wealth of empirical characterizations of forgetting, two things have remained more elusive: a model that accounts for the varied mathematical forms empirically measured curves of forgetting may take (either how retention of performance deteriorates over time, or in the savings methodology used by Ebbinghaus, how much additional time to relearn to criterion changes over time), and the mechanisms via which forgetting occur, either through the natural decay of a memory trace, or through some interference mechanism - either proactive interference where new learning is disrupted through competition with previously learned items, or retroactive interference where learning following initial learning disrupts the initial learning.

Memory Consolidation

Interestingly, while memory has long been hypothesized to be composed of multiple interacting systems (Atkinson & Shiffrin, 1968), the empirical search for a forgetting function has made no great distinction between the form that would be used at short time scales as at long ones. In addition, very few studies (only 20 out of the 210 data sets analysed by Rubin and Wenzel (1996)) have looked at forgetting intervals that include multiple retrieval points before and after time scales when consolidation processes have been shown to occur (McGaugh, 2000) in multiple learning paradigms. Further, as performance on some tasks have been shown to improve following consolidation occuring during human sleep (Stickgold, 2005), the use of a solely monotonic function to represent retrieval performance across a consolidation 'boundary' seems to be unlikely to properly pertain.

Until recently, a body of evidence had accumulated purporting to show that sleep promoted consolidation of a procedural learning task in a procedural tapping task (Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002, 2003), but with much smaller improvement in children (Wilhelm, Diekelmann, & Born, 2008), and in learning musical keyboard sequences (Duke & Davis, 2006). However, in attempting to provide a causal role for sleep in this enhancement, mixed evidence has been advanced showing that either Slow Wave Sleep (Milner, Fogel, & Cote, 2006; Backhaus & Junghanns, 2006; Nishida & Walker, 2007) or conversely Rapid Eye Movement Sleep plays a critical role for motor learning (Marshall & Born, 2007). Further, some have argued that consolidation of motor learning is a purely time based phenomenon, where inevitably sleep will have resulted in more time passed, and hence more consolidation (Vertes, 2004; Robertson, 2004). More recently, the entire phenomenon of procedural learning

consolidation during sleep has been thrown into question. Rieth, Cai, McDevitt, and Mednick (2010) showed that in a replication of the procedural learning task used by Walker et al. and others, taking into account non-sleep factors such as massed practice and circadian confounds entirely moderated the effect of sleep on motor learning improvements. More damningly, Pan and Rickard (2015) have shown fairly convincingly with a meta-analysis that the improvements reported across a range of procedural learning consolidation studies could be a result of several non-sleep factors, including time of testing, duration of training, elderly status, and averaging artifacts of incremental improvement across blocks separated by sleep. Here the improved performance due to practice at the end of a post-sleep block will cause a seeming improvement post-sleep, whereas the far lower performance at the beginning of the pre-sleep block will show a much lower performance than the final performance level before sleep.

In the literature, it seems that little attention has been paid to the effect of sleep on perceptual learning, but overall the impact appears to be positive (Mednick, Nakayama, & Stickgold, 2003; Mednick, Cai, Kanady, & Drummond, 2008), with the exception of naps induced by Ambien (Mednick et al., 2013) which seems to be worse than naps under the influence of a placebo.

Consolidation of episodic and declarative learning has been linked to Slow Wave Sleep (Gais & Born, 2004; Tucker et al., 2006), with a strong correlation between the prevalence of sleep spindles and the amount of consolidation of declarative memory (Schabus et al., 2004; Mednick et al., 2013). These consolidated memories are thought to become more reliant on cortical networks, as opposed to the hippocampal representations that underlie their initial encoding (Frankland & Bontempi, 2005).

Further evidence to support the consolidation of declarative memories has been found in high school vocabulary learning (Gais, Lucas, & Born, 2006), word-pair associates tasks in children (Wilhelm et al., 2008), and word list recall (Mednick et al., 2008). Finally, due to a wealth of evidence showing similar consolidation effects to Slow Wave Sleep from NMDA receptor antagonists, benzodiazepines, alcohol, and acetylcholine antagonists, it has been suggested that hippocampus dependent memories seem to benefit from reduced retroactive interference (Mednick, Cai, Shuman, Anagnostaras, & Wixted, 2011), whereby there is less competition for encoding resources in the hippocampus (Wixted, 2004b) (due to a limited number of newly formed hippocampal neurons which are integral to the initial encoding (Frankland, Köhler, & Josselyn, 2013; Sadeh, Ozubko, Winocur, & Moscovitch, 2014; Déry, Goldstein, & Becker, 2015)), which also happens to be induced by slow wave sleep.

In summary, therefore, it seems that while not all learning will benefit from sleep induced consolidation, the explicit declarative and episodic learning that underpins much formal education will benefit from sleep induced consolidation - whether from a good night's sleep or from a nap (Mednick et al., 2008, 2013).

Mechanisms of Forgetting

As previously discussed, there are three primary potential mechanisms that could (alone or in combination) drive forgetting over time. First is that there is simply a decay process associated with the encoding strength of a memory, this could be driven by a global garbage collection process, as suggested by Altmann (2009), an analogue equivalent of the garbage collection processes employed by many digital computer programming languages, where memory locations that have no extant referent can

be safely emptied to create space for other representations. In the analogue version, however, rather than do a costly search for active referents, a continual degrading of the representation is hypothesized, which is then periodically counteracted by a currently active referent in retrieval of the memory.

The other two mechanisms are both kinds of interference, as detailed by Wixted's 2004b review, proactive interference, whereby previously learned material provides cue competition for the act of recall, and retroactive interference, whereby learning other material after initial learning interferes with ongoing encoding and consolidation of that material. In long term memory, it has long been seen as uncontroversial that interference effects are the sole drivers of forgetting over time (McGeoch, 1932), however for more short term memory (over the span of seconds to multiple minutes), while there has been no suggestion that interference plays no role in forgetting, there has been much argument over whether simple decay over time has an additive causal role in forgetting.

Interference in Long Term Memory

Wixted (2004b) argues against the traditional story of proactive interference (interference from previously encoded memories) being the primary driver of forgetting, instead suggesting that retroactive interference prior to consolidation of the memories is the primary cause of forgetting. This hypothesis is driven by findings that sleep (as in section 1.5.2), and amnesic substances like benzodiazapines and alcohol both improve subsequent recall compared to ordinary wakefulness, and hence other memories formed in that time are competing for a limited resource in the hippocampus before memories can be consolidated into areas of the neocortex. Wixted (2004a) also uses

this hypothesis to explain the temporally graded nature of retrograde amnesia, where more temporally distant memories are less prone to retroactive interference, having undergone more cycles of consolidation prior to the onset of retrograde amnesia.

Sloman, Hayman, Ohta, Law, and Tulving (1988) found no difference in the kind of interpolating activity (verbal against non-verbal) in the effectiveness of retroactive interference in recall of primed fragment completion at a week delay, indicating that retroactive interference may be non-specific in nature, implying that it interferes with primary encoding, rather than acting as cue competition at the time of recall.

Frankland et al. (2013) extend Wixted's hypothesis to include an account for the causal role of hippocampal neurogenesis. Frankland et al. contend that short term memories are subject to retroactive interference due to continued long term potentiation of the new neurons, meaning that any new learning that occurs during this phase of the neuron's life cycle will be creating competing associations. Meanwhile, in the long term, integration of new hippocampal neurons into existing hippocampal circuits will result in reduced cue associated recall for previously learned items, hence inducing forgetting. While this does give a plausible mechanism for accumulation of retroactive interference, it does not match with Wixted (2004a)'s explanation of retrograde amnesia, as there is no account of consolidation processes.

Work by Rangel et al. (2014) in rats suggests that over a time scale of about three weeks, neurogenesis in the rat dentate gyrus allows for differentiation of spatial contexts, as inhibition of neurogenesis with a non-lethal agent induced reduced contextual selectivity in the rats. As such, it appears the initial hyperexcitability of the newly formed neuron allows for a specific encoding of a particular spatial environment, as long as significant exposure to other environments does not occur

within the three week window (presumably analogous to the period of hyperexcitability of the new neuron). Akers et al. (2014) also conducted work in a variety of rodents, showing that inducing neurogenesis in adult mice induced forgetting, preventing neurogenesis in infant mice prevented forgetting, and that inducing neurogenesis in guinea pigs and degus (that generate all granule cells prenatally) induced forgetting. These latter results lend support to Frankland et al. (2013)'s hypothesis, that over time the addition of new cells in the hippocampus will tend to drive forgetting. Further, due to the high levels of neurogenesis occuring in infants, this is also explanatory of the phenomenon of infant amnesia.

Déry et al. (2015) conducted human work, using measures of depression and stress (with low levels of each acting as a proxy measure for higher levels of adult neurogenesis), finding better recognition of repeats within blocks, and less false recognition of lures at a two week delay for participants with lower depression and stress scores. This indirectly supports a role for adult hippocampal neurogenesis for providing pattern separation in the short term for newly learned memories, and stabilization for longer term retention.

Sadeh et al. (2014) claim that memories based more on recollection are more prone to decay over time than interference (either due to changes in LTP in the hippocampus or neurogenesis there), whereas memories based on familiarity will be more prone to proactive interference than decay - which would explain findings of the role of proactive interference in the paired-associate forgetting literature. Sadeh, Ozubko, Winocur, and Moscovitch (2016) investigated this hypothesis finding some evidence to suggest support for a contrast in recollection against familiarity in memory. In a high interference condition there was more of an impact on familiarity ratings than

on recollection ratings, whereas decay over time had more of an impact on recollection rather than familiarity. This suggests that target relevant proactive interference impacts familiarity based recall more (due to extrahippocampal encoding), whereas non-specific retroactive interference impacts recollection more due to interference with encoding.

In summary, there is broad consensus that non-specific retroactive interference exposure in close temporal proximity to original learning, is causally implicated in forgetting in long term memory. Proactive interference may have additional effects, and may be more specific to judgments of familiarity of items (and hence recognition responses), rather than explicit free recall.

Temporal Decay in Short Term Memory

Much of the debate about the existence of a temporal decay process in short term memory has been driven by results from word list learning complex span tasks (Lewandowsky, Oberauer, & Brown, 2009) with different intervening distractor tasks between items in order to attempt to independently vary distractor duration and cognitive load. Proactive interference seems to occur in these tasks, as negative probes from a previously learned list are rejected more slowly than novel lures. Lewandowsky et al. (2009) review a range of reports that seem to indicate that proactive and retroactive interference are sufficient to account for forgetting effects in short term memory, with no need for temporal decay. Berman, Jonides, and Lewis (2009) found that adding additional intervening time between trials had very little effect on the reaction time difference induced by proactive interference, but adding a single study-test trial of the same duration eliminated the effect. Lewandowsky, Geiger, and

Oberauer (2008) found very little effect of time (up to five seconds) on subsequent recall, with articulatory suppression in place. Cowan et al. (2006) found no effect of time when children were allowed to free recall world lists at their own pace, as compared to 'as quickly as possible' - the children responded with similar accuracy in both conditions, but took significantly longer in the self paced condition. Oberauer and Lewandowsky (2008) found very little effect of increasing time on accuracy, but increasing the amount of intervening suppressive activity (including decision making tasks) delayed recall of last list items by up to fourteen seconds, but again found no effect on accuracy. Lewandowsky et al. (2009) describe the time-based resource sharing model (TBRS) (Barrouillet, Bernardin, & Camos, 2004) that predicts that short term memory is continually refreshed, unless interrupted suppressed by a high cognitive load task, at which point time based decay will diminish the strength of the representation. Oberauer and Lewandowsky (2008) found that a distractor reduced accuracy, but additional intervening distractors made no additional difference, contrary to the predictions of TBRS. In addition, they reanalyzed Portrat, Barrouillet, and Camos (2008)'s data showing that additional forgetting in the high load condition seemed to be due to increased errors, so self-correction after error may have been responsible for disruption.

Barrouillet, Portrat, Vergauwe, Diependaele, and Camos (2011) concede to Lewandowsky et al. that generally interference is sufficient to explain forgetting in short term memory, but not on timescales sufficiently short to be specifically working memory (i.e. timescales on the order of tens of seconds, and where participants have task demands to maintain some sort of active representation of the memorandum). However, reanalyzing data from Baddeley and Scott (1971) which purports to show a

two trace model (one a temporally decaying trace that disappears after six seconds, and the second a more stable trace), Neath and Brown (2012) are able to account for the changes using their SIMPLE model (see section 1.5.2), implying that the temporal distinctiveness by itself is sufficient to account for the observed forgetting (which implies that it is interference from temporally closely coded representations that are responsible for forgetting). This uncoupling from temporal decay is further backed up by the work of White (2012) who found participants performed much worse on word list recall with an intervening difficult arithmetic task, than an arithmetic easy task. However, an arithmetic task that started difficult and then became easy showed recovery of performance over time, once the easy task had come into play. This tends to imply that retroactive interference is responsible for decay in short term memory performance, as opposed to trace decay over time. However, proponents of TBRS might contend that once the task became easier, the active rehearsal process would increase the memory strength again. In fact, Barrouillet, Paepe, and Langerock (2012) conducted two similar experiments, one with consonant recall, another with spatial location recall, and in both, participants showed better performance with a digit multiplication intervening task, as compared to a word based multiplication task. They conclude that because no interference possible in the spatial task, the longer time taken to do the word task is causally responsible for the increased decay - however, if the retroactive interference is indeed non-specific in nature, this becomes less clearly the case.

Oberauer and Lewandowsky (2013) conducted a range of experiments, finding that complex span memory was unaffected by the time spent on a visual search task. This was then replicated for judgement of spatial fit task (whether a block would fit

through a gap presented at some distance from the block), and no effect of the difficulty of the judgment was apparent on memory. In follow up experiments Oberauer and Lewandowsky increased the time pressure (how much time is available to perform the distractor task) that participants experienced, but with approximately constant cognitive load (the ratio of time that attention is captured by the distractor task, and the total intervening time), this lead to reduced memory performance, potentially explaining previous studies that have shown decreased memory performance due to distractor tasks where cognitive load and time pressure have been conflated. Further, free time following a time pressured distractor task had a beneficial effect on subsequent complex span performance. In their final experiment, Oberauer and Lewandowsky manipulated search set size and time pressure in a visual search distractor task - the former had a strong impact on measured cognitive load, but the latter, none. In spite of this, time pressure had a significant impact on memory, and search set size had none. In all, only within domain interference and time pressure seem to have a detrimental effect on working memory performance, with no place for purely time based decay. Oberauer and Lewandowsky (2014) conducted a further experiment confirming that with fixed cognitive load and time pressure per distractor task, participants suffered from no complex span performance detriment when they engaged with eight intervening distractor tasks as opposed to four, implying no role for temporal decay in working memory.

In summary, in spite of a lively debate and controversy in recent years, the weight of evidence seems to rest with a purely interference based account of short term memory.

The Mathematical Form of Forgetting Data

The last twenty five years have seen increasing interest and discussion about the precise Mathematical form of the forgetting curve. Wixted and Ebbesen (1991) administered a word recall task over a span of less than a minute and a face recognition task over a span of two weeks, both in humans, and a delayed match to sample task (over several seconds) in pigeons. They ran comparison fits across multiple two parameter functions (linear, exponential, hyperbolic, logarithmic, power, and exponential square root - an exponential power function with the power in the exponential fixed at $\frac{1}{2}$) to the empirical data and determined that the fits to data favoured the power law, which, in spite of the mathematical peculiarity of increasing to infinity at time zero, produced the best fit across the data sets.

Extending this endeavour of comparative empirical fits for two parameter candidate functions to fit empirical forgetting data, Rubin and Wenzel (1996) conducted a meta-analysis of 210 data sets from the forgetting literature. Rubin and Wenzel fit not only the candidate functions used by Wixted and Ebbesen, but also a further 105 two parameter functions available for fitting in the "Table Curve 2D" (1994) software package. This investigation further highlighted the inadequacies of a linear model to adequately predict how retention reduces over time, and the exponential and hyperbolic functions appeared to perform less well on fits to data than the logarithmic, power, exponential-square root, and hyperbolic-square root. However, the data were inadequate to distinguish between these four. The conclusion of Rubin and Wenzel is not in favour of any one particular mathematical description, but rather that in order to better characterize how retention reduces over time, more data is required at more time intervals and measured with higher accuracy, in order to disambiguate competing

Table 1.1: Table of candidate functions to fit empirical forgetting data

Candidate Function	Equation	No. of parameters
Linear	$y = a - bt$	2
Exponential	$y = a\exp(-bt)$	2
Hyperbolic	$y = \frac{1}{a+bt}$	2
Logarithmic	$y = a - b\log(t)$	2
Power	$y = at^{-b}$	2
Exponential-Square Root	$y = a\exp(-b\sqrt{t})$	2
Hyperbolic-Square Root	$y = \frac{1}{a+b\sqrt{t}}$	2
Exponential-Power/Weibull	$y = a\exp(-bt^c)$	3
Hyperbolic-Power/Pareto-II	$y = \frac{1}{a+bt^c}$	3
Sum of Exponentials (Rubin, Hinton, & Wenzel, 1999)	$y = a\exp(t/T_1) + b\exp(t/T_2) + c$	5
Two Trace Model (Chechile, 2006)	$y = 1 - b(1 - \exp(-dt^2))(1 - \exp(-at^c))$	4

functions. This conclusion is confirmed by the Bayesian analysis of the same data sets by Lee (2004). Lee follow the prescriptions of Roberts and Pashler (2000), who urge that goodness of fit measure is not very meaningful unless accompanied by an understanding of the flexibility of the theory creating the model (i.e. how much the data set could have varied and still been fit by the theory), and the flexibility of the data set being fit to (i.e. how many different theories the data set would provide a good fit for). Lee use a functional form complexity measure (see Equation 1.1) using a Laplacian approximation of the area of a function's parameter space over which the function continues to give good fits to the observed data.

$$\sqrt{\det\left(\frac{1}{2}\begin{bmatrix} \frac{\partial^2}{\partial m^2}\sum_i(y_i-\hat{y})^2 & \frac{\partial^2}{\partial m \partial b}\sum_i(y_i-\hat{y})^2 \\ \frac{\partial^2}{\partial b \partial m}\sum_i(y_i-\hat{y})^2 & \frac{\partial^2}{\partial b^2}\sum_i(y_i-\hat{y})^2 \end{bmatrix}\right)} \tag{1.1}$$

This measure summarizes the local curvature of the residuals for the fitted function, $\sum_i(y_i-\hat{y})^2$, within the parameter space. The more pronounced the curvature of the residuals in this parameter space, the narrower range of parameters that is able to be fitted to the data available. This can be extended to arbitrarily parameterized functions using an N-dimensional Hessian matrix, however, across different numbers of parameters the complexity will not provide as robust a measure, as the higher dimensional parameter space will reduce the marginal probability of the data occurring given the more highly parameterized model. Using this measure, they conclude that while the hyperbolic function may not produce quite as good a fit as the other functions, the range of its parameters that provide robust fits to the observed data (which in turn implies that the particular parameter selection is less susceptible to noise in the data) mean it is to be preferred as a lower complexity function (beyond the simpler

conception of complexity simply as number of free parameters). In line with Rubin and Wenzel, they suggest that to properly disambiguate the functional forms would require more precise retention data than is currently available in the literature.

Anderson and Tweney (1997) argue, correctly, that arithmetic averaging of data (either across subjects or across items) could cause an inflation of power law fits to observed data, if the underlying data was in fact produced by exponentials with different decay rates (i.e. different rates of forgetting for individuals, or items). However, in their own reanalysis of short term retention experiments examining the impact of proactive interference in a Brown-Peterson paradigm (Brown, 1958; Peterson & Peterson, 1959), while they found some evidence of that arithmetic averaging across subjects inflated the goodness of fit of the power law, use of geometric averaging still gave a better fit for the power law. A similar reanalysis was conducted by Wixted and Ebbesen (1997) in response to the criticisms of Anderson and Tweney, and found that geometric averaging did not overturn the power law advantage Wixted and Ebbesen had found in Wixted and Ebbesen (1991). Further, while Anderson and Tweney (1997) maintain that their failure to find an advantage for the exponential function is a result of occluded arithmetic averaging across items that has occurred early in data processing, Wixted and Ebbesen (1997) show that with decay rates for exponential functions drawn from the normal and exponential distributions, no preferential fit for the power law was found on the averaged exponential functions. The only way such a power law advantage was found was with a Weibull distribution with extreme variability - which would imply that different items have highly differential rates of forgetting. It is also worth reflecting on what a refusal to average across subjects and items would result in - unless the recall task is suited to a graded performance criterion,

for a particular subject performance will have to be averaged across items, as otherwise all that will be found is a series of subject-item singletons that will either be correctly recalled or not (given the perturbing influence of the probe - see subsection 1.5.1 - measurements for the same item cannot be repeated unproblematically, as a correct response will strongly enhance subsequent recall). The conclusions of Wixted and Ebbesen (1997) are backed up by the analysis of Myung, Kim, and Pitt (2000), which show that any wide variation in parameter values for the individual forgetting rates, combined with a non-linear model, and arithmetic averaging, will result in the power law artifact noted by Anderson and Tweney (1997). This is demonstrated by a response curve analysis, whereby arithmetically averaging over data from widely ranging exponential parameters (represented by extrema on the response curve for the exponential function in Figure 1.1) will result in data sets that are closer to being on the power law response curve - extending this graphical reasoning to an N-dimensional space that encompasses the data space for all subjects shows how, with large individual differences, power law artifacts can arise from arithmetic averaging.

Anderson (2001) extended this analysis to consider the results of averaging over item traces that are linear or logarithmic (with asymptotes of performance), as well as exponential and power law components. Their simulations show that in these circumstances, with high variability, then the underlying memory traces (whether different traces for different items, or traces within different systems for the same item) if arithmetically averaged, could give the appearance of an empirical fit to a power law more reliably than any of the underlying components. This simulation work is complemented by analytical work by Murre and Chessa (2011) showing that if an exponential component is used and arithmetically averaged over, and if the learning

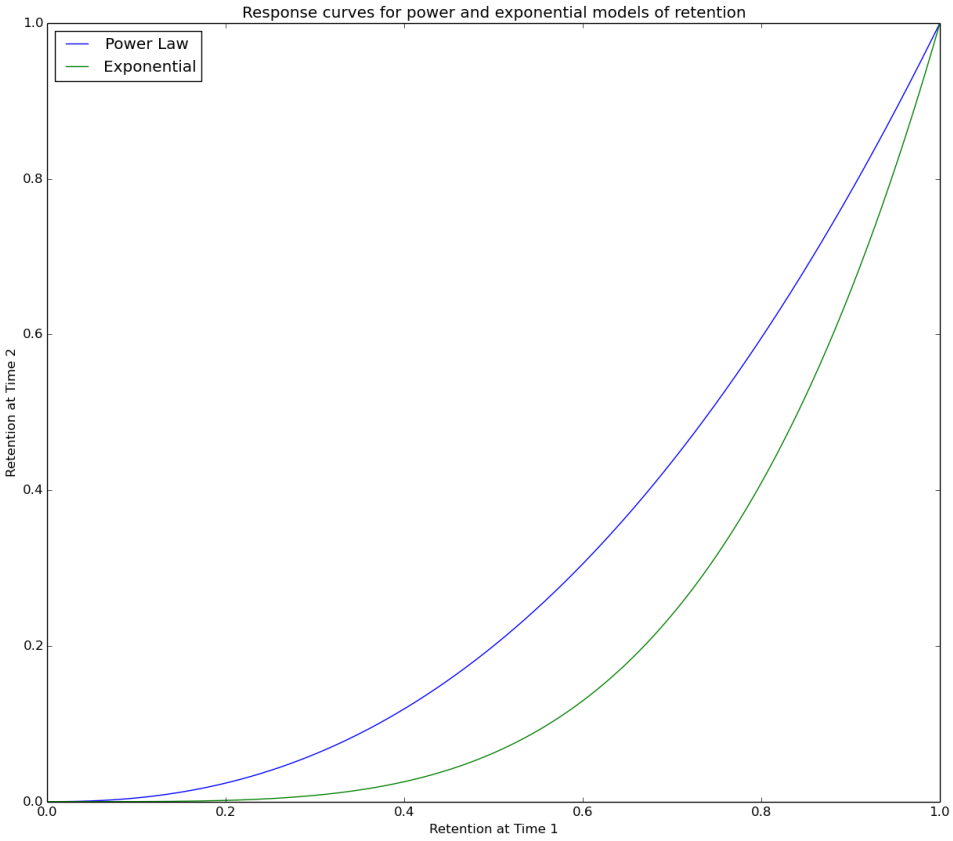


Figure 1.1: Response Curves for Exponential and Power models of retention

rates are distributed according to 'a gamma distribution, a uniform distribution, or a half-normal distribution'(Murre & Chessa, 2011, p.595) then artifactual power law fits will arise.

In addition to concerns about aggregating across subjects that arise from Rubin and Wenzel (1996) (and to a lesser extent, Wixted and Ebbesen (1991)) a theoretically motivated argument is put forward by Wickens (1998). Arguing that a purely empirical approach fails to properly encode potentially important aspects of forgetting, such as heterogeneity in the memorability of different items, consolidation of memory for items, and competition between items for memory resources (both proactive and retroactive interference). They suggest the use of a Pareto-II function (an extension of the Hyperbolic-Square Root function identified by Rubin and Wenzel (1996), but with a free parameter for the power that the t parameter is raised to, rather than fixed at $\frac{1}{2}$) - this is done by construction from multiple component exponential functions to represent different processes for memory (i.e. memory stores that have differential characteristic time scales, and hence different decay rates, and differential decay rates for different items) and then assuming a Gamma distribution for the decay rates. This is in line with the conclusions of Murre and Chessa (2011), who showed that a Gamma distributed mixture of exponential functions would also produce good fits to a power law model.

A similar conclusion is reached by Rubin, Hinton, and Wenzel (1999) in their follow up paper, where they introduce a simpler sum of exponentials to represent working memory, long term memory, and an above chance asymptote (see Table 1.1, where $\log 2T_1$ is the half-life for working memory, and $\log 2T_2$ is the half-life for longer term memory). In fits to data with delays at up to ten minutes, the five parameter

function performed better than previously considered two parameter functions. Rubin et al. attempted to account for the increased complexity by using an adjusted R^2 measure for the additional degrees of freedom, which still gave a better fit, but did not analyze the complexity using the more advanced measures used by Lee (2004), so it is possible that the considerably greater flexibility of the five parameter model is insufficiently accounted for by the adjusted R^2 measure.

A similar Bayesian model selection technique was deployed by Averell and Heathcote (2011) to model data from a cued recall and stem completion experiment covering retention intervals from 1 minute to 28 days. Data sets for individuals across the time range were best fit by exponential functions, and within an experimental session retention was fit well by an exponential and a Pareto-II fit. However, when the Bayesian model selection process was used, the addition of an above chance asymptote parameter to capture the longer term behaviour, led to a greater increase in model complexity for the exponential, as opposed to the power law fit. Using the model selection, a power law with an above chance asymptote of performance was preferred.

As shown in the foregoing, one important factor in the complexity of mathematical fits of functions is the inclusion of an asymptote for above chance performance (if performance is at chance levels, it can be included as an analytically determined parameter, with no extra model complexity). In addition to the findings of Averell and Heathcote (2011), Averell and Heathcote (2009) had previously found an apparently above chance asymptote in word list recognition after very brief exposures, with performance being indistinguishable at seven and twenty eight days, but both significantly different from chance performance. Further, in observational studies of basic science knowledge in medical professionals, for knowledge that had no self report

of rehearsal, Custers and ten Cate (2011) found some retention of learning after more than twenty five years. For further discussion of naturalistic studies see section 1.5.2.

An alternative theoretically motivated argument based on discriminability over time (in parallel to sensory discriminability on other dimensions) by White (2001) endorses an exponential square root function, firstly by arguing that the exponential function has a 'a constant rate of decrement'(White, 2001, p.196) - which presumably refers to the property of the exponential to reduce by the same proportion over an equivalent period - meaning that the current rate of decay is determined entirely by the current probability of recall, and not some more long lived trajectory in memory. Secondly, White argues that the time be scaled to the square root due to a random walk acting on the temporal discrimination process itself. While White presents several data sets that fit this model well, there is no clear advantage over any of the previously considered functions, and the random walk analogy within the discrimination process seems weaker than the memory chain analogy advanced by Rubin and Wenzel, whereby memory is imagined to be a sequence of chains, any one of which could break and prevent recall of the memory - this analogy brings us to the more general Weibull distribution, which in the meta-analysis of Rubin and Wenzel, generally fit to an approximation of the exponential square root function pushed by White.

Hazard Functions

An alternative means of examining forgetting data to understand its mathematical form is advanced by Chechile (2006), the mathematical function usually determined by aggregation of observation data is the cumulative probability function of the failure of memory - i.e. how likely is it that the memory will have failed after

a certain amount of time. A parallel consideration to the cumulative failure rate is the instantaneous failure rate of the process, i.e. given that this particular memory has not failed so far, what is the probability density of failure of the memory. The exponential function has a constant hazard function, with the consequence that the future likelihood of failure is unrelated to the time since the memory was formed, but is a consistent function of any previous likelihood of failure. Most other candidate forgetting functions (c.f. most functions in Table 1.1, although some need modification to allow a hazard function that is defined at zero, such as the power function) have monotonically decreasing hazard over time, with the consequence that failure of a memory becomes less likely within the same time span as time without failure from the original memory formation increases. The hazard function is mathematically defined as:

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \tag{1.2}$$

where $f(t)$ is the probability density function of failure, and $F(t)$ is the cumulative probability function for failure (and hence the $1-F(t)$ denominator is the probability of survival). However, there are considerable empirical challenges to measuring the values of the hazard function directly, due to the fact that the probability density function, $f(t)$, requires very temporally fine grained measurements in order to estimate (as it is a density function, small changes over time are needed to estimate the instantaneous density, whereas the cumulative probability function can be estimated from more distributed data). To address this issue, Chechile shows, by application of L'Hôpital's rule, that for a hazard function to be monotonically decreasing, the function $u(t) = \frac{F(t)}{t}$ must be at a maximum at $t = 0$. As this function is a derivative of the cumulative probability density function that is ordinarily estimated experimentally, this is more

achievable. Using a modified Brown-Peterson procedure (Brown, 1958; Peterson & Peterson, 1959), Chechile showed evidence that the peak was very shortly after $t = 0$ rather than directly at it, indicating that when measured over very short time spans, the hazard function is not monotonically decreasing. This has the consequence of suggesting that memory over very short time scales rapidly accumulates failure probability over time. Taking this, and the possibility of above chance asymptote of long term memory store into account, Chechile proposes a two trace model, one that allows for this very rapidly decaying and volatile trace (a Weibull cumulative distribution function: $F(t) = 1 - exp(-dt^2)$), and the other that allows for longer term storage (a Weibull subprobability distribution: $F(t) = b(1 - exp(-at^c))$, where c is less than 1), the resulting cumulative failure probability is the product of these two cumulative failure probability functions:

$$F(t) = b(1 - \exp(-dt^2))(1 - \exp(-at^c)) \tag{1.3}$$

In summary, current empirical data is amenable to fitting a wide range of different mathematical forms, with particular fits being particularly amenable to certain aggregations of data (for example, the power law). Empirical fitting unmotivated by any theory, will provide very little insight, therefore, except which form may be of most use in predictions of retention. For the purposes of computer assisted instruction, these kinds of empirical fits may be sufficient to provide useful predictions about future performance, but more theoretically grounded models will be of more use in attempting to better understand the more detailed time course of memory (see section 1.5.2 for further discussion of Cognitive Models.)

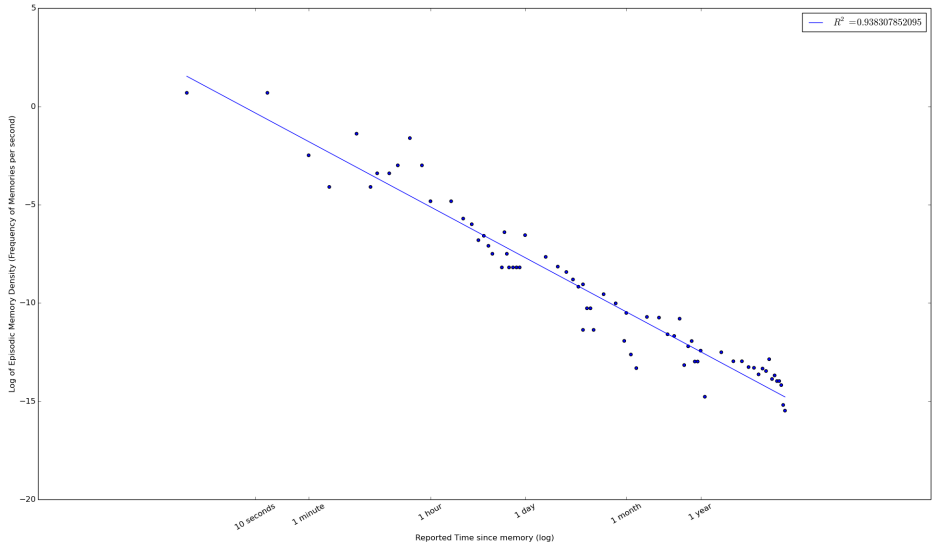


Figure 1.2: Fit of episodic memory density over time calculated from Crovitz and Schiffman (1974)

Forgetting in the Wild

In an attempt to study forgetting in a more ecologically valid, naturalistic setting, researchers have examined the density of recallable episodic memories, and then used the density of recalled memories over time as an index of forgetting over that time period. Such an experiment by Crovitz and Schiffman (1974) found a power law relationship between the frequency of memories and the time delay at which they occurred. The use of absolute frequency, rather than a density measure of frequency of memories per unit time, makes these results questionable, however, as it causes older memories to have more weighting than newer ones because they are aggregated over a longer period. Working from their original reported data, and calculating a density metric, it is possible to see that a log-log linear relationships still pertains (see Figure 1.2), fitting with their previous conclusion of a power model fit.

Rubin (1982) followed a similar methodology to Crovitz and Schiffman (1974)

(except that Rubin properly used a temporal density metric, rather than raw frequency), and found a similar power law distribution for autobiographical memories elicited by a range of methods.

Another autobiographical approach was taken by Bahrick, Bahrick, and Wittlinger (1975), with people asked to free recall names of people from pictures in their high school yearbook. While they demonstrated the general applicability of reduction of recall over time in this cross-sectional approach (people for whom high school graduation was in the much more distant past had somewhat poorer recall), the data do not seem to lend support for a unified mathematical model of declarative forgetting over time, as Rubin and Wenzel's later 1996 analysis of this data shows, a power model fit did not explain as much of the variance as a linear or exponential fit - implying that the exact time course that forgetting takes may well depend on other factors, such as how overlearned the original memories were, as they were in this case, where name and face associations would have been repeatedly learned in high school.

Roediger and DeSoto (2014) extended this cross-sectional autobiographical memory paradigm to examine the retention of the names and temporal sequencing of past Presidents. Their data seem to lend support for differential rates of forgetting for different Presidents - with Kennedy showing far higher persisting recall probability than other Presidents who came shortly before or soon after him. This and their other results for convergent asymptotes of recall probability across three generations (whereby Presidents from Coolidge to Washington show consistent recall probability across generations) highlight that forgetting curves calculated in naturalistic studies are representative of the broader information environment - Kennedy will show slower 'forgetting' in the population, due to the persistence of mentions of him in media and

broader culture.

In contrast to autobiographical, and broader cultural knowledge, such environmentally induced rehearsal of academic skills will tend to vary far more from individual to individual. A broad review of retention of academic knowledge in naturalistic circumstances was conducted by Custers (2010), unfortunately, while Custers correctly notes in his review of studies in medical students that 'In general, differences in retention between basic sciences can probably be accounted for by differences in rehearsal or reinforcement during the RI' (Custers, 2010, p.122), he takes at face value the appealing conclusions of a range of studies by Bahrick that claim a three stage forgetting process, particularly for material that has been learned and rehearsed over a period of around three years - rapid, exponential decline for up to six years, relatively stable performance for up to thirty years, finally followed by a steeper loss in performance, probably related to age. However, for this middle period of 'permastore' (Bahrick, 1984) to be validated, there would have to be no intervening rehearsal, and above chance performance in all these studies.

The data reported in Bahrick (1984) seem to show that with very little self-reported rehearsal, some level of performance was retained from college Spanish classes over long durations, with the strongest predictor of delayed performance being the highest class level. In a less clear study, Bahrick and Hall (1991) looked at how high school Algebra skills were maintained over time. Participants were recruited across a range of ages across the adult life span. Factors considered in performance on the administered Algebra test were: the highest level of Mathematics class taken, the retention interval since high school Algebra, a coding for intervening rehearsal (a sum of rehearsal from reported leisure and professional activities), and gender. From their

Table 1.2: Levels of Mathematic Rehearsal for Different Levels of Class Taken, extract from Table 2 of Bahrick and Hall (1991)

	Level of Rehearsal					
Highest level of Mathematics	0	1	2	3	4	5
Below Calculus	445	135	51	9	10	1
Calculus	102	41	5	2	3	3
Above Calculus	18	6	4	0	1	4

hierarchical regression Bahrick and Hall conclude that the level of class taken is the most important factor (accounting for roughly 40% of the variance observed in recall on the test) in common with the conclusions of Bahrick (1984), drawing the further conclusion that this extended acquisition period for Algebra (whereby they must reuse Algebra knowledge in subsequent Mathematics courses) is solely responsible for the maintenance of performance in Algebra for up to 50 years. While they maintain that this result holds even for participants who have had no rehearsal in the intervening years, a Pearson's chi-squared test (Pearson, 1900) of the independence of amount of rehearsal compared to level of Mathematics (see data used in Table 1.2) clearly indicates that the two variables are not independent ($p < 0.0001$), implying that the order in which these variables were added to the hierarchical regression will have an impact on the seeming contribution. As such, these data seem insufficient to draw the strong conclusion that three years of extended study are sufficient to prompt lifelong learning, although continual lifelong rehearsal clearly is.

Conway, Cohen, and Stanhope (1991) found above chance recognition, and greater than zero recall performance on facts and concepts learned in a Cognitive Psychology class, up to ten years after original learning. Unlike Bahrick (1984), Bahrick and Hall (1991), there was little variation in level of learning as all students took

only one class in Cognitive Psychology, although a small amount (10%) of variance in retention was explained by course work grade (Conway, Cohen, & Stanhope, 1992) - the insensitivity of retention to the final course grade was interpreted by Bahrick (1992) as being indicative that the temporal sequencing of initial learning affects the asymptote of later retention. Broadly, this conclusion seems consistent with accounts of the testing effect and repeated learning (see subsection 1.5.1), and are further supported by retention recorded over four and eleven months by Semb, Ellis, and Araujo (1993), and over sixteen years by Ellis, Semb, and Cole (1998) - both of whom showed improved retention induced by students acting as tutors, independent of student achievement. Unfortunately, missing from the studies of both Semb et al. and Ellis et al. is any attempt to capture the amount of possible rehearsal that may have occurred in the intervening period, or to ascertain if this course material was subsequently used for further courses (mirroring Bahrick and Hall (1991)). Finally, a cross-sectional study of medical students, clerks, and doctors (Custers & ten Cate, 2011) found retention of basic Science knowledge learned in medical school for up to fifty-five years after graduation on a short answer test. Their results broadly mirrored the pattern reported by Bahrick (1984), however, their reliance on self-report of rehearsal of specific items (as opposed to the more general measures used by Bahrick (1984)) on their test seems inadequate to make their claim that unrehearsed knowledge followed precisely the same pattern as Bahrick (1984). It seems more likely that a significant proportion of basic Science knowledge acquired in medical school is at least occasionally rehearsed through a doctor's career.

Recent work by Ridgeway, Mozer, Bowles, and Stone (2016) has examined relatively short term forgetting in a more ecologically valid setting than a traditional

laboratory experiment, within a dataset from the Rosetta Stone language learning software. The learning content consists of a mixture of vocabulary and grammar learning, and Ridgeway et al. (2016) found some evidence to favour the power law over the exponential-power law in this data set. This kind of work opens up the possibility that investigation of more ecologically valid learning environments, with large associated datasets, may be able to shed additional light on the empirical form of the forgetting function. The evidence of such forgetting in naturalistic learning motivates further investigation of forgetting and spacing effects in curricular contexts. In addition, as well as showing evidence the power law model, Ridgeway et al. (2016) used data driven techniques to use additional information about students in predictive models, as is standard practice in intelligent computerized tutors. The improved predictions based on the combination of both the power law predictions and data driven predictions show that there is additional information to be discerned from the use of theoretically driven modeling, and that it is not just completely colinear with existing information that can be inferred from a data set of user interactions.

In summary, it seems that forgetting in more naturalistic settings may frequently result in asymptote performance above chance, depending on the precise nature of original learning. Further more, a power law model seems to provide good performance in predicting forgetting over time scales that are relevant to formal education contexts (Custers, 2010; Ridgeway et al., 2016) for declarative fact learning, however, the decline in performance in Algebra for those who did not take further Mathematics courses seems to be consistent with a range of potential functions, it remains an open question, therefore how best to predict the time course of forgetting for Mathematical learning.

Cognitive Models of Forgetting

In order to give more theoretically grounded accounts of memory, rather than simply searching for the functional form of empirically observed data (c.f. section 1.5.2), it is possible to build cognitive and neurocognitively inspired models, and test their predictions against observed data. Three such models, each motivated by different aspects of memory, are described here.

Memory Chain Model

This neurocognitively inspired model, first advanced by Chessa and Murre (2002) and later developed in Chessa and Murre (2007) to model spacing regimes in advertising campaigns, assumes two distinct memory stores, one mapping to initial hippocampal encoding, the second to neocortical representations. In addition, consolidation from the hippocampal store to the neocortical is assumed to take place over time, with the rate of consolidation proportional to the strength of the hippocampal store. A new representation for an item is created for every exposure to a stimulus, with the representation in each store modeled as a nonhomogenous Poisson point process with an exponential decay function. As such, in Chessa and Murre (2007) the intensity of the representation in the hippocampal store is modeled with the following form, where μ_1 is the initial encoding intensity, and a_1 is the decay rate of the trace:

$$r_1(t) = \mu_1 \exp(-a_1 t) \tag{1.4}$$

The neocortical analog second store is more complex, due to its interrelation with the intensity of the first, hippocampal store analog (again, μ_2 is initial intensity of

encoding, and a_2 is decay rate):

$$r_2(t) = \frac{\mu_1 \mu_2 a_2}{a_1 - a_2}(\exp(-a_2 t) - \exp(-a_1 t)) \tag{1.5}$$

where the decay rate parameter a_2 for the second store is assumed to be lower than that for the first store $a_2 < a_1$. The resulting intensity of the representation is a simple sum of the two store intensities:

$$r(t) = r_1(t) + r_2(t) \tag{1.6}$$

Lastly, the probability of recall in the case of a single representation is given by:

$$p(t) = 1 - \exp(-r(t)) \tag{1.7}$$

which, expanded, becomes:

$$p(t) = 1 - \exp(-\mu_1 \exp(-a_1 t) - \frac{\mu_1 \mu_2 a_2}{a_1 - a_2}(\exp(-a_2 t) - \exp(-a_1 t))) \tag{1.8}$$

For multiple learning trials, it is assumed that there is a maximum intensity r_{max}, which is greater than μ_1, that can be accumulated in the first store during initial encoding, with learning duration l and learning rate v, the initial encoding intensity is given by:

$$\mu_1(l) = r_{max}(1 - \exp(-\frac{vl}{r_{max}})) \tag{1.9}$$

This simplifying assumption rather undercuts the intuition that would seem logical in driving their independent modelling of different encoding of the same item - that

each encoding lays down a separate trace, instead each encoding 'tops up' the current intensity of encoding, as shown by equation (6)(Chessa & Murre, 2007), which shows the intensity accumulated during a learning trial for the Lth learning trial:

$$\mu^{(L)}(l) = (r_{max} - \sum_{i=1}^{L-1} \mu_{(i)} \exp(-a_1(t_L - t_i)))(1 - \exp(-\frac{vl}{r_{max}})) \tag{1.10}$$

which, has the intensity during encoding reaching a peak sum of r_{max}, however, as the rate of exponential decay depends only on the current intensity, this new sum of r_{max} would decay exactly like a single trace, as all the constituent traces in this sum share the same decay constants. The resulting recall probability, for a total of L learning trials, is:

$$\begin{aligned} p(t) = 1 - \exp\Bigg(- \sum_{i=1}^{L} \mu^{(i)} (\exp(-a_1(t_L - t_i + t)) \\ + \frac{a_2}{a_1 - a_2} (\exp(-a_2(t_L - t_i + t)) - \exp(-a_1(t_L - t_i + t)))) \Bigg) \end{aligned} \tag{1.11}$$

The above model provided good fits to recognition data for spaced and massed exposure to advertising material (Chessa & Murre, 2007), it is possible, by inspection of the model, to understand how spaced exposure would be preferred by the model - as exposure on too short a time scale would result in a very low encoding intensity (as the previous sum of encoding intensities would still be close to r_{max}), which, in turn, would mean that the highest achievable encoding intensity in the secondary, slower decaying, store would be far lower. In further model fitting, Murre, Chessa, and Meeter (2013) found a very good fit to paired associates data from Rubin et al. (1999), using the single trace model, Equation 1.8 (Rubin et al. found a similar but simpler sum of exponentials to provide the best model fit for their data also). Murre et al. conducted

further model fits to both human and animal data for subjects with atypical function (either as a result of lesions or neurodegenerative disorders), good fits were found, again using the single trace model, Equation 1.8, but it is unclear from their analyses how much of the goodness of fit, particularly in the human data (where experimental counterbalancing is not available), is a result of the noise inherent in the data, and hence its ability to fit a wide range of potential models (c.f. Roberts and Pashler (2000)). In summary, the model seems to provide a coherent account of forgetting, with the capacity to explain multiple exposures to the same item, while being independently motivated by consolidation and a plausible neurocognitive account of initial encoding and longer term memory. However, contrary to section 1.5.2, the model makes no attempt to lend a mechanistic role to interference, seemingly preferring a temporal decay mechanism - however, this may be a simple matter of convenience, as the original formulation was created for advertising data (Chessa & Murre, 2007), and, so, on relevant timescales, measurement of intervening interference was simply not possible.

Temporal Ratio Model

The Temporal Ratio Model advanced by Brown, Neath, and Chater (2007), which they name SIMPLE, in contrast to the Memory Chain Model, makes no assumptions about the underlying neurocognitive instantiation of memory, instead, similarly to White (2001), preferring an analogy between memory and sensory discrimination as its basis - with time being one relevant dimension along which memories are distinguished. As such, and again in contrast to the Memory Chain Model, the model assumes no role for memory trace decay in forgetting - instead, difficulty in recall

comes from a difficulty in disambiguating two memories that are close on the temporal dimension. The final important assumption of the model is that items become less distinguishable as they become more distant from the present - so, two memories that were formed thirty seconds apart will be more distinguishable ten seconds after the second memory was formed, than ten days after the second memory was formed, as such the confusability of two items is proportional to the ratio of their temporal lags from the present. In the foregoing example, $\frac{10}{40}$ gives a much smaller confusability than $\frac{864000}{864030}$. As such, the same feature that we see in empirical studies of forgetting, of items becoming more difficult to recall as they retreat into the past, is recovered on the basis that interfering memories accumulate proportional to time passing. The confusability can be more formally recast as psychological similarity, which makes the exponential nature of relationship more evident:

$$\eta_{i,j} = \exp(-c|M_i - M_j|) \tag{1.12}$$

This similarity measure could be extended to a multi-dimension measure, which would allow for greater nuance in interference from other memories, aside from temporal disambiguation. A single memory trace will therefore be discriminable as an inverse proportion of the sum of its similarity to all other memories:

$$D_i = \frac{1}{\sum_{j=1}^{n} \eta_{i,j}} \tag{1.13}$$

For constrained serial recall, where participants are probed for correct ordering of stimuli, and only errors of displacement and not omission are possible, Equation 1.13

will also yield the probability of correct recall. In the case of free recall, when omission errors can occur, they assume that items with low discriminability (Equation 1.13) will be the most likely not to be recalled, yielding:

$$P(R_i|D_i) = \frac{1}{1 + \exp(-s(D_i - r))} \tag{1.14}$$

with s as the slope of the sigmoid, and r as the threshold for recall. Brown et al. found good fits to a range of historical data for constrained and free serial recall tasks, in the former, only one free parameter was needed, c from Equation 1.12, while for free recall tasks, s and r from Equation 1.14 were included as free parameters. In addition, follow up work has shown strong fits to data in short term memory (Lewandowsky, Nimmo, & Brown, 2008; Oberauer & Lewandowsky, 2008; Lewandowsky et al., 2009; Morin, Brown, & Lewandowsky, 2010; Neath & Brown, 2012), and simulations using SIMPLE have reproduced consolidation effects across multiple timescales (Lewandowsky, Ecker, Farrell, & Brown, 2012). Finally, SIMPLE, like the Memory Chain Model, assumes that each new learning event encodes a new memory trace, however, unlike the Memory Chain Model, there is no immediatley apparent interaction between previously formed memories, and the newly formed trace - any testing or spacing effects must therefore arise as a consequence of 'eggs in more baskets' approach, where having the memory encoded at multiple locations on the temporal dimension increases the net discriminability.

Due to its rejection of decay as a mechanism in memory, and placing interference as the sole causal agent in forgetting, SIMPLE is an appealing model - however, beyond interference, it makes little contact with the underlying mechanisms that might account for its effects (c.f. section 1.5.2). While SIMPLE can qualitatively predict consolidation

across multiple timescales from its assumptions (Lewandowsky et al., 2012), it provides a unifying mathematical description of what might well be quite disparate processes. This may well be the same in the case of its accounts of forgetting, where the Memory Chain Model has a more neurobiologically appealing analog available. A proponent of SIMPLE may wish to make the claim that it is merely formulating regularities that are present due to, for example, temporal encoding in the hippocampus as a result of neurogenesis (Rangel et al., 2014), but, assuming that such encoding takes place at a specific timescale (which appears to be selective around the three week mark for rats (Rangel et al., 2014)), it is not clear why SIMPLE predicts such a continuous and exponential drop off in discriminability as time lag increases. Further, the formulation of SIMPLE described here is restricted to modelling word list serial recall paradigms, which are far from covering all possible memory tasks. While its appeal to interference as the sole causal mechanism is appealing, the inability to exhaustively measure all interference (or even to characterise the relevant dimensions for interference in memory by which we would have to multidimensionalize Brown et al.'s similarity metric, Equation 1.12), means that it is not clear how to extend the model to more applied domains, or longer time scales where the interference is not controlled - without a probabilistic model that reduces interference to a function of time.

Multiscale Context Model

Akin to the Memory Chain Model, the Multiscale Context Model was primarily advanced to model spaced learning, and to predict optimal spacing regimes for learning declarative facts. Pashler, Cepeda, Lindsey, Vul, and Mozer (2009) use two parallel

descriptions of the model to describe its behaviour, in one, it is cast as a neural network, in the other as a set of leaky integrators. The mathematical description of the model is similar to a sum of exponentials, as proposed by Rubin et al. (1999) and Murre et al. (2013), but without the limitation on the number of units imposed by psychological or neurocognitive models of memory, and with an added x_i term that represents the current activity of the integrator, however, if $x_i(0)$ is set to 1 at initial encoding, then a sum of exponentials is recovered:

$$s_N(t) = \sum_{i=1}^{N} \gamma_i \exp(-\frac{t}{\tau_i}) x_i(0) \tag{1.15}$$

Pashler et al. (2009) refine the representation of their formulation in order to give a good approximation to a power function leading to the following substitutions:

$$\tau_i = \mu\nu^i, \gamma_i = \frac{\omega\xi^i}{\sum_{j=1}^{N} \xi^j} \tag{1.16}$$

where $\nu > 1$ and $\xi < 1$ - as such, for high i the characteristic decay time (τ_i) will be much larger, but the contribution to the weighting in the sum for the trace, γ_i, will become much smaller. Going forwards, I will represent the current activity of the integrator as $x_i(t)$, as this makes the update rule more easily parseable. This is more simply construed as:

$$s_N(t) = \sum_{i=1}^{N} \frac{\omega\xi^i}{\Xi_N} x_i \tag{1.17}$$

where x_i is the current activity of the integrator, given by:

$$x_i = \exp(-\frac{t}{\mu\nu^i}) \tag{1.18}$$

and Ξ_i is the sum of the powers 1 to N of ξ:

$$\Xi_N = \sum_{j=1}^{N} \xi^j \tag{1.19}$$

The probability of recall is given by $min(1, s_N)$, meaning that integrator activity can increase in a potentially unbounded way. This quirk of the model leaves open the possibility of the kind of long term 'permastore' of items reported by Bahrick (1984). The update rule is applied to the activity of the integrator, when a retrieval attempt is made:

$$\Delta x_i = \epsilon(1 - s_i) \tag{1.20}$$

where s_i is given by the following formula:

$$s_i = \frac{\sum_{j=1}^{i} \gamma_j x_j}{\sum_{j=1}^{i} \gamma_j} \tag{1.21}$$

which, due to the cancellation of the ω and Ξ_N terms, further simplifies to:

$$s_i = \frac{\sum_{j=1}^{i} \xi_j x_j}{\sum_{j=1}^{i} \xi_j} \tag{1.22}$$

and $\epsilon = 1$ for unsuccessful retrieval and $\epsilon = \epsilon_r > 1$ for a successful retrieval. Essentially, this means that each integrator is updated according to the weighted sum of the current activity of all the shorter time scale integrators that are being modeled. The Multiscale Context Model provides good fits to empirical forgetting data (unsurprisingly given the

use of a sum of exponentials that approximates the power function), but also provides accurate quantitative predictions of the impact of additional learning trials (using an update parameter value of $\epsilon_r = 9$, which was hand tuned on one data set, but then applied to six others without further tuning). The Multiscale Context Model can be readily applied to any learning paradigm, which is amenable to some measure of recall, and as such is appealing from the stance of practical application (see subsection 1.5.3 below for discussion of Lindsey, Shroyer, Pashler, and Mozer (2014)'s work in this area) and also generalizability across paradigms and time spans. However, unlike the Memory Chain Model it does not posit any biological analog that might be responsible for the leaky integrators that have temporally distinct characteristics. Additionally, it employs temporal decay as the causal driver of forgetting, at odds with the consensus of the literature (see section 1.5.2), however, as discussed in section 1.5.2 the difficulty of measuring interference effects over anything but the shortest timescales, mean that for practical purposes modelling must be done based on the passing of time, rather than the accumulation of interference - however, this does run the risk on examination of such models that assumptions operative in the model that are actually about the nature of the probabilistic accumulation of interference, are mistaken for assumptions about the nature of memory itself.

Summary

Each of the cognitive models explored in the foregoing give good model fits to the data in their domain. SIMPLE, while appealing that it places interference at the center of the causal picture of forgetting, is somewhat limited in its broader application (the main work looking at longer timescales, as in consolidation on multiple

timescales is that by Lewandowsky et al. (2012), which uses simulations of the model to make qualitatively similar curves to laboratory data) due to the difficulties of empirically measuring interference effects, and finding the relevant dimensions for similarity. As such, the two models that employ temporal decay, the Memory Chain Model, and the Multiscale Context Model are applicable in longer time scales and more practical domains, such as memory for advertising campaigns (Chessa & Murre, 2007) and Spanish vocabulary learning (Lindsey et al., 2014). The key surface feature that is different between the two is the use of consolidation in the Memory Chain Model. It is possible that the Multiscale Context Model could be embellished to add consolidation, but it is unclear whether it would be best modelled as a stochastically occurring update (perhaps with a weaker update multipler, ϵ), or modelled in the continuous manner used by the Memory Chain Model, which may be inappropriate without a clear analog between the leaky integrators of the Multiscale Context Model and possible neurobiological analogs that it would be appropriate for consolidation to occur between.

1.5.3 Spacing, Massing, and Optimizing Review Schedules

Until relatively recently in pedagogical practice (as shown by the design of Mathematics textbooks), it was thought that the most efficient way for a student to learn Mathematics in a way that facilitated later retrieval was overlearning - the continued practice of a procedure after mastery has been achieved. This massed (as opposed to spaced) practice model explains the design of Mathematics textbooks, where, by chapter, exercises are massed by a small number of procedures that need to be applied. By contrast, a spaced learning methodology would require intermingled

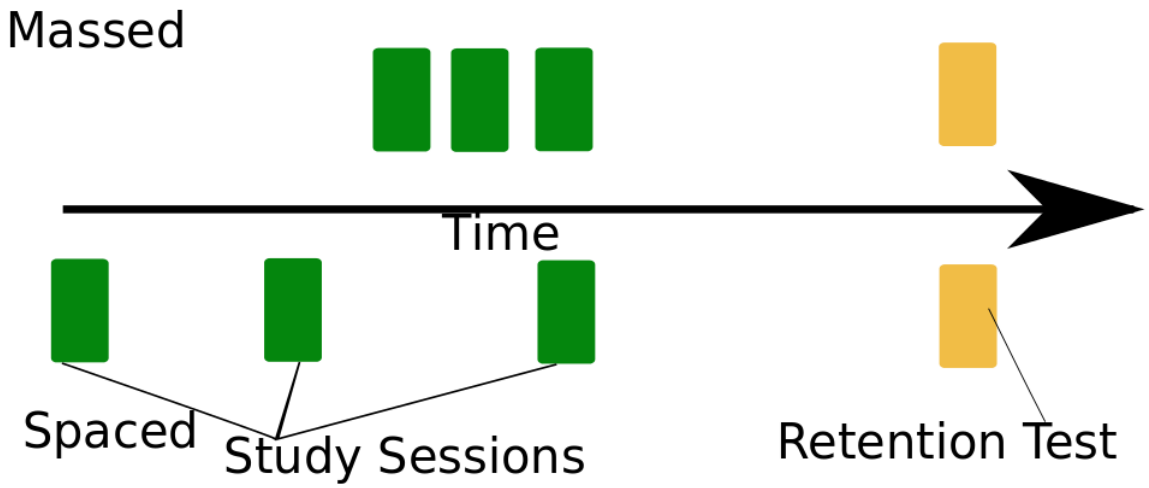


Figure 1.3: Spaced vs Massed Learning regimes

exercises requiring application of different kinds of procedure, but with procedures recurring multiple times over several study sessions.

Spacing has been a core component to the recent advances in our understanding of the Science of Learning. Rohrer and Pashler (2007), drawing on work by Rohrer and Taylor (2006, 2007), identify the empirical support for using such spaced learning episodes (as opposed to the usual massed practice encouraged by the design of Mathematics textbooks and many online tutoring programs) in the learning of mathematics. While Rickard, Lau, and Pashler (2008) examined the role of spacing in promoting retrieval over calculation in mathematics, later work has focused almost exclusively on declarative fact learning (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008, 2009; Lindsey, Mozer, Cepeda, & Pashler, 2009; Lindsey et al., 2014; Pashler et al., 2009; Lewis, Lindsey, Pashler, & Mozer, 2010; Sobel, Cepeda, & Kapler, 2011). This work has shown that for retention at longer time intervals, massing the study sessions (as seen in Figure 1.3) will lead to poorer performance at a temporally distal retention session.

In addition, the modeling and empirical work of Pashler et al. (2009) shows that

the optimal particular spacing regime deployed during learning is dependent on exactly when the retention test occurs - roughly speaking, the further out the anticipated retention test, the wider the spacing that should occur between learning sessions (and each session should be scheduled with an exponentially expanding interval). For highly temporally proximal retention tests (as is common in many classroom learning environments, particularly with the implementation of year end high stakes tests) this implies that the optimal learning strategy is precisely that already employed intuitively, cramming. However, tests that occur in the future (when students might actually be asked to use what they have learned in school) will be highly negatively impacted by these kind of learning regimes.

The work on declarative fact learning has shown considerable learning gains in applied contexts by the use of optimized review strategies (Sobel et al., 2011; Lindsey et al., 2014), with delayed retention tests by Lindsey et al. (2014) (following a long summer vacation) showing increases in retention by using spacing regimes that followed the mathematical models of Pashler et al. (2009), rather than more simplistic spacing regimes that did not account for individual student performance (such as might be provided through whole group instruction). The combination of these demonstrated benefits of the application of spacing effects to fact learning, along with the early results of Rohrer and Taylor (2006, 2007) that show that the same kind of spacing phenomena may produce similar results in Mathematics learning, compels verification and testing of similar techniques in Mathematics learning.

In addition, Rickard et al. (2008) found that the use of a small number of highly repeated exercises, gave a greater shift to a retrieval strategy (as indexed by lower reaction times) during the training period, but showed a reduced use of retrieval

strategy during a test period, when compared to training with a larger number of exercises that were repeated less frequently. This shows that in addition to the temporal spacing effects outlined above, there may be additional effects of interleaving practice, as opposed to massing practice, within an individual session. While the effect in Rickard et al. (2008) seems to be more akin to declarative fact learning (where a retrieval of the result of a previously performed calculation is occurring), it echoes the work of Kornell and Bjork (2008) and Kornell, Castel, Eich, and Bjork (2010), which primarily examined inductive category learning by examples - where interleaved examples gave better performance than massing examples of a single type.

Stafford and Dewar (2014) have extended the spacing analyses carried out above to a domain beyond declarative fact learning by looking at the impact of spacing in the context of a fast-paced computer game. The game requires players to quickly parse a visual scene, and then react with mouse clicks to set the next target for migration of a neuronal axon - players show learning within a single session and across multiple sessions. In three separate analyses, Stafford and Dewar (2014) showed that spreading out practice across a larger time frame increased the scores achieved on the game (using a bootstrapped distribution to provide a control for the number of attempts); that players who played more than fourteen times, and who spread their first ten plays over more than 24 hours, as opposed to massing them in 24 hours, scored higher on subsequent plays; and that players who played their first 6 plays within two hours and their 15th-20th plays in a separate two hour window, and are players with similar habits, had divergent learning curves, when subdivided into a group who spaced their practice by over 6 hours and a group that did not - the group of players that began to space their practice after the 6th play improved more rapidly

than the group who that did not.

1.5.4 Interleaving against Blocking Instruction

The temporal spacing of learning seems to be important - however, as discussed in section 1.5.2, it appears quite plausible, especially over longer time scales, that time itself is merely strongly correlated with the actual causal mechanism by which learning degrades over time, interference. As such, in addition to temporal spacing, it may well be equally important to how well something is remembered, how much interfering material is learned in close temporal proximity to the target item. In addition, if we take the natural extension of the SIMPLE model (see section 1.5.2), and add further dimensions of similarity beyond time of learning, it seems plausible that learning very similar materials at the same time may actually increase interference and hence degrade retention (whether by interrupting encoding, or through cue competition at the time of recall).

Interleaving and Blocking Mathematics Instruction

Due to current practices in Mathematics (where students frequently practise on large blocks of questions that all make demands on application of the same Mathematical procedure), work has been done to examine the impact of interleaving the practice of Mathematical skills - Mayfield and Chase (2002) required participants to learn five basic algebra rules with which they were unfamiliar (as established by a pretest), through blocked practice of the skills. However, one group had review sessions that combined multiple skills in a single fifty question review, while other groups had review sessions only on a single skill at a time. A subsequent post test

showed better performance on application problems where participants had to choose the right skill and apply to the problem.

Rohrer and Taylor (2007) found similar results from interleaved practice of multiple novel mathematics problems, where after initial learning, subsequent practice was interleaved or blocked by problem type. Participants in the interleaved condition performed significantly better on a one week delayed post test. Taylor and Rohrer (2010) followed up on this study, due to concerns about the confounds of temporal spacing effects in the results (by interleaving practice, without additional controls, temporal spacing is induced due to the intervening of other problem types), in this study, by adding filler distractor tasks into the blocked condition, time between practice of the same problem type was equated across conditions. Again, a positive effect of interleaving was found, on a one day delayed post test, but only for correctly identifying the problem type, and not for actual execution of the required task. This implies that the positive effects of interleaving on Mathematics problems may be due to increased discriminability of problem types, rather than enhanced application of the correct procedure once chosen. As the problems in this experiment had a high level of superficial similarity (they were all about calculating different features of prisms), it is possible that the positive effect of interleaving would disappear from more dissimilar Mathematics problems.

Rohrer et al. (2014) conducted a classroom based study on seventh graders in order to attempt to verify the existence of the effect in more ecologically valid conditions. Participants had their usual Mathematics lessons that were supplemented with either interleaved or blocked practice (the manipulation was within subjects and counterbalanced, with one half of students receiving one half of the problems

interleaved, while the other half received them blocked, and vice versa). Assessment was made by means of a surprise test two weeks after the nine week instruction period. Participants performed nearly twice as well (72% against 38%, $d = 1.05$) on problems that they practised in an interleaved manner, in spite of the fact that the problems were not superficially similar as in Taylor and Rohrer (2010). In addition, in spite of the fact that the interleaved practice induced a spacing effect, the problems for which there was the largest difference in delay between the last presentation of the item and its test showed the smallest effect for interleaved practice - rather, the items learned closest to the test showed the largest effect of interleaved practice, implying that the spacing effect alone could not be responsible for the observed effect. Rohrer, Dedrick, and Stershic (2015) replicated this study with another group of seventh grade participants, this time testing either at a one day delay, or a thirty day delay. While the effects were more modest in this replication ($d = 0.42$ for one day delay, $d = 0.79$ for thirty day delay), interestingly, the effect was more pronounced at longer delays, indicating that the interleaved practice has an effect that helps reduce forgetting over time (perhaps because recall is less effortful during blocked practice), further, the previous finding at two weeks may have simply been near the peak of a U-shaped function of benefit for interleaved instruction (where forgetting has happened more rapidly for blocked instruction, and so the difference between the two is maximal at the two week delay).

This work was further conceptually replicated in the ASSISTments adaptive tutoring system (Feng, Heffernan, & Koedinger, 2009) by Ostrow, Heffernan, Heffernan, and Peterson (2015), where student performance on a follow up homework was improved for otherwise low performing students by interleaving practice problems

within a session, as opposed to blocking (students were randomly assigned regardless of ability, and divided into low and high ability post hoc). No difference was found for higher performing students in the system, presumably because of ceiling effects on performance.

Chapter 2

Mathematics Learning and Computer Assisted Instruction

2.1 Mathematics Learning

2.2 Current State of Computer Assisted Instruction for Mathematics

Khan Academy ("The Khan Academy," n.d.) is one of many online Mathematics products that provides a range of computer assisted instruction in Mathematics. Others include Pearson Success Net ("Pearson Success Net," n.d.), the ALEKS Mathematics tutor ("ALEKS – Assessment and Learning, K-12, Higher Education, Automated Tutor, Math," n.d.), which draws on the work on knowledge spaces by Falmagne, Koppen, Villano, Doignon, and Johannesen (1990), and a range of platforms differentiated by school grade from Carnegie Learning ("Home - Carnegie Learning," n.d.). These

products all have one feature in common, the ability to automatically grade student responses to questions, and give immediate feedback about whether they were correct or incorrect. In addition, all the platforms listed above have some measure of adaptivity, whereby students can have an initial diagnostic placement test to determine their current level of knowledge, and also be moved onto different materials depending on the ongoing updated estimate of student knowledge. Some computer assisted instruction software is described as an intelligent tutoring system (VanLehn, 2011), meaning that not only does it give feedback at the level of responses to questions, but it also gives corrective feedback in the incremental steps that students engage in while solving a problem. For example, if a student was solving an algebra equation, the student would have to enter each step of the equation, and would give corrective feedback if the student made mistakes along the way. As described above, VanLehn (2011)'s meta-analysis found improvements in this kind of more fine grained feedback over and above the more simplistic correct/incorrect feedback that Khan Academy takes. However, in spite of this demonstrated effectiveness in approach, the intelligent tutoring systems of Carnegie Learning have not achieved as widespread usage as Khan Academy, which, by 2014, was recording over ten million unique visitors to its website a month (Hirasaki, 2014).

2.3 Primary Data Source: Khan Academy

Khan Academy is a free online resource for learning a range of topics, focused mainly on the US K-12 curriculum. Due to its having ten million unique users monthly, it provides a large dataset of student engagement with interactive exercises with feedback. While many of the exercises are devoted to Mathematics, there are

also many exercises in a range of other domains, many of which are focused more on declarative fact learning - for example, US and World History.

The Khan Academy Mathematics exercises cover a large range, from basic arithmetic (starting at symbolic to non-symbolic mapping), to function graphing and calculus. As such, as well as providing a large dataset, it is very diverse in terms of the kind of learning activities that students are engaging in, and the kinds of Mathematical understanding that the exercises are trying to promote. This means that analysis of the Khan Academy data for spacing effects, gives the opportunity to examine the impact of spacing on declarative fact learning (within primarily non-Mathematical domains), procedural learning, and more abstract categorical and conceptual learning, within the domain of Mathematics.

An initial analysis (Tibbles, 2015) of the data has revealed that a straightforward analysis of spaced vs massed practice across all exercise types in the Khan Academy data set shows a negative correlation with overall performance - i.e. a greater number of study sessions, while controlling for total attempts, has a negative impact on performance in the final session. The confounds in the dataset are due to the mastery model employed by Khan Academy, where students who are initially successful in material will either see it very few times again, or not at all, having been deemed to have mastered due to a short streak of good performance, along with good performance in other areas - in addition the foregoing analysis fails to take into account the temporal duration of the spaced practice, and makes no account for any exercise or user specific effects that may influence performance.

2.3.1 User Stories

In order to better elucidate the nature of the data being analysed, it is helpful to imagine several different users that engage with different Khan Academy exercises (these imagined users are based on observed patterns in the data, and hence will be useful to understand for the foregoing analyses). An individual Khan Academy exercise is composed of multiple assessment items, or questions, which are displayed in varying orders to users as they engage with the exercises.

Spaced Review Mechanics in Khan Academy

Khan Academy currently uses spaced review as part of their mastery model (Faus, 2014) (as depicted in Figure 2.1), with users initially gaining 'practiced' upon completion of a mastery criterion, then after approximately 16 hours, being prompted for mastery review, then upon gaining mastery, being prompted with review cards at longer and longer intervals. In addition, to gain some additional information about mastery, individual questions from exercises with which the user has not engaged will be shown to the user in order to increase the information available for future recommendations.

The current data set spans data collected on Khan Academy from 2012 to 2014, meaning that during some of this period, this spacing mechanism was in place, but not all data will have been produced as a result of it. However, due to the increasing popularity of Khan Academy over time, more data will have been collected in later periods when this mastery system was active. A final caveat is that the above recommendations will only be forced upon the user if they engage with Khan Academy 'missions' - curated paths through sets of content where sets of content are

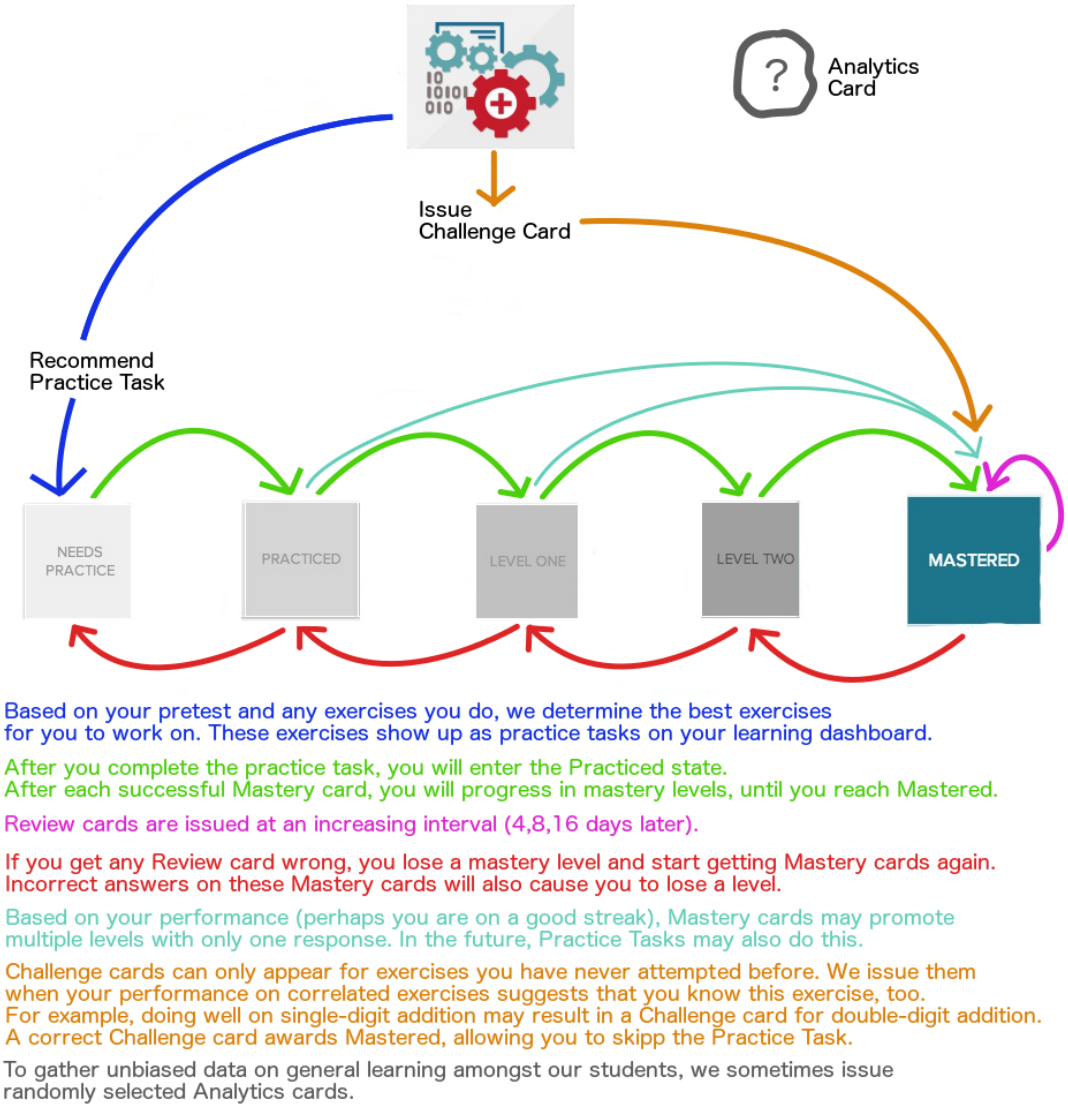


Figure 2.1: Infographic of the Khan Academy Mastery model in 2014 (taken from Faus (2014))

presented in sequence to the learner. Otherwise, users would have to seek out these recommendations on their individual dashboard.

Declarative Fact Learners

While much of the focus of the content on Khan Academy is on Mathematics learning, there are also approximately 4500 questions outside of the Mathematics domain that are multiple choice in nature. Of these questions, not all are explicitly focused on declarative fact learning, as many are comprehension style questions focused on SAT Writing preparation and MCAT preparation, in addition, there are many Science questions that use multiple choice response, but are attempting to assess conceptual understanding, rather than rote factual recall.

In order to focus solely on declarative fact learning, data for declarative fact learning was restricted to multiple choice questions that focused on recall of a single fact. A total of 67 questions were used, each one checked by the researcher to ensure it focused on the recognition of a particular fact among distractors. Additional questions were admissable by the criterion set, but questions with fewer than 1000 responses in the data set were excluded from the analysis, so that sufficient data would be available for analysis on a per question basis. Also excluded was a string of more than 7000 correct responses made by a single user on a single question (by the nature of the response pattern, it appeared to be the result of an automated test procedure run by Khan Academy).

Now we turn to how these data have been gathered - a user would come to the Khan Academy website (or perhaps, already have been engaging with other content - either videos, exercises, or reading articles), and navigate to an exercise (a collection

of related questions) on a particular topic. Therefore, the user will either have come directly to the exercise, from a link in a related piece of content, or because they have been prompted to practice the exercise by Khan Academy's recommendation engine (described above).

The data under examination concern the user's engagement with an individual question. In the case of the declarative fact learning questions, they are all in exercises that contain between six and nine questions. When a user starts an exercise, she is presented with a question from that exercise. The user will answer it and be told that their answer was either correct or incorrect. In the case that they got the answer incorrect, if the user wishes to continue attempting the exercise, they must now choose the correct answer from the available options. In addition, the user can also look at a hint that, in the case of the multiple choice questions, will reveal the correct answer. It is therefore relatively straight forward for a user to perseverate on the exercise, and engage with multiple different questions within the exercise.

When users engage with these exercises they are subject to either a 'five in a row correct' mastery criterion, or a requirement to get all the questions in the exercise correct. As such, in a single session, if a user perseverates to achieve the mastery criterion in the face of incorrectly answering questions, they could confront the same question (with randomized answer order) more than once in a single session of engagement with the exercise. However, if a user gets the answers correct first time round, they are far less likely to view the same question again within that session, unless they continue engaging with the exercise after achieving mastery.

Once a user achieves mastery on the exercise they are prompted to move on to the next piece of content in the sequence that the exercise is in. Students are therefore

likely to return to an exercise only when prompted to do so by the Khan Academy recommendation engine, at an interval of approximately sixteen hours.

Each log entry contains information about the time spent from the loading of the question to the submission of the answer, whether the user's initial submission was correct or incorrect, whether a hint was used, and the time at which the question was loaded by the user. The vast majority of users interact twice with an individual question (see Figure 3.1), indicating that either they fail to achieve mastery first time around, and hence see the question again, or they successfully master the exercise and return to practice it. If the former were true, then performance on the initial engagement should be lower than performance on the subsequent engagement. When testing this hypothesis with a McNemar test (McNemar, 1947) for dependent samples on the performance from first to second conditions, we see that the mean of 0.54332 for the first engagement in a sequence two or more in length is significantly smaller ($p = 1.4822 \times 10^{-323}$) than the mean of 0.80243 for the subsequent attempt. In spite of this significant difference, however, more than 50% of users who do correctly answer the question in their first engagement, then return to answer it again subsequently (this is either due to more practice, or because they fail to achieve mastery on the exercise as a whole). In exploring the data, two groups of users seem to appear:

Same Day Users engaged with a question more than once, but never repeated a question on a subsequent day (assuming US/Central timezone, as most Khan Academy users are based in the US), meaning any sequences of engagement happened, presumably, within a single user session with Khan Academy. These users seemed to engage with questions, and never return to them.

Different Day Users did return to questions on the following day. However,

while they had these longer delays, they also still engaged with repeated questions with shorter delays similar to the short delay users. In general, the long delay users showed slightly higher performance on initial engagement with the questions, meaning that examining performance at longer delays without accounting for the two user types could be slightly confounded by the increased performance of these users.

Mathematics Learners

In contrast to the declarative fact learners detailed above (although, it is possible that the actual users themselves do overlap these two groups), a typical user engaging with Mathematics exercises in Khan Academy is practicing a more complex task than recall or recognition of one particular declarative fact. For example, in the exercise 'Recognize fractions 1', a user is shown a pictoral representation of a pure fraction or mixed number, such as a whole shaded segmented circle, and a partially shaded segmented circle (see Figure 2.2). The user then has to encode that pictoral representation into a pure fraction or mixed number and input the quantity. In other exercises, users will compare fractions, add, subtract, multiply, and divide fractions, building up a suite of rules that must be applied in correct orders to properly achieve the desired result.

Unfortunately, unlike much Intelligent Tutoring Software, Khan Academy does not allow users to input intermediate steps as they solve Mathematics problems, meaning that the only data that is gathered is the correctness or incorrectness of the response - making it harder to view graded levels of performance within an exercise. In addition, within the Mathematics exercises, there are a larger number of questions than in the declarative fact learning. This makes repeats of individual questions less

common (this is especially true in the current data set, where approximately 500 of the Mathematics exercises were algorithmically generated, rather than relying on hand crafted questions).

Further, within the Mathematics exercises, due to the larger amount of data on which to train predictive models, Khan Academy has been able to create more finely tuned mastery criteria for each exercise (Hu, 2011), based on Item Response Theory (Reckase, 2009). This means that, following the recommendation engine described above, users are more likely to engage in shorter runs of engagement, if they are broadly successful in their engagement, and then return to fulfill the next level of their mastery criteria at a later time.

This is reflected in the data by contrast with the declarative learners - there, the data from short duration and long duration learners is comparable in the order of magnitude of their quantity. For a typical Mathematics exercise, however, this does not appear to be the case, with far more data being logged for long duration users than short duration.

In addition, due to the nature of the subject matter, it seems highly likely that the Mathematics exercises are being taken in the context of other instruction in Mathematics - while there may be some independent adult learners brushing up on their pre-algebra, it seems far more likely that the users are engaged in formal education programs, and are using Khan Academy as an additional resource to aid in their learning. So, while the data has the virtue of being spaced, and hence giving the opportunity to investigate forgetting and spacing effects in the real world, the potential for student interaction with other learning materials outside of the Khan Academy exercises means that any forgetting effects may be modulated by engagement

with other materials.

This confound is troubling from a Scientific perspective, as it could potentially hide important results about the nature of forgetting for Mathematics learning - conversely, this kind of real world noise, for practical applications of forgetting and spacing, ideally could be accounted for, as any educational technology tool is very rarely the single source of instruction in any educational context - the information available to it will always be incomplete, therefore.

2.4 Secondary Data Source

In addition to the Khan Academy data, another data source will be used as a replication set to verify results found in the Khan Academy data.

2.4.1 KA Lite

As described in subsection 1.3.1, KA Lite (Alexandre, 2014; Tibbles & Alexandre, 2014) is a platform that allows for the creation of a web like experience, but without the need to connect to the Internet, in perusing and engaging with Khan Academy exercises and videos. As such, it provides an interesting replication set to the Khan Academy data collected over a similar time frame, as the exact same exercises are used in the KA Lite platform, as on the online Khan Academy website. However, the populations served by KA Lite are very different and more geographically diverse than those using Khan Academy, which while having representation in many different countries, are primarily based in the United States. By contrast, many of the users of KA Lite are from India, sub-Saharan Africa, and Latin America. As such, the KA Lite data provides one dimension of generalization for any findings within the Khan

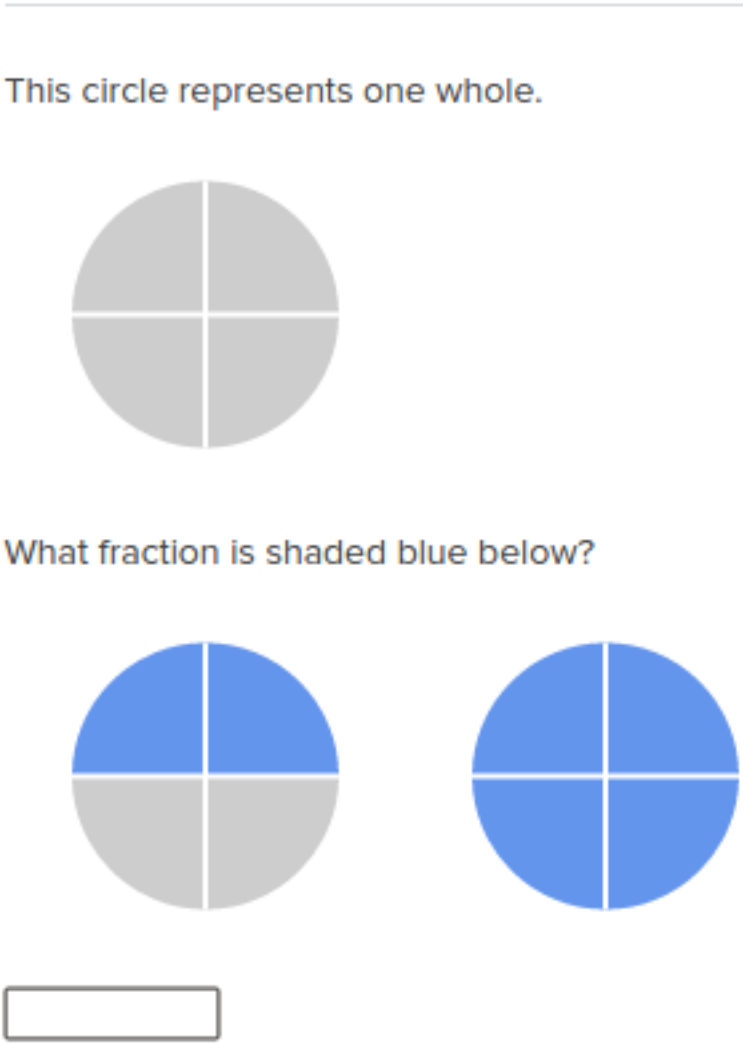


Figure 2.2: One example question from the Khan Academy 'Recognize fractions 1' exercise

Academy data, by generalizing over other populations.

Due to the nature of the KA Lite platform, much of the user data that is collected within the offline installations of KA Lite never makes it back to the central aggregation server from which this data set is derived - if the offline KA Lite instance is never reconnected to the Internet, then no synchronization of the learning data can be accomplished. Data that has been synchronized to the KA Lite central server, therefore, is generally from larger organizations deploying KA Lite, as central aggregation of the user data allows them to more closely monitor and observe how KA Lite is being used across classrooms in their deployments. As such, a significant proportion of the data in the KA Lite data set is from use in more formal schooling settings (in contrast to Khan Academy), with 45% of the Mathematics learning data in the KA Lite data set coming from one regional deployment in India.

Chapter 3

Forgetting in Declarative Learning

3.1 Introduction

As detailed in the foregoing (c.f. subsection 1.5.2, subsection 1.5.1, and subsection 1.5.3) much laboratory and some applied work has been done on the decline of memory performance over time, and the impact of spaced repetition of testing has on improving performance. Much of this work has been done in the context of paired associations (even in applied work, in the work of Lindsey et al. (2014) in teaching foreign language vocabulary) and serial word list recall, however, some work has been done in the realm of declarative fact learning (most notably that of Cepeda et al. (2008) where participants were required to learn obscure declarative facts). As such, there are good reasons for thinking that any applied declarative fact learning would show similar patterns of forgetting, and impacts of spaced repetition. In the Khan Academy data set, in the non-Mathematical content areas, such as History and Science, there are sets of questions focused on declarative fact learning. Unlike much of the testing effect literature (see subsection 1.5.1), however, all of the Khan

Academy declarative fact learning questions are based on multiple choice questions, which does seem to have a beneficial effect (Kang et al., 2007), but does not produce as robust learning gains over time as cued or free recall (Glover, 1989; McDaniel et al., 2007; Halamish & Bjork, 2011), especially when corrective feedback is applied to both test kinds (Kang et al., 2007). In addition to replicating laboratory results, it is particularly interesting to see what other variables seem to have an important contribution to predictive models of student performance, over and above temporal intervals and spaced repetition practice.

Aims

- Do laboratory inferred models of forgetting replicate in real world data sets?
- What are the important predictor variables for forgetting over time in real world data sets?

3.2 Data Selection

To select questions that primarily probed declarative fact learning, a database of the Khan Academy questions was queried in order to generate a set of questions from topics that were outside of the Mathematics domain. Following this, the questions themselves were filtered down only to only those that were multiple choice in form (specifically, coded to use a 'radio' response in the rendered HTML for the question). The Mathematics domain was specifically excluded as many of the multiple choice Mathematics questions on Khan Academy are used as a way to prevent onerous and

complex user input, rather than recall of a previously encountered fact. This yielded 2513 candidate questions.

Due to the lack of data on independent study sessions, only data where a user engaged with the same question more than once was used for analysis, so as to give an initial time point for learning (either a correct response or an incorrect response followed by feedback), and then a further time point for recall. In addition, as there is the possibility that automated processes (including automated testing processes) have been responsible for the creation of some of the Khan Academy data, any single user who recorded more than 941 responses (three standard deviations above the mean) across all these questions (where the mean number of responses per user was 32.85, and the standard deviation is 302.57) was removed from the data set. This subselection process left 128592 users out of the 128760 in the original data set.

After this automated selection process the data were then further subsampled by only using questions with more than one thousand responses. The remaining 619 questions remaining after this subsampling were then checked by hand to ensure that each question was focused on the recall of one specific fact (as opposed to 'all of the above', 'which of the following answers is false', or text comprehension type questions). The text of all sixty seven remaining questions were included after this stage. Many of the excluded questions were comprehension type questions taken from the MCAT preparation section, which resulted in a large reduction in the number of questions, also excluded were SAT Reading, Writing, and Mathematics questions that were either comprehension or not clearly factually recall related. The remaining questions were in the topics of History, Art History, and some questions on magnets and compasses that required simple recall of facts, as opposed to more conceptual Physics questions.

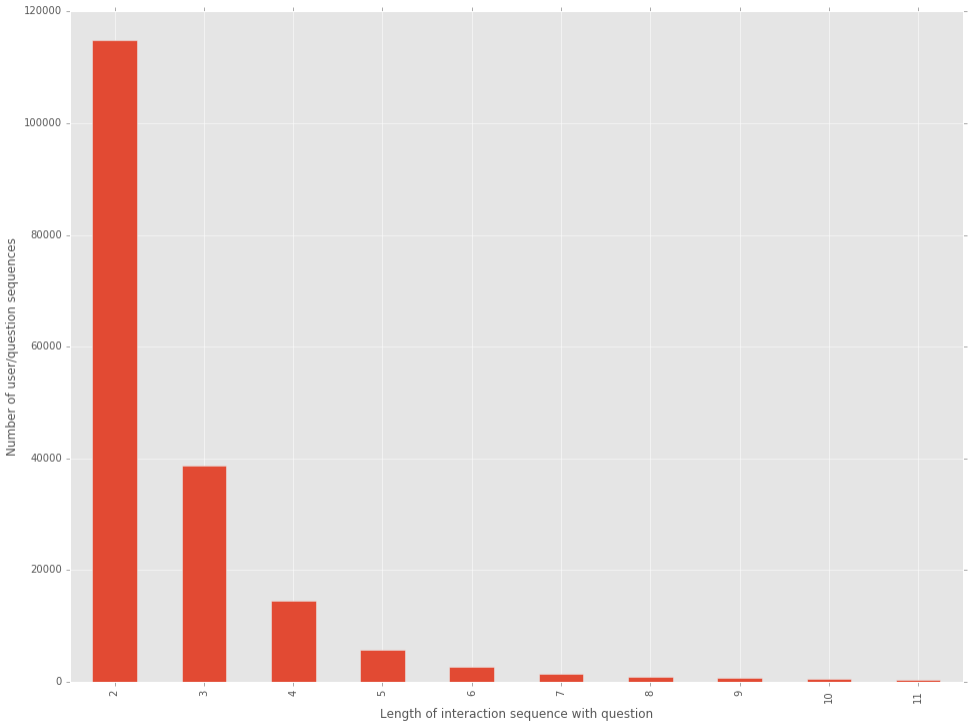


Figure 3.1: Histogram of user/question engagements in declarative fact learning data

3.3 Single Fact Data

The remaining single fact data is primarily composed of once repeated engagements with a question, any single engagements with the question are discarded, as a test-retest sequence is minimally required to establish a duration over which forgetting can occur. In Figure 3.1 the distribution of the number of different sequences can be seen (any sequences longer than the mean plus three standard deviations of sequence length - 11 - were discarded).

For the purposes of initial curve fitting, only the first test-retest sequences for a particular user and question combination were used, so as to avoid the confound

of repeated testing that might add additional variation to subsequent forgetting behaviour. Further, in order to replicate a criterion of 'one correct answer', only questions on which the initial attempt was correct are included. It is possible that this criterion is too stringent, as, when engaging with Khan Academy questions, learners must correctly answer a question in order to proceed to the following question. Further, for declarative multiple choice questions in particular, there are no hints or additional material, so the learner must provide the correct answer without being guided by the software (this is in contrast to the Mathematics exercises, where inline worked example hints are available to guide the student to the correct answer). However, as a learner could simply engage in guessing to produce the correct answer, it seems likely that the learner's memory representation is significantly less robust in the case where the initial response was incorrect.

Examining the distribution of the test-retest pairs (Figure 3.2 and Figure 3.3), it is evident that the vast majority of the test-retest pairs are coming from questions to do with World War I History, and Compass Physics. This distribution of test-retest pairs demonstrates that many of the responses for the Art History topics are due to repeated engagement (this may well be due to the high number of 'check all that apply' type questions that are in the Art History exercises, in addition to the subselected single fact questions - this will cause more student mistakes, and hence the student will have to continue engaging with the exercise and be re-exposed to the single fact questions also). Further, it seems possible that for at least some of the questions it is plausible to do analyses on a per question basis, so as to avoid aggregating across different items.

As can be seen in Figure 3.4, the vast majority of users are only represented

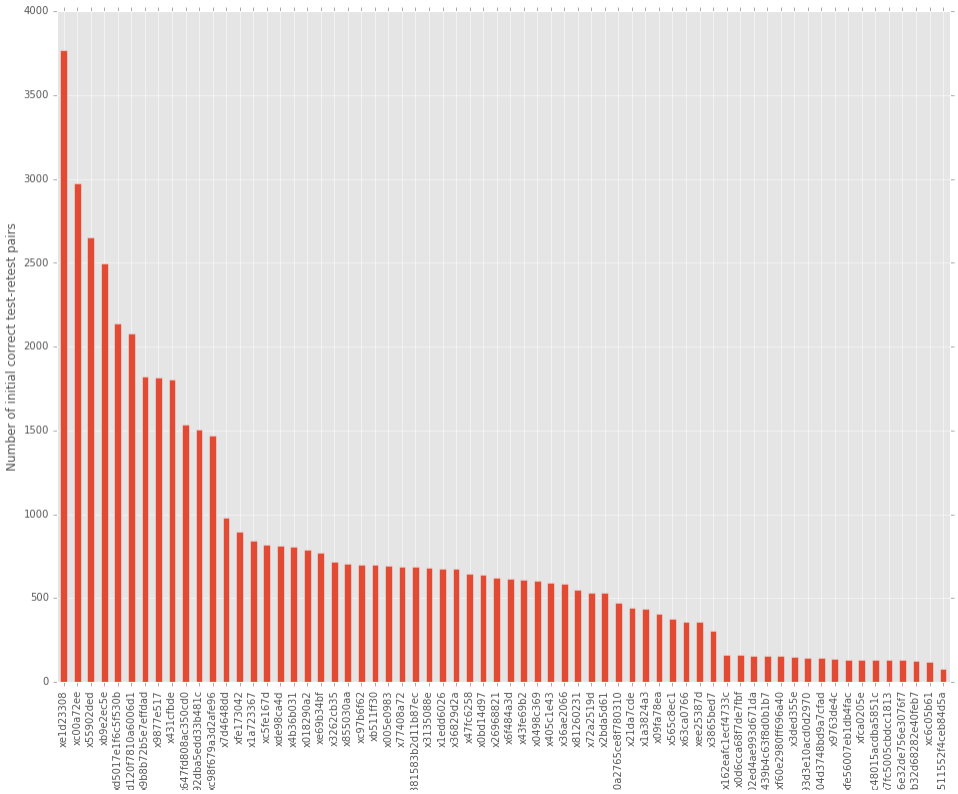


Figure 3.2: Distribution of declarative fact learning test-retest pairs across questions

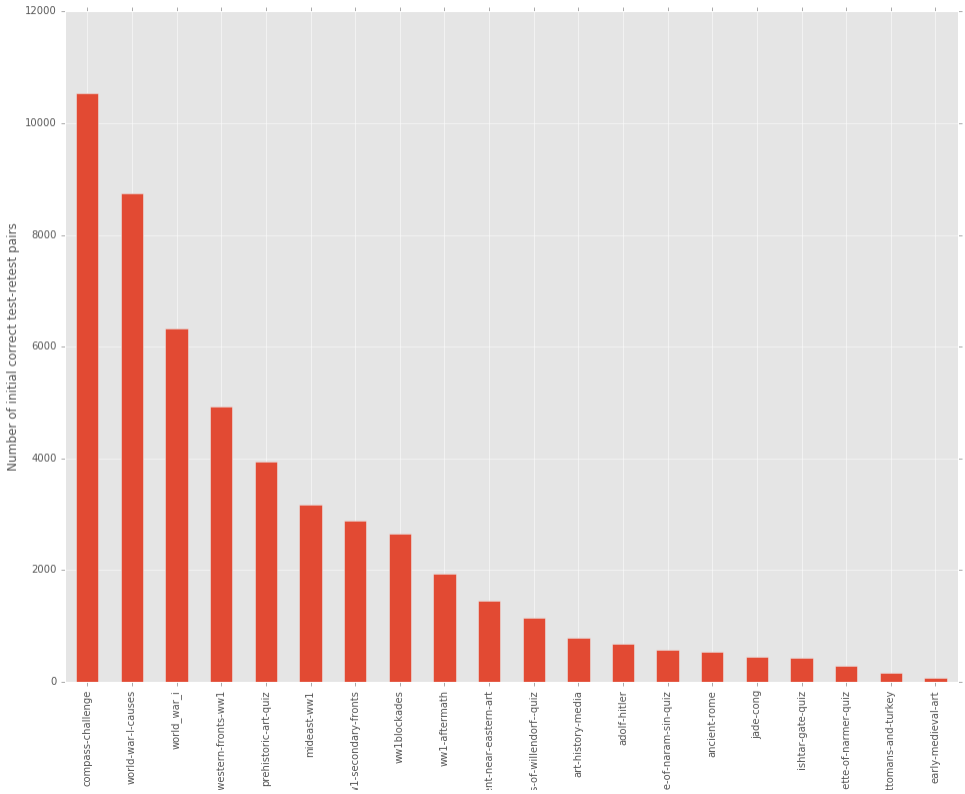


Figure 3.3: Distribution of declarative fact learning test-retest pairs across exercises

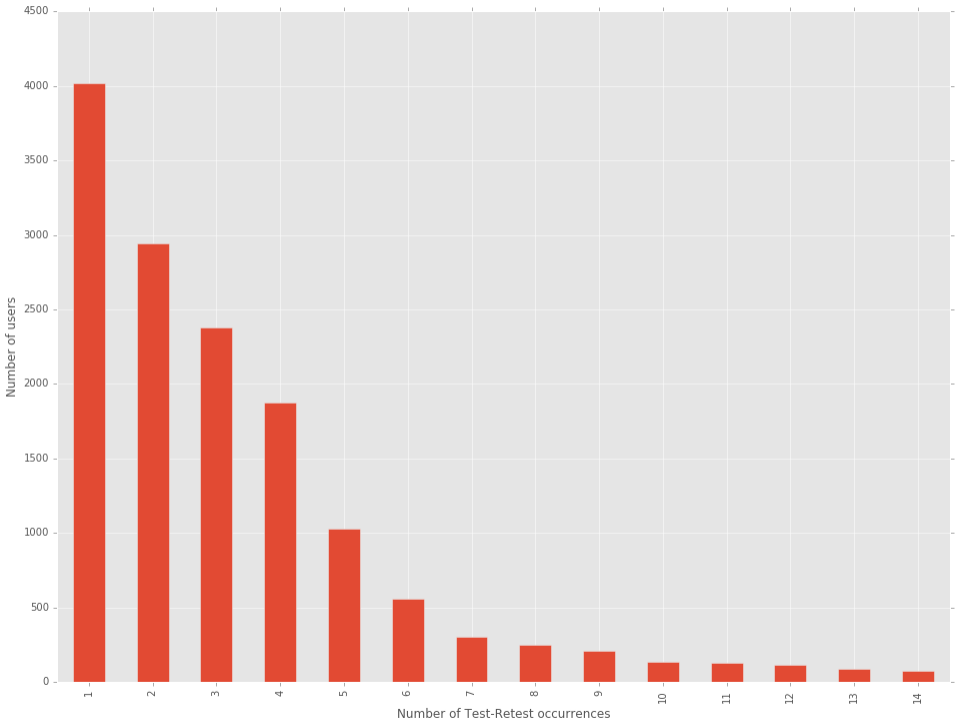


Figure 3.4: Distribution of declarative fact learning test-retest pairs across users

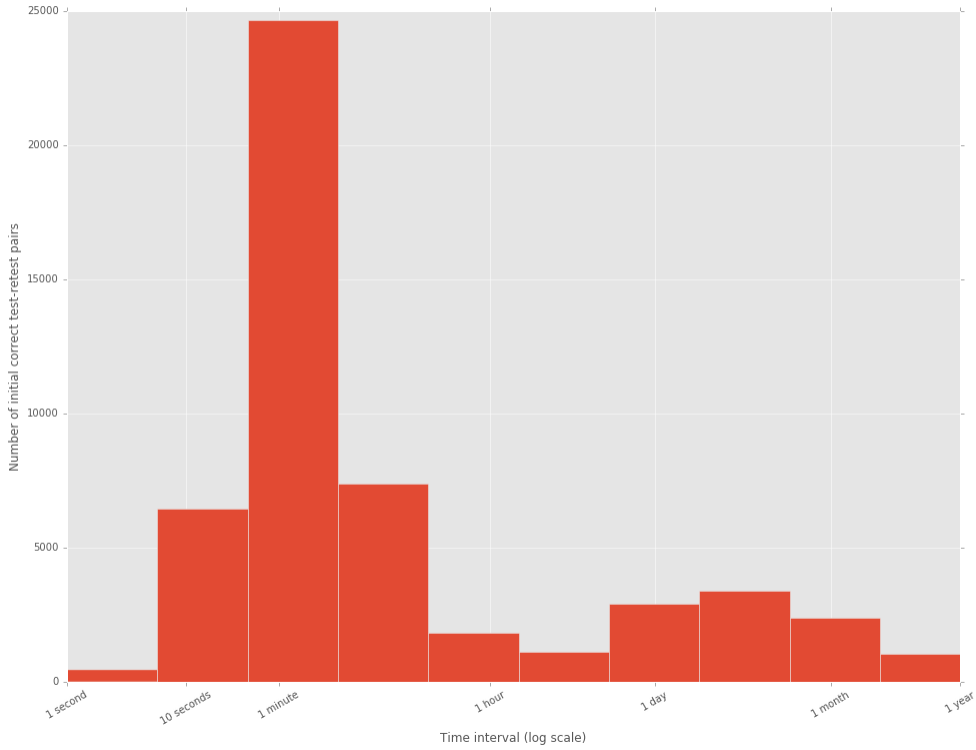


Figure 3.5: Distribution of declarative fact learning test-retest time intervals

once in this sample data set, and very few users are represented more than six times in the data. As such, doing an analysis on a per user basis for the test-retest subsample is impractical, due to the limited number of data points over which aggregation could occur. Finally, an examination of the distribution of time intervals for the test-retest pairs (see Figure 3.5) shows that a large proportion of the test-retest pairs have time-intervals on the order of minutes, with a much smaller proportion of the data at longer time-intervals.

3.4 Retention Curves

As in the broader literature, in order to provide meaningful model fits, data has to be aggregated over multiple collected data points. This is due to the fact that any particular response to a question produces a binary response, either the learner got the question correct or incorrect. However, in order to model how the likelihood of forgetting the data must be aggregated across multiple responses at similar time intervals (in most laboratory experiments sufficient data is collected for a particular user, and the items deemed sufficiently similar that aggregation within a user across items is feasible). In the Khan Academy data, and other real world data sets, users are far less likely to engage with highly similar material, and, as seen above, the vast majority of users do not engage with multiple questions more than once. In aggregating these data, therefore, we are collapsing across users, and potentially dissimilar items (although, as previously discussed, they are all recognition based multiple choice questions). To aggregate around different time points, the data are ordered by time interval and then binned, each bin giving a data point with a mean proportion correct, and a mean time interval. The data plotted in a range of bin sizes from 100 to 1000 are shown in Figure 3.6, all error bars are omitted from the plots, as the standard error is inversely proportional to the mean value (the lower the mean proportion correct, the higher the variance and hence the standard error, as only values of 0 and 1 form the distribution), and inversely proportional to the square root of the bin size. From a first glance at the data, it is again evident that shorter time scale engagements dominate the data set, and further, that the variance of longer time scale performance is significantly higher.

From the literature, it is to be expected that aggregation across multiple users

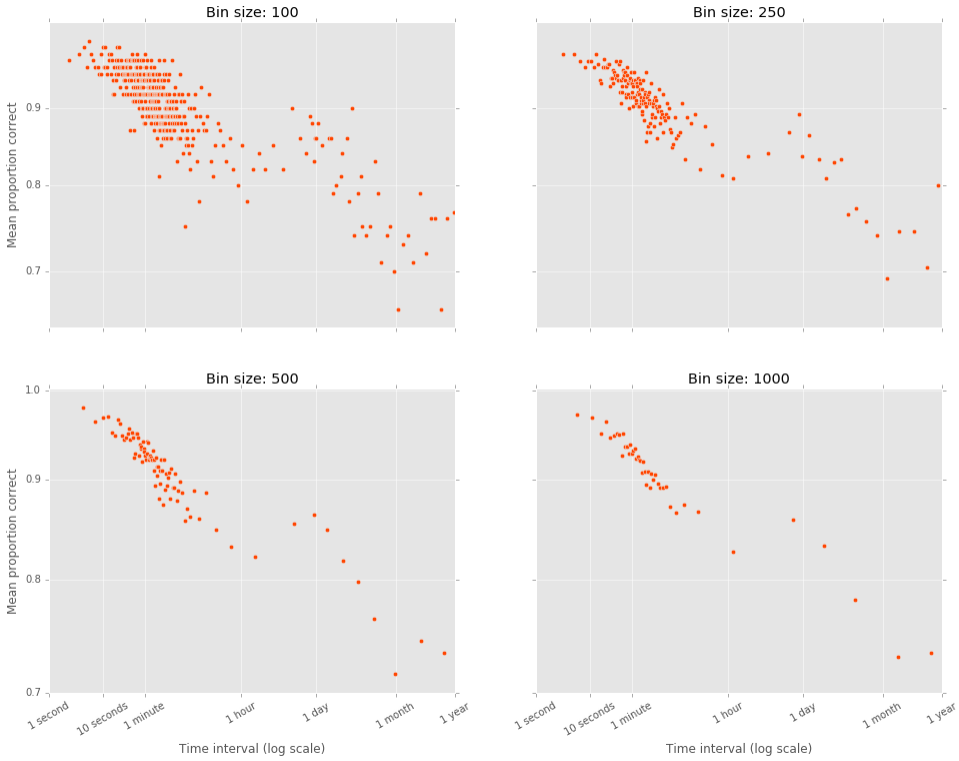


Figure 3.6: Scatter plots of aggregated declarative fact learning data points at bin sizes from 100 to 1000

and data sets will favour a power model fit. This broad favouring is evidenced in Table 3.1, where as the bin size used to aggregate across test-retest data that are clustered around the same time interval, we see a steady increase in the goodness of fit measure for the power law, logarithmic, and the sum of exponentials of Rubin et al. (1999). It is also interesting to note that the particular choice of bin size can have a drastic impact on the other models, for example the two trace model of Chechile (2006) fares comparably to the best fitting models until a bin size of 500, when the number of aggregated data points falls below 100. This seems to potentially be a consequence of the higher number of parameters the function has (four, as opposed to the two parameters of most of the other models), which can also be seen in the unstable performance of the three parameter hyperbolic-power function, the similarly parametered exponential-power function failed to converge during model fitting. Interestingly, the five parameter sum of exponentials function of Rubin et al. (1999) does not suffer from these issues, but may fair better due to the summative, rather than multiplicative contribution of its constituent functions. Performance for logarithmic and power functions is almost indistinguishable across the range of bin sizes, due to the fact that across the range of output variables $[0, 1]$ the functions are almost equivalent, as the power function can be equivalently written as $\log y = \log a - b \log t$, and across the range $[0, 1]$ $\log y \approx y$, which gives $y = \log a - b \log t$, the form of the logarithmic function.

Examination of the exact form of the fits for a bin size of 100 (where every candidate function converges on a somewhat predictive fit), as seen in Figure 3.7, shows that the candidate functions that perform well at this bin size appear to be fitting more to the shorter time interval data (which as noted above has a significantly

Table 3.1: R^2 values for different bin sizes for aggregating pure declarative data across all users and questions

Bin size	10	25	50	100	250	500	1000
Exponential	0.057	0.101	0.121	0.189	0.197	0.288	0.381
Exponential-Square Root	0.115	0.213	0.278	0.381	0.412	0.501	0.570
Hyperbolic	0.063	0.112	0.137	0.208	0.214	0.305	0.396
Hyperbolic-Power	0.211	0.402	0.000	0.710	0.000	0.000	0.000
Hyperbolic-Square Root	0.123	0.228	0.299	0.406	0.440	0.528	0.592
Linear	0.052	0.090	0.108	0.173	0.182	0.273	0.368
Logarithmic	0.207	0.392	0.537	0.694	0.798	0.876	0.915
Power	0.208	0.396	0.543	0.700	0.806	0.883	0.922
Sum of Exponentials-Rubin	0.201	0.384	0.529	0.677	0.785	0.850	0.879
Two Trace Model	0.204	0.387	0.532	0.685	0.790	-3.461	-3.337

higher density of data), whereas the functions that fit worse are fitting more to the longer time interval data, which the higher performing candidate functions are able to mostly ignore due to its sparsity and higher variance. To more quantitatively examine the impact of this differential in data density across time intervals, the data can be progressively subsampled to enrich it for longer time intervals. Arbitrarily, time intervals less than an hour are deemed short, whereas those longer than an hour are considered long. In the unenriched data set, the ratio of short to long time interval data points is 3.54.

The data was subsampled, where short interval data were randomly sampled without replacement, to give a range of short to long interval data ratios from the unenriched data set ratio of 3.5 in increments of 0.5 down to 1. Each candidate function was then fitted to the data subsampled at each of these ratios (using a bin size of 100 to ensure most potential for relative comparisons of fits between functions), the R^2 values for the fits are shown in Table 3.2. The subsampling does not seem to affect the relative performance of the best performing models from the unenriched

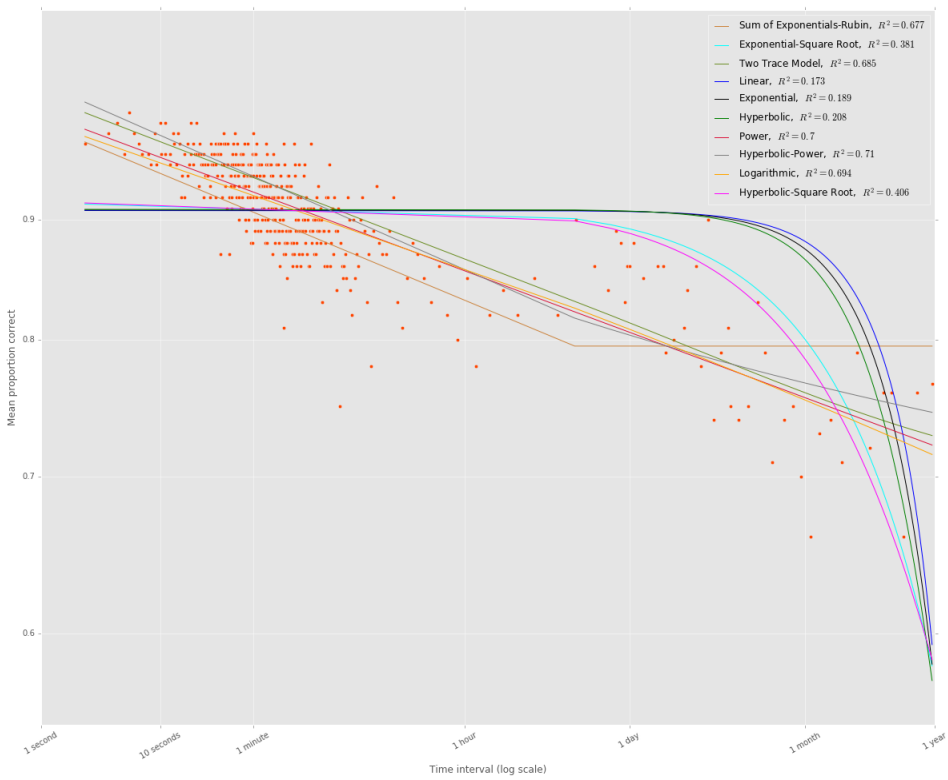


Figure 3.7: Declarative fact learning candidate function fits for a bin size of 100

Table 3.2: R^2 values for different ratios of short to long data, aggregated at a bin size of 100, for pure declarative data across all users and questions

Ratio	3.5	3.0	2.5	2.0	1.5	1.0
Exponential	0.133	0.190	0.189	0.086	0.226	0.205
Exponential-Square Root	0.327	0.389	0.400	0.305	0.463	0.447
Hyperbolic	0.150	0.210	0.208	0.099	0.248	0.225
Hyperbolic-Power	0.689	0.705	0.000	0.000	0.000	0.000
Hyperbolic-Square Root	0.354	0.415	0.428	0.337	0.494	0.479
Linear	0.119	0.173	0.172	0.075	0.206	0.188
Logarithmic	0.668	0.691	0.715	0.678	0.752	0.742
Power	0.675	0.697	0.720	0.685	0.754	0.744
Sum of Exponentials-Rubin	0.662	0.670	0.691	0.667	0.812	0.810
Two Trace Model	0.664	0.686	0.715	0.683	0.751	0.000

data set, with logarithmic, power, sum of exponentials, and the two trace model still performing comparably at most ratios. As with the varying bin size in the unenriched data set, the two trace model seemed more sensitive to changes in the size of the data set and failed to produce a good fit when the data set was at its smallest at a ratio of 1. As an example of the resultant fits, Figure 3.8 shows the fits at a ratio of 1.5 (or $3:2$). When compared with Figure 3.7, it is evident that only the sum of exponentials function of Rubin et al. (1999) has changed its model fit significantly as a result of the subsampling, with far more sensitivity evident towards the more variant data at longer intervals.

This leads to the question, which has been underexplored in the literature, of whether short and long interval forgetting should even be expected to follow the same pattern of forgetting over time, or whether the presumably complex environmental and neurobiological interactions that presumably underpin forgetting, might present a more complex pattern that may change over time? In order to consider this in more detail, rather than simply enriching the overall data set, it is helpful to consider the

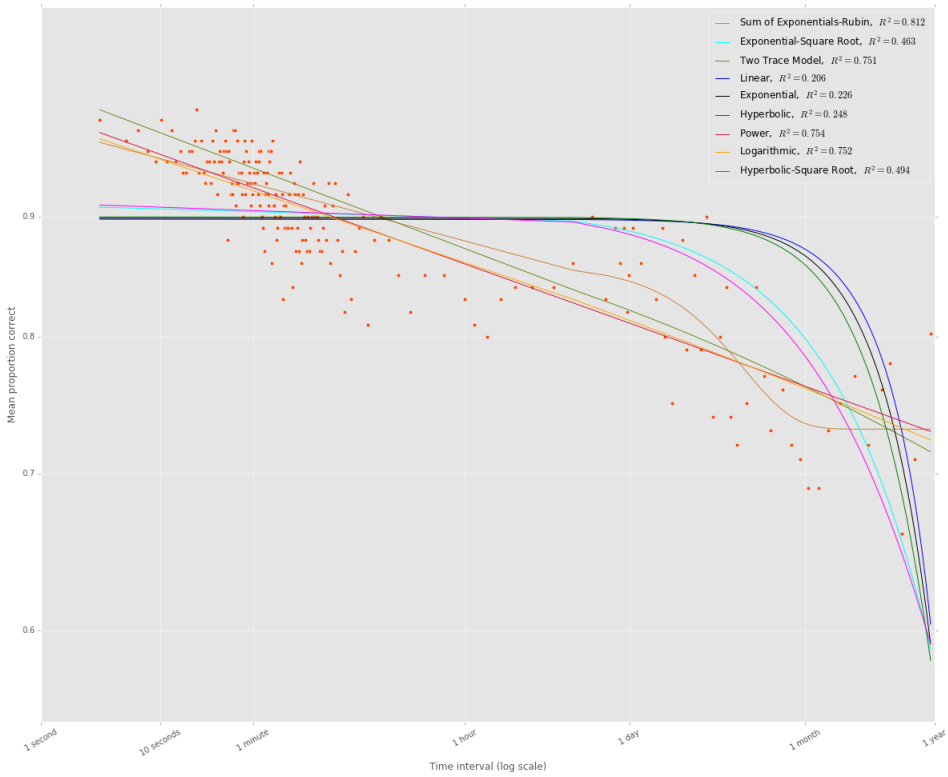


Figure 3.8: Declarative fact learning candidate function fits for a bin size of 100, subsampled to give a ratio of 3:2 short:long interval test-retest data points

long and short interval sequences separately.

As noted in section 1.5.2, the power law, by its nature, will tend to fit well to an aggregation of exponential functions. By the simultaneous aggregation across users and questions, it seems likely that the power law is being overly favoured compared to other candidate functions. In order to partially address this concern, fits can be made on a per question basis, to see if this reduces the favourability given to the power law by the calculated R^2 values. Questions with more than 1000 test-retest pairs were included in the analysis, in order to provide at least ten data points for fitting at a bin size of 100. The results are shown in Table A.1 and are summarized in Figure 3.9. The results continue to show reliable performance from a power law fit (potentially because data is still being aggregated across multiple users) and for a logarithmic fit, even when model performance is scaled by the best fitting model for a particular question (which should favour functions that fit particularly well for a smaller number of questions - see Figure A.1). Performance for the sum of exponentials model is occasionally very high, but generally performance is surprisingly unreliable. The two trace model suffers due to the smaller data sets, and fails to provide good fits for most questions. Plots for fits that produced an R^2 in excess of 0.5 are displayed in Figure 3.10. As noted above the roughly equivalent performance of the power and logarithmic fits is unsurprising, given the characteristics of the logarithm function across the $[0, 1]$ range.

3.5 Interference Effects

It is to be expected that interference effects should have an effect on the retention function as interference accumulates. While interference is highly correlated

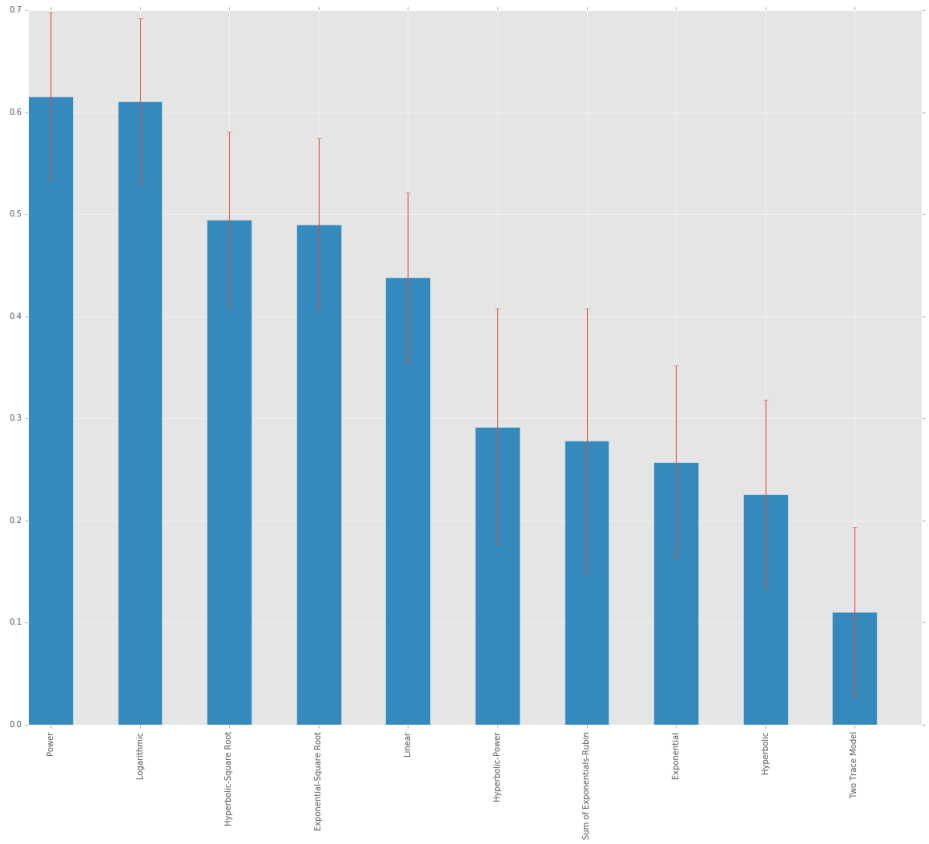


Figure 3.9: Declarative fact learning candidate function mean R^2, on a per question basis.

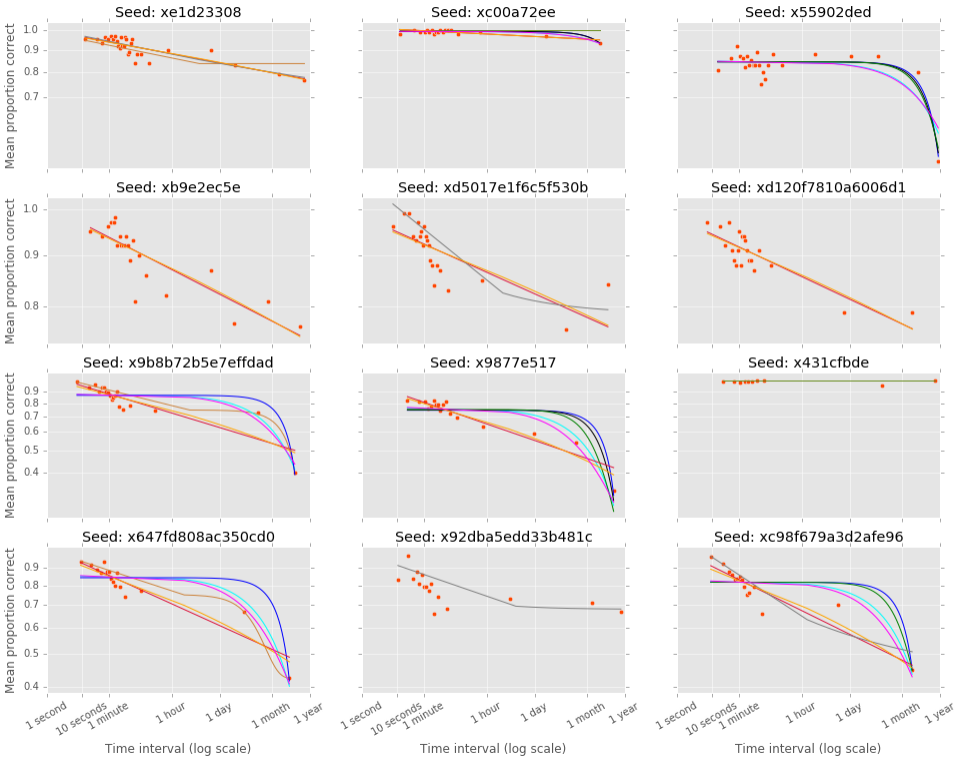


Figure 3.10: Declarative fact learning candidate function fits for a bin size of 100, on a per question basis for fits with R^2 in excess of 0.5.

Table 3.3: R^2 values for different bin sizes for aggregating pure declarative data across all users and questions, fitted only against a holdout data set using interference measures as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.112628	0.297288	0.450050	0.670366	0.830241	0.874063	0.938081
Ridge regression test score	0.104502	0.312417	0.456159	0.661868	0.842560	0.890357	0.953952
Ridge regression training score	0.108264	0.297333	0.446311	0.647943	0.830191	0.868137	0.941248
exercise_interference	-0.023719	-0.027995	-0.027577	-0.024027	-0.027780	-0.026991	-0.028171
other_interference	-0.011786	-0.022666	-0.021470	-0.023923	-0.031159	-0.031751	-0.035403

with the passing of time, it is possible it could subsume the effect of time passing, as well as provide additional information about retention. Further, from the literature, it seems that any interference between a test-retest pair should be problematic for later retention, as the consensus seems to be that retroactive interference is non-specific in nature.

Using the logarithm of time as a regressor variable (due to the high performance of the power and logarithm fits in the foregoing), a multivariable retention model can be constructed with interference measures constructed from intervening answers to questions for the user. Due to the size of the overall data set, only interference effects from other multiple choice questions are considered due to the difficulties of calculating interference effects across the whole data set for a user.

Using only interference effects as regressor variables (both within an exercise, and across all exercises), it is possible to produce comparable model performance to a purely time-based fit. The results of both an unregularized ordinary least squares fit, and one using ridge regression are shown in Table 3.3, both were trained on 80% of the available data, and the regression test score was calculated on the hold out set.

This result is unsurprising, as the accrual of interference is highly correlated with the passing of time, however examination of the correlation matrix for the data show a very small correlation -0.0374 between time interval and interference from

Table 3.4: R^2 values for different bin sizes for aggregating pure declarative data across all users and questions, fitted against a holdout data set using log-time and interference measures as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.221018	0.413086	0.601506	0.759940	0.870159	0.915474	0.947155
Ridge regression test score	0.181942	0.382333	0.527917	0.713035	0.881972	0.918165	0.967459
Ridge regression training score	0.221692	0.414312	0.602927	0.762008	0.871500	0.914508	0.938656
exercise_interference	-0.009500	-0.010548	-0.010843	-0.012771	-0.017427	-0.017473	-0.017011
other_interference	0.005929	0.002346	0.001502	-0.002811	-0.013924	-0.015682	-0.016786
time_interval	-0.046184	-0.043713	-0.043124	-0.038616	-0.025238	-0.024602	-0.019925

Table 3.5: R^2 values for different bin sizes for aggregating pure declarative data across all users and questions, fitted against a holdout data set using only log-time as a regressor variable.

Bin size	10	25	50	100	250	500	1000
R Squared	0.211635	0.398859	0.581067	0.732870	0.830882	0.876994	0.898951
Ridge regression test score	0.173063	0.348130	0.489759	0.660098	0.816130	0.836124	0.846368
Ridge regression training score	0.211861	0.399292	0.581670	0.733636	0.832065	0.878775	0.901916
time_interval	-0.047908	-0.048253	-0.048346	-0.048307	-0.048794	-0.051473	-0.050390

the same exercise, whereas the correlation for interference from other exercises is 0.445. When the time interval is added back in as a regressor (see Table 3.4), we see a very minor improvement in performance at large bin sizes, but more pronounced improvements at coarser aggregations.

Finally, we can compare this with similar models fitted for time alone and see that combining the regressors results in a modest improvement in performance (see Table 3.5).

3.5.1 Summary

In summary, it seems that the current data set is amenable to replication of the literature on replication curves, and that the data also seem to accord with mechanistic explanations on the role of interference in forgetting over time. Given the predomination of the shorter time scales within the declarative fact learning data set,

the correlation of interference with time is unsurprising, as much of the data is accrued within timescales of less than an hour - which it would seem is not unreasonable for a learner to be engaging consistently with exercises across that timespan. However, in spite of strong regularization, using both time and interference together as regressors gives better performance than either of the two alone.

Chapter 4

Forgetting in Mathematics Learning

4.1 Introduction

It has been assumed that the forgetting of Mathematical knowledge will behave similarly to the laboratory tested paradigms outlined above (for example, Lindsey et al. (2014) suggest that their spaced scheduling system is agnostic to content type, and hence could work equally well for any identified 'knowledge component'). However, while this is a reasonable working assumption in the absence of evidence, it is important to verify and quantify the extent to which retention in Mathematics learning resembles retention in other kinds of learning. As noted by Rohrer et al. (2014), learning in Mathematics can be readily conceptualized as a two step procedure, firstly the recognition of the problem type, and hence the correct procedure to apply, and secondly the correct recall and performance of that procedure in the context of this problem instance. The implication here is that while the recognition part of this

process may be very similar to the kind of recognition occurring in the declarative fact learning exercises, the correct recall and application of the procedure may have different characteristics. Further, it is unclear how generalization from applying a procedure to multiple individual instances of a problem to being able to apply the procedure to novel, hitherto unencountered problems, is achieved. Indeed, in early arithmetic, work by Rickard et al. (2008) seems to suggest that performance of arithmetic problems tends to result in memorization of results (where if a calculation is performed frequently enough the result will be stored for future use, rather than repeatedly recalculated) - so it is unclear in Mathematics if fluency arises from a generalization of the calculation procedure to multiple instances, or the application of the generalized procedure to multiple instances a sufficient number of times to produce memorization of results. It is also possible that both of these processes occur as a learner masters a Mathematical procedure (either serially, or in parallel).

Aims

- Does forgetting in Mathematics occur in a similar way to declarative fact learning in the Khan Academy data set?
- Is learning of individual question and answer associations directly analogous to declarative fact learning, in terms of retention?
- Does the learning of recognition of the correct procedure to apply, and application of that procedure (i.e. generalization of the procedure to multiple instances) exhibit different characteristic forgetting than declarative fact learning?

4.2 Mathematics Exercises in Khan Academy

The Mathematics exercises available on Khan Academy are directed at covering the whole of the US Common Core Curriculum standards ("Common Core State Standards Initiative — The Standards — Mathematics," n.d.) from Kindergarden to 12th Grade. As such, they cover a comprehensive range of Mathematics topics from counting all the way up to calculus. In addition, there is a wealth of college level Mathematics material available. An example of one such question from the Mathematics exercises can be seen in Figure 2.2. The vast majority of Mathematics exercises on Khan Academy require some sort of free response from the student (as opposed to the predominantly recognition based multiple choice questions used in the declarative fact learning questions), with either a numeric or symbolic input being required. Some Mathematics questions do use recognition based multiple choice, but even among the multiple choice questions, the correct answer frequently requires the application of conceptual knowledge (such as 'Which of these three triangles are isosceles?', for example).

Within the current data set, there is a mixture of data from exercises constructed in different ways. One kind of exercise (the first kind deployed on the Khan Academy site) are procedurally generated exercises, where each exercise is templated and then the values filled in by use of a seed value. In this way, exercises are procedurally generated, but repeatably so, if the same seed is applied. Unfortunately, it is not precisely apparent which seed values are unique, and which produce an equivalent exercise to the other, so doing item level analyses on these exercise types is challenging, as it is not clear whether two interactions have been with the same exercise (as the seed is an incrementing integer, there is no repetition of these seeds within a user, and hence

no repeated interactions). The second kind of exercise uses hand crafted questions (since the collection of this data set, these have completely replaced the procedurally generated questions on Khan Academy), which each have a unique identifier, and as such, repeated attempts on these particular questions results in data that can be compared in a similar way to the declarative fact learning data.

4.3 Data Selection

Data Density

Most of the accrued data on the Khan Academy platform falls in the pre-algebra to algebra and geometry range, roughly aligning with Mathematics instruction that usually occurs in upper elementary through early high school in the US education system.

Temporal Distribution

As with the declarative fact learning data, many of the time intervals between subsequent engagements with Mathematics questions and exercises are less than an hour, but there is also a longer tail of the distribution that means there is sufficient data to examine retention over time intervals from months to up to a year.

STEM Pipeline

From a pragmatic point of view examining the prealgebra section of the curricu-

lum is of particular interest, as it is one area where students frequently begin to struggle, with the lack of fluency in pre-algebra resulting in a difficulty transitioning to algebraic Mathematics in late middle school and early high school ([ )p.8]council˙adding˙2001. Further, in the secondary, KA Lite dataset, the exercises covering these topics also have the highest data density, and hence will serve as a useful comparator for the Khan Academy data in a different population.

4.4 Item Level Analysis

In order to understand the relationship between retention functions for learning Mathematics and declarative fact learning it is instructive to initially carry out parallel analyses to the declarative fact learning analyses for individual items within Mathematics exercises, and then to compare to the same analyses carried out at a higher level of granularity, with time intervals taken between repeated attempts at the same exercise (rather than repeated attempts at the same question within an exercise).

4.4.1 Item Level Data for Adding and Subtracting Fractions with Like Denominators Word Problems

Only thirteen of the forty nine exercises within the pre-algebra topic are amenable to item level analysis, although the responses to these questions make up nearly a quarter of the 200 million responses in the data set. Analyses for one of these problems, with approximately 4.5 million responses are described here.

In spite of the use of individually crafted questions, in parallel to the declar-

ative fact learning exercises described above, the length of interaction sequences is considerably shorter for Mathematics exercises. This is due to the differing mastery criteria used, whereby users are not necessarily expected to get every question correct in an unbroken streak (unlike in the declarative fact learning exercises) and the higher number of questions per exercise, meaning that even if a learner engaged with an exercise for twenty attempts, far fewer of those attempts would involve the learner attempting questions that they had previously engaged with - whereas in a declarative fact learnign exercise, this would almost certainly have involved multiple repeats of the same question. The distribution of these sequences is summarized in Figure 4.1 - as with the declarative fact learning data, the single item occurrences have been excluded (although, to an even greater extent they make up the majority of the data).

In contrast to the broader sweep of the declarative fact learning data, the distribution of test-retest pairs (see Figure 4.2) is far more evenly spread across different questions, with a noticeable drop off for a subset of questions that, by inspection of the current Khan Academy questions being used in this exercise, it seems were removed from use at some point during the data collection period.

Similarly to the declarative fact learning questions, the number of test-retest sequences for a particular user on an item is also very low (see Figure 4.3), meaning that aggregation by user, as opposed to by item, would be difficult within this particular data set.

Finally, an examination of the distribution of time intervals for the test-retest pairs (see Figure 4.4) shows that while a large proportion of the test-retest pairs have time intervals on the order of minutes, in contrast to the declarative fact learning data the distribution of time intervals is nearly bimodal, with a high number occurring at

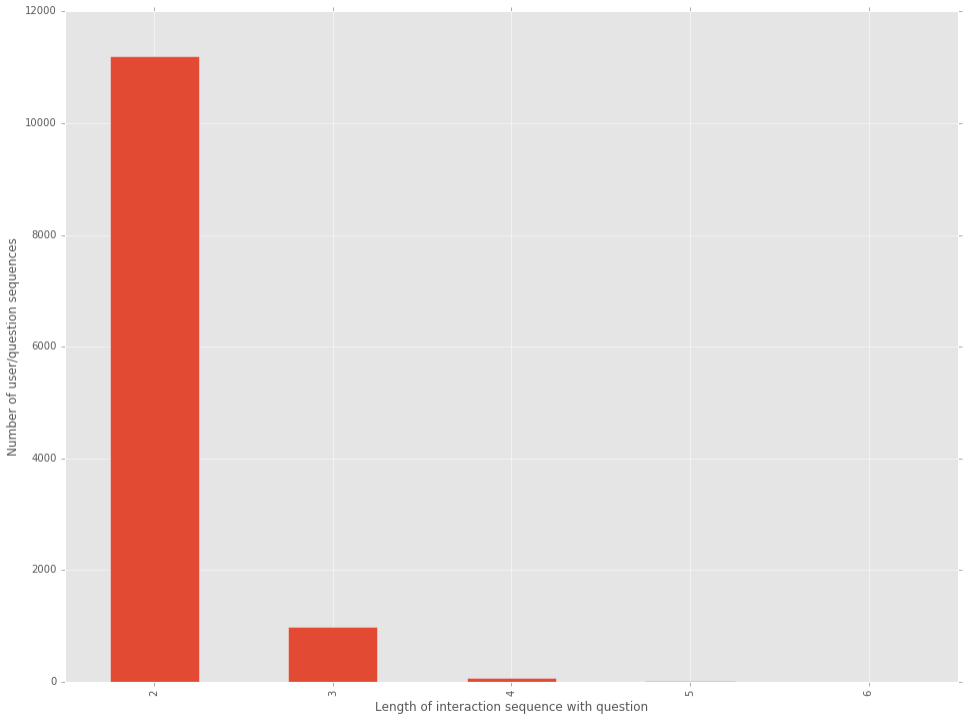


Figure 4.1: Histogram of user/question engagements in Adding and Subtracting Fractions with Like Denominators Word Problems data.

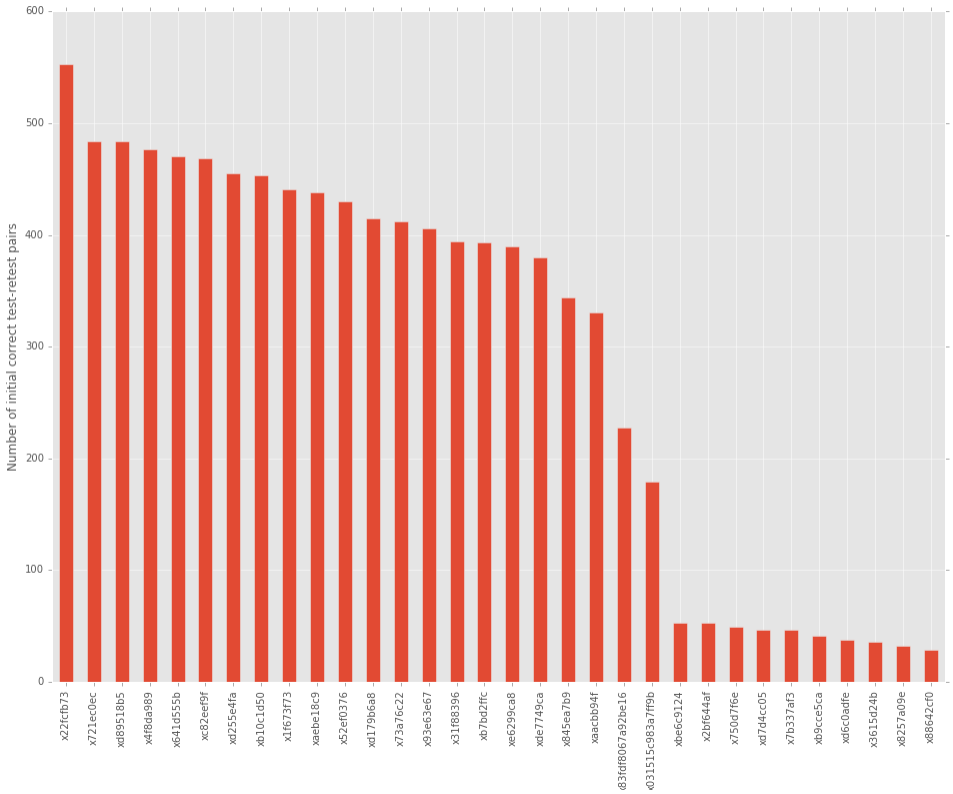


Figure 4.2: Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems test-retest pairs across questions

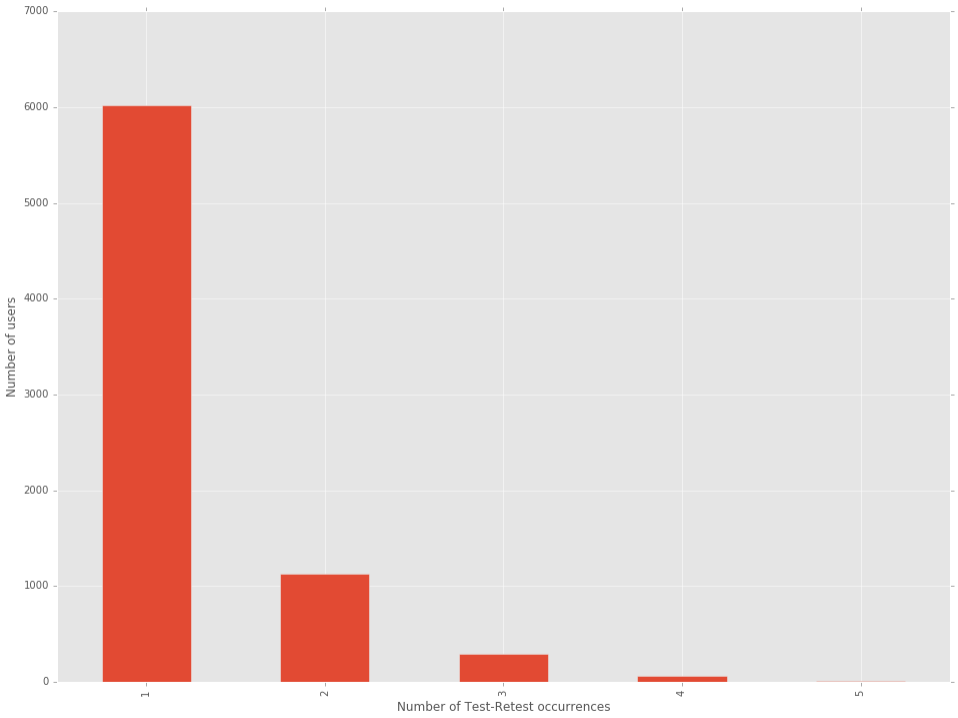


Figure 4.3: Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems test-retest pairs across users

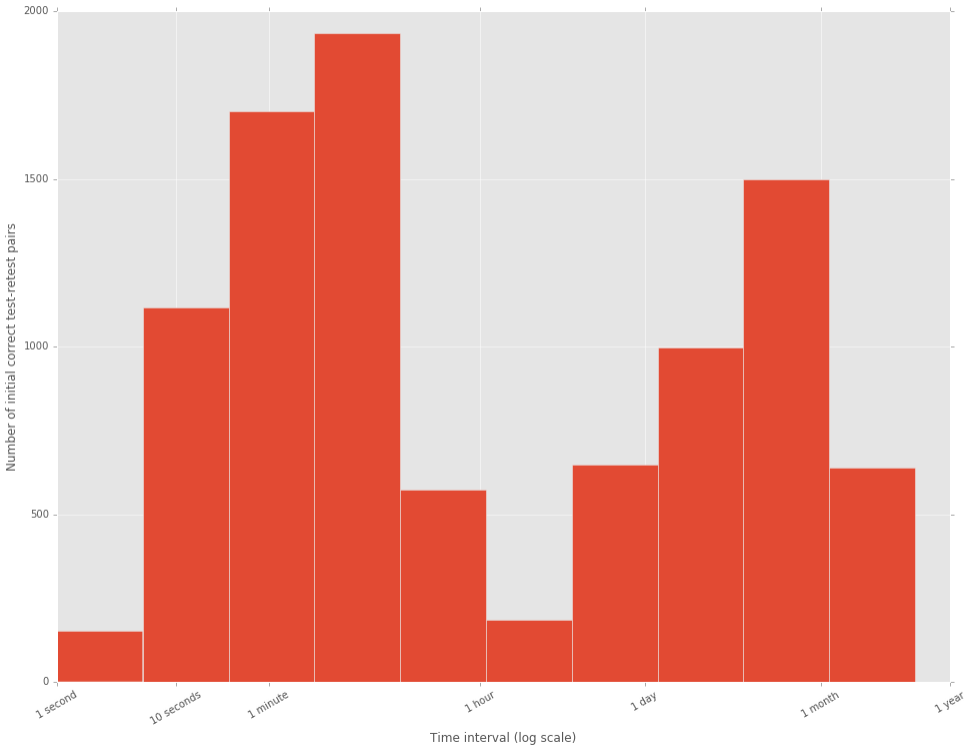


Figure 4.4: Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems test-retest time intervals

significant delays, on the order of many days. While the ratio of test-retest pairs of short intervals to long intervals for the declarative fact learning data set was 3.54, for the current data it is 1.37.

4.4.2 Retention Curves

There are two possibilities for an analysis of the retention curves at an item level for Mathematic exercises - either they are amenable to similar fits to the declarative fact learning, in which case it seems likely that individual items are learned in a way similar to individual items on the declarative fact learning questions, or a markedly

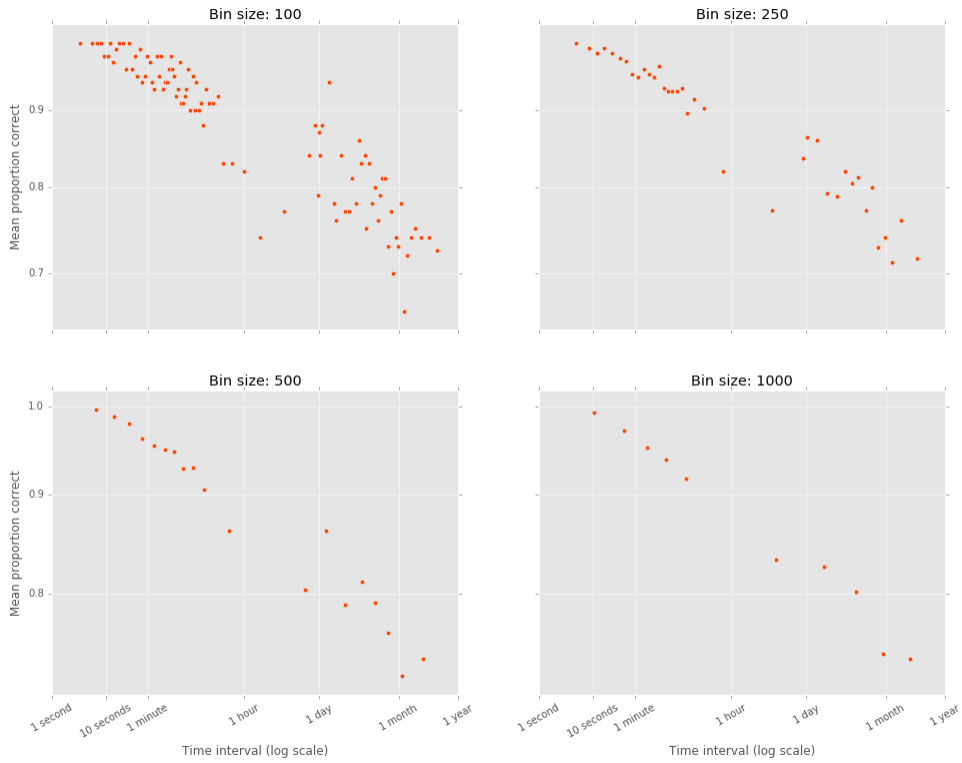


Figure 4.5: Scatter plots of aggregated Adding and Subtracting Fractions with Like Denominators Word Problems data points at bin sizes from 100 to 1000

different pattern is apparent in the individual item data, suggesting that the underlying learning that is happening is at the exercise or procedural level, rather than at the level of the individual items.

If the level of learning is procedural, rather than individual item based, then we should expect to see a positive influence of interference from the same exercise on future engagements with an individual question - which is the converse of the effect seen in the declarative fact learning analyses, where in the purely interference and the interference and time based models we saw a negative contribution to performance from interference.

Table 4.1: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions

Bin size	10	25	50	100	250	500	1000
Exponential	0.203	0.355	0.474	0.345	0.408	0.428	0.477
Exponential-Square Root	0.321	0.506	0.597	0.615	0.678	0.698	0.719
Hyperbolic	0.221	0.377	0.487	0.385	0.444	0.458	0.498
Hyperbolic-Power	NaN	0.000	0.656	NaN	NaN	NaN	NaN
Hyperbolic-Square Root	0.332	0.518	0.602	0.644	0.707	0.727	0.746
Linear	0.185	0.335	0.465	0.309	0.374	0.401	0.458
Logarithmic	0.416	0.609	0.650	0.851	0.922	0.958	0.986
Power	0.415	0.607	0.644	0.851	0.921	0.958	0.984
Sum of Exponentials-Rubin	0.394	0.572	0.593	0.821	0.883	0.899	0.902
Two Trace Model	0.414	NaN	NaN	-1.715	-1.841	-1.919	-2.082

In the scatter plots (Figure 4.5) of performance in test-retest trials, we see a similar pattern over time to the declarative fact learning, although, even at smaller bin sizes for aggregation, we see less variance in performance at higher time lags, presumably because of aggregation over a more homogeneous item set, as all questions are word problems for the same mathematical procedure. Making fits to candidate retention functions (Table 4.1), similarly to declarative fact learning, we see increasingly improving fits across bin sizes for aggregation, with the best model performance seeming to come from the logarithmic and power law fits again.

Examining sample fits for a bin size of 100 (Figure 4.6), in spite of the more equal ratio of short to long interval data points available, it appears that the shape of the best fitting functions is still largely constrained by the shorter interval, lower variance time points. Fits to individual questions also show good average performance for the power and logarithmic retention functions (Figure 4.7), implying again that their performance is not due to aggregation over particular items, but possibly still due to aggregation over individual user characteristics. Fits with an R^2 in excess of

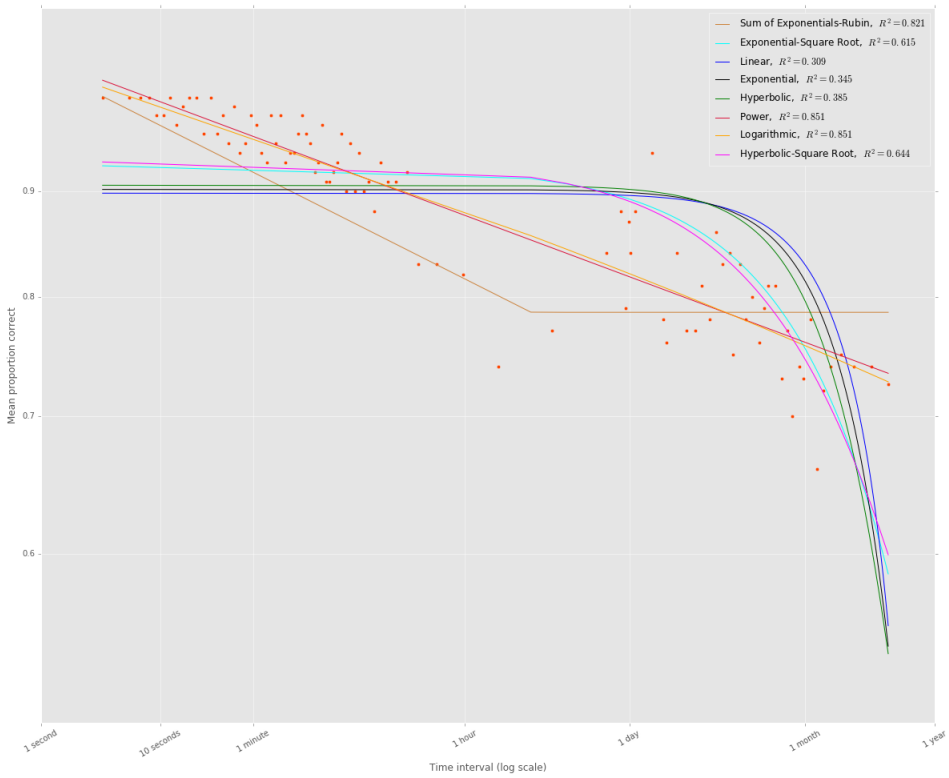


Figure 4.6: Adding and Subtracting Fractions with Like Denominators Word Problems candidate function fits for a bin size of 100

0.5 are plotted in Figure 4.8, with the majority of the fits being logarithmic and power function fits.

4.4.3 Interference Effects

As suggested previously, if learning happens more on the procedural level for Mathematics exercises, it seems likely the previously observed interference effects from the same exercise should reverse in direction. When examining a model that includes interference from the same exercise, however, it seems that this reversal does not occur, at least in the subset of the data for test-retest trials (Table 4.2).

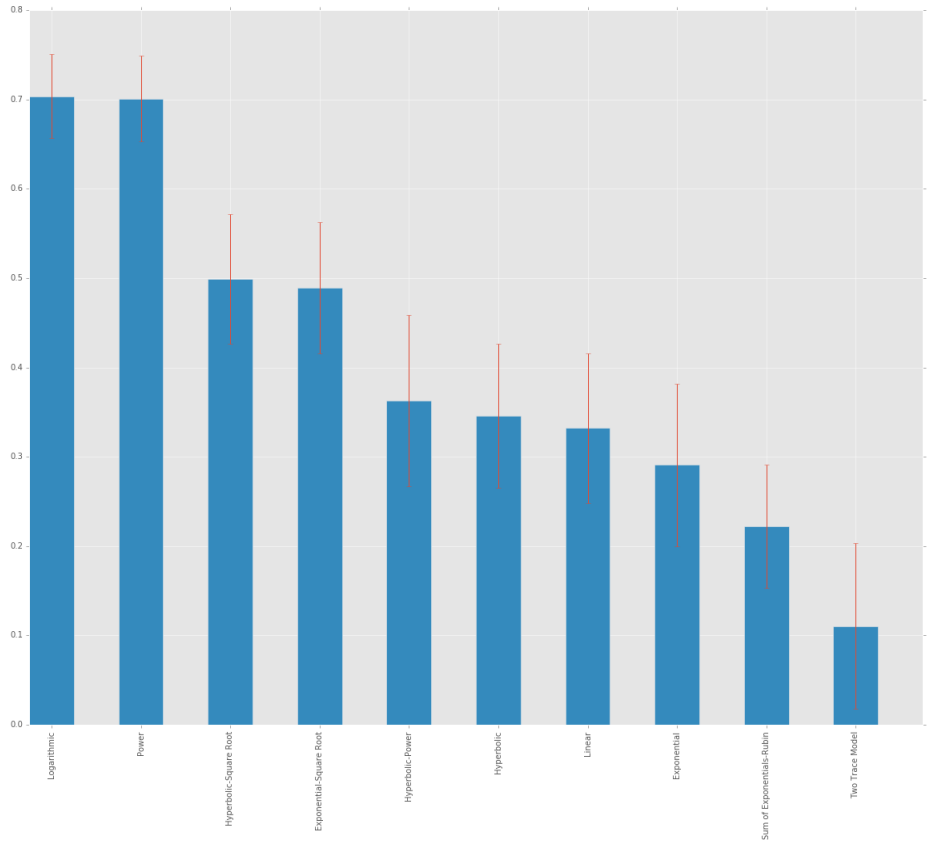


Figure 4.7: Adding and Subtracting Fractions with Like Denominators Word Problems candidate function mean R^2, on a per question basis.

Table 4.2: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set using log-time and exercise interference as a regressor variable.

Bin size	10	25	50	100	250	500	1000
R Squared	0.248671	0.410405	0.530027	0.607699	0.671858	0.738196	0.826468
Ridge regression test score	0.229176	0.357353	0.502330	0.641442	0.700757	0.618459	0.887817
Ridge regression training score	0.250634	0.414063	0.534925	0.612332	0.656939	0.658911	0.869851
exercise_interference	-0.036468	-0.049359	-0.054135	-0.053597	-0.049957	-0.041850	-0.057081
seed_time_interval	-0.044253	-0.029738	-0.024927	-0.021151	-0.015500	-0.014921	-0.030749

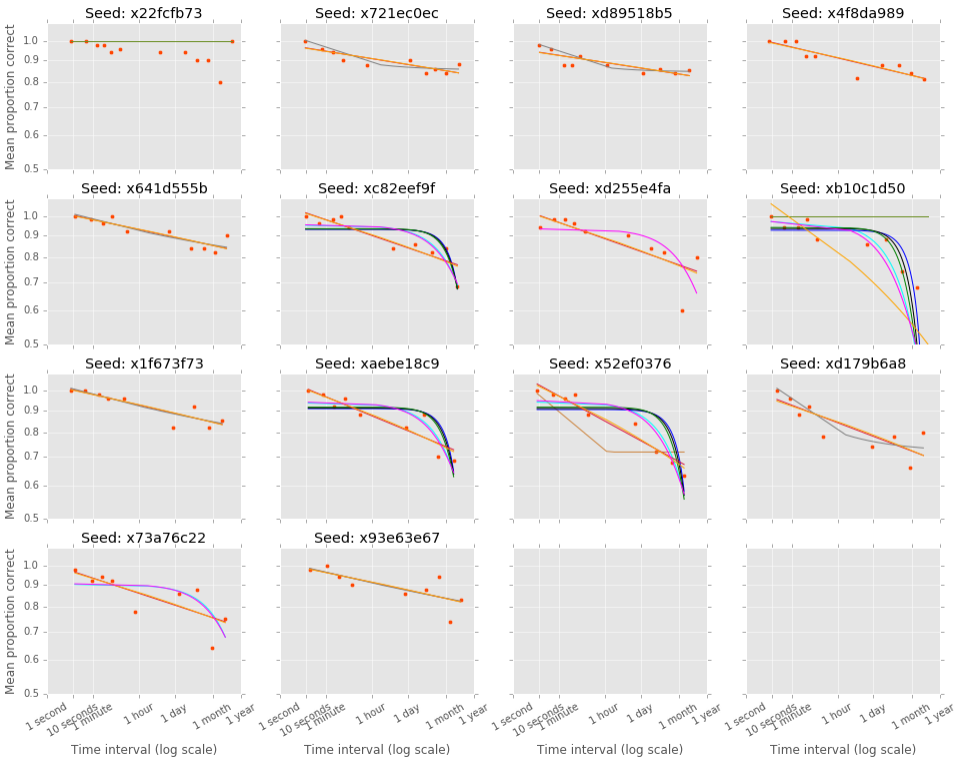


Figure 4.8: Adding and Subtracting Fractions with Like Denominators Word Problems candidate function fits for a bin size of 50, on a per question basis for fits with R^2 in excess of 0.5.

4.4.4 Replication in KA Lite Data

The KA Lite data is significantly more sparse than the data from the Khan Academy site, due to the majority of its usage occurring in offline settings - meaning the majority of learning data never has the opportunity to be synchronized back to the central aggregation server from which the KA Lite data has been derived. Due to this, fits have not been made on a single exercise basis, but rather across all of the Mathematics exercises in the KA Lite data set, as the data is too sparse in the data set to make meaningful fits otherwise. This has the consequence that while most test-retest pairs in the data are bound to only a specific user, there is a longer tail of users who have many test-retest pairs in the data set (see Figure 4.9), however, meaning it may be possible to do meaningful per user fits. As with both of the Khan Academy data sets thus far considered, the majority of question interaction sequences (beyond singletons) are of length two, with a rapidly diminishing proportion of higher length interaction sequences (see Figure 4.10).

Finally, an examination of the distribution of time intervals for the test-retest pairs (see Figure 4.11) shows that while a large proportion of the test-retest pairs have time intervals on the order of minutes, similarly to the Khan Academy mathematics data the distribution of time intervals is nearly bimodal, with a high number occurring at significant delays, on the order of many days. While the ratio of test-retest pairs of short intervals to long intervals for the KA Lite data is 1.

Retention Curves

A cursory examination of the data for initial test-retest pairs in Figure 4.12, shows that the retention across the data is markedly different in the KA Lite data, as

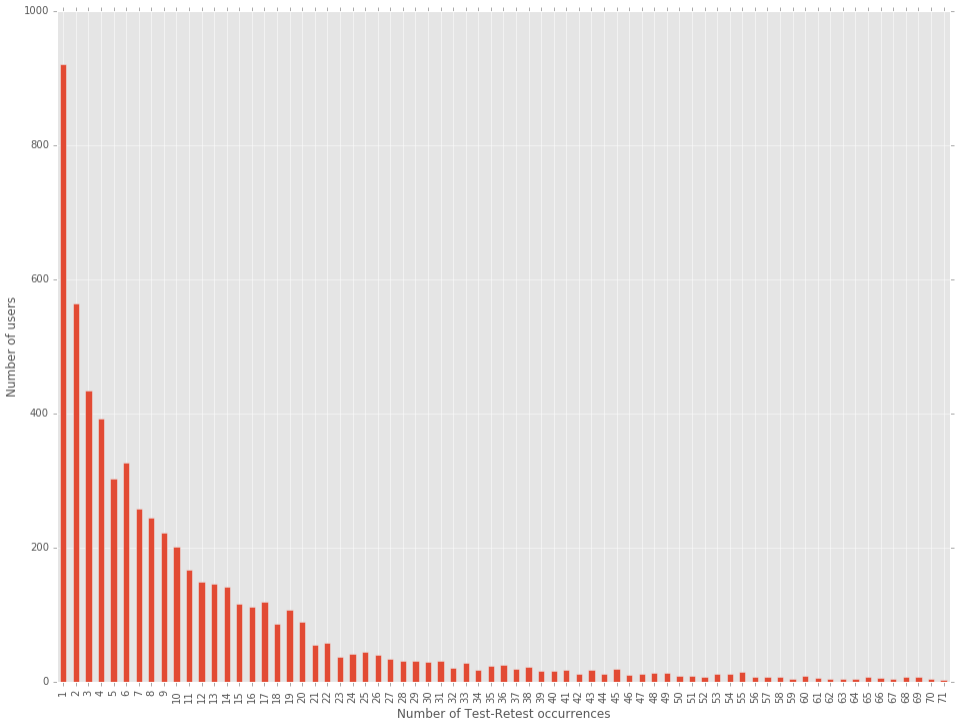


Figure 4.9: Distribution of KA Lite test-retest pairs across users

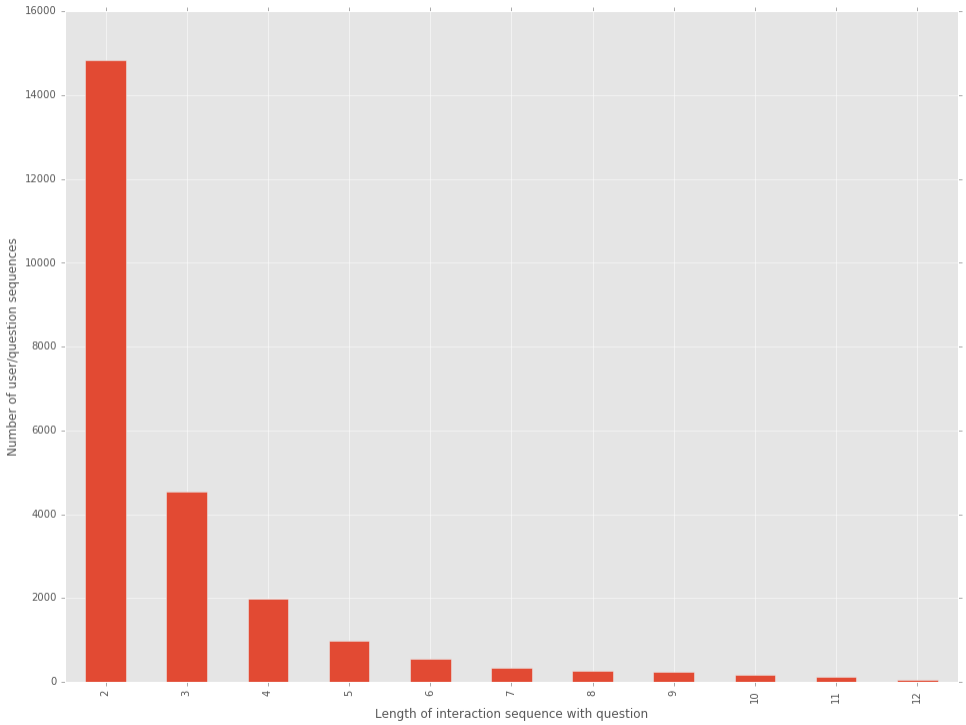


Figure 4.10: Histogram of user/question engagements in KA Lite data.

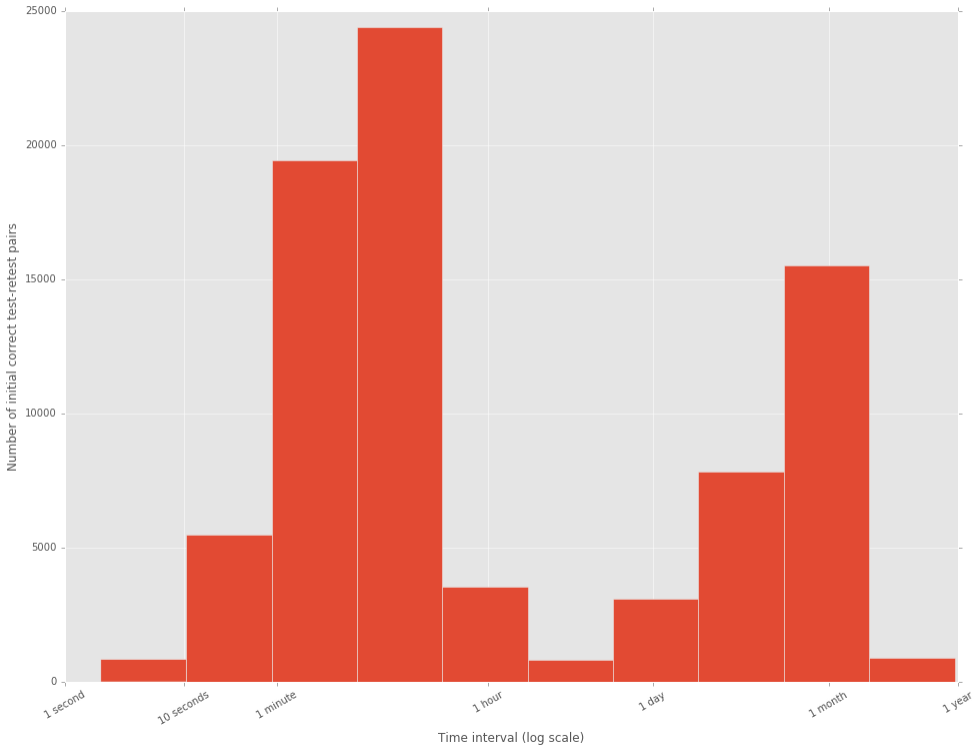


Figure 4.11: Distribution of KA Lite test-retest time intervals



Figure 4.12: Scatter plots of aggregated KA Lite data points at bin sizes from 100 to 1000

compared with either the Khan Academy declarative fact learning or Mathematics data. While there appear to be strong similarities at short time scales, at longer time scales, the retention in the KA Lite data either stays level with that at shorter durations, or rises. One plausible explanation for this is that a large amount of the data is collected in the context of learners using the software in formal education contexts (while the majority of the Khan Academy data is self driven, even if the learners may well be in formal education concurrently). As such, there may well be exogenous factors, such as teacher mediated instruction and question sets, that are driving improved performance for learners outside of the software.

This is further evidenced by candidate function fits to the KA Lite data, with

Table 4.3: R^2 values for different bin sizes for aggregating KA Lite data across all users and questions

Bin size	10	25	50	100	250	500	1000
Exponential	0.006	0.014	0.023	0.040	0.052	0.106	NaN
Exponential-Square Root	0.018	0.040	0.066	0.102	0.135	0.197	0.266
Hyperbolic	0.007	0.015	0.026	0.043	0.055	NaN	NaN
Hyperbolic-Power	0.086	0.190	0.325	0.482	0.672	0.000	0.000
Hyperbolic-Square Root	0.020	0.044	0.073	0.111	0.146	0.209	0.275
Linear	0.006	0.012	0.022	0.038	0.049	0.104	0.176
Logarithmic	0.078	0.172	0.294	0.434	0.600	0.726	0.751
Power	0.081	0.178	0.305	0.451	0.624	0.753	0.778
Sum of Exponentials-Rubin	0.087	0.193	0.331	0.492	0.689	0.817	0.878
Two Trace Model	0.088	0.194	0.332	0.493	0.697	0.830	NaN

every function at every bin size failing to achieve anything close to a good fit to the data, as shown in Table 4.3.

4.5 Exercise Level Analysis

4.5.1 Exercise Level Data for Adding and Subtracting Fractions with Like Denominators Word Problems

While the same underlying user data is used for this analysis, the test-retest pairs that are used as the basis for data aggregation are grouped subsequent engagements with the same exercise (which practices the same Mathematical skill), as opposed to subsequent engagements with precisely the same question. As such, the mean length of engagement sequences (Figure 4.13) is naturally much longer than for the item level data, with a longer tailed distribution, as some learners will engage with the same exercise many times.

Another consequence of this change in aggregation scheme is that the time

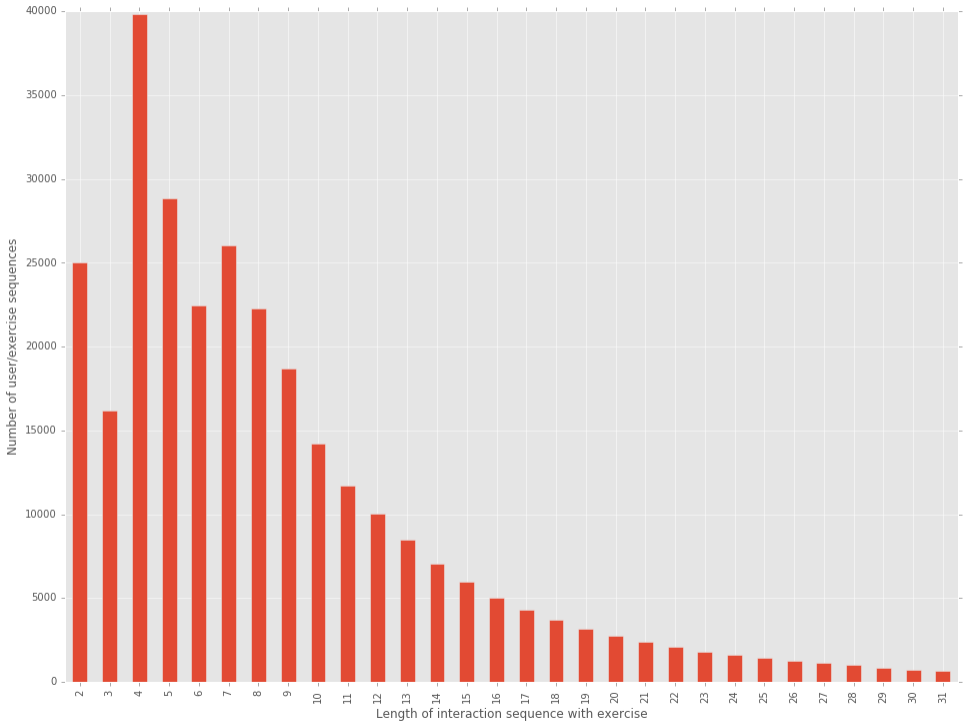


Figure 4.13: Histogram of user/exercise engagements in Adding and Subtracting Fractions with Like Denominators Word Problems data.

intervals between initial test-retest pairs are more heavily distributed within shorter time intervals, with a ratio of approximately 5 : 1 of shorter to longer time interval data points - the distribution is shown in Figure 4.14. However, one important additional consequence in looking at retention is that users who have very long retention intervals have not engaged with the exercise at all on the Khan Academy site in the interim, whereas in the item level data, there has frequently been repeated interaction with similar (but not the same) questions between engagements with the same question. If learning these kinds of mathematics procedures happens largely by generalization from many examples, then it is expected that we should see better long term retention in the item level data, than in the exercise level data.

4.5.2 Retention Curves

Scatter plots of the aggregated data at various bin sizes (Figure 4.15) show a markedly different pattern from both the equivalent item level data from the Khan Academy, and the item level data from KA Lite. This is shown even more starkly in the candidate function fits (Table 4.4) where all of the models, except the sum of exponentials model, fail to provide a good fit to the data. Examining these fits graphically (Figure 4.17) shows that the sum of exponentials function is able to achieve good performance due to its flexibility in curvature across different time scales (after initial declining performance, it is able to flatten out to minimize mean squared error on the highly varied but, on average, steady performance at longer time scales).

If the data are subdivided and fits made separately to long and short intervals, we see even more clearly that any goodness of model fit is being driven by the denser short interval episodes. In examining the model fits (Table 4.5), we see significantly

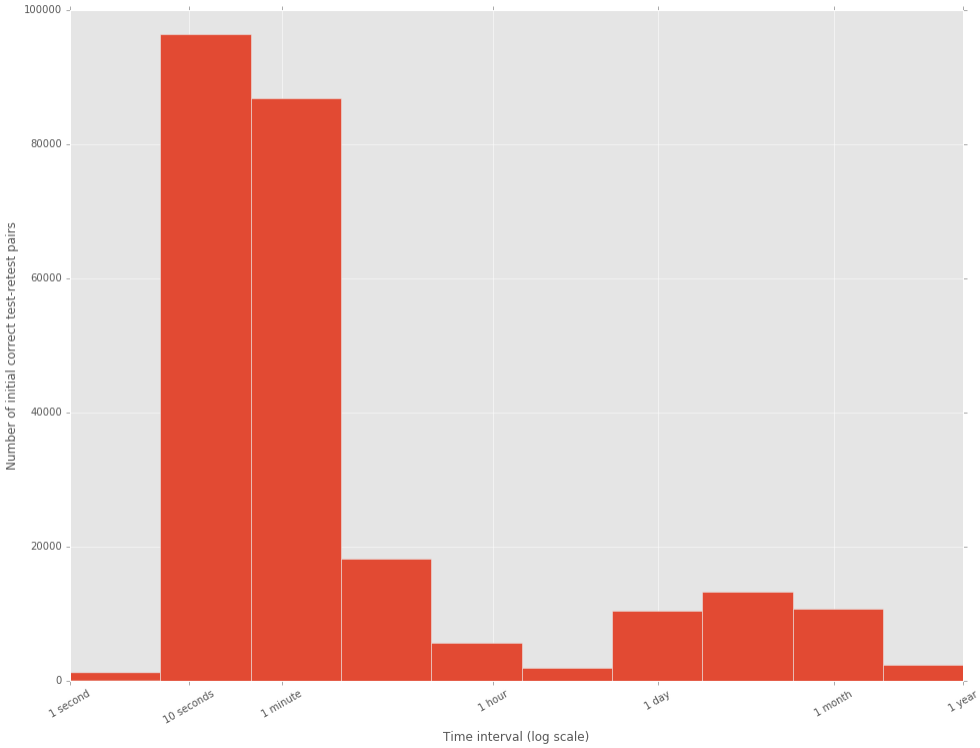


Figure 4.14: Distribution of Adding and Subtracting Fractions with Like Denominators Word Problems exercise test-retest time intervals

Table 4.4: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users by exercise

Bin size	10	25	50	100	250	500	1000
Exponential	0.016	0.021	0.024	0.025	0.022	0.026	0.023
Exponential-Square Root	0.044	0.059	0.066	0.071	0.071	0.074	0.071
Hyperbolic	-1.715	-2.302	-2.601	-2.790	-2.933	-2.975	-3.043
Hyperbolic-Power	0.196	0.272	0.174	0.633	0.679	NaN	0.715
Hyperbolic-Square Root	0.048	0.064	0.072	0.077	0.078	0.081	0.078
Linear	0.013	0.018	0.020	0.020	0.018	0.022	0.021
Logarithmic	0.181	0.243	0.275	0.294	0.306	0.312	0.313
Power	0.203	0.272	0.308	0.330	0.343	0.349	0.350
Sum of Exponentials-Rubin	0.532	0.714	0.000	0.864	0.947	0.917	0.925
Two Trace Model	NaN	-1.859	NaN	NaN	-2.355	-2.392	0.953

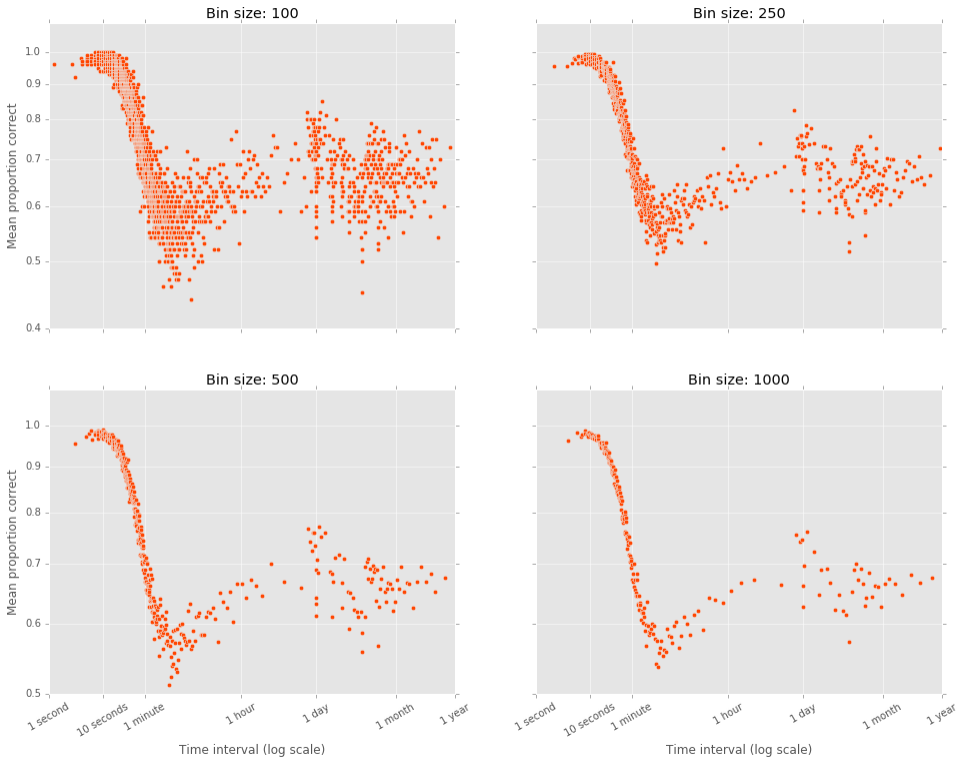


Figure 4.15: Scatter plots of aggregated Adding and Subtracting Fractions with Like Denominators Word Problems exercise data points at bin sizes from 100 to 1000

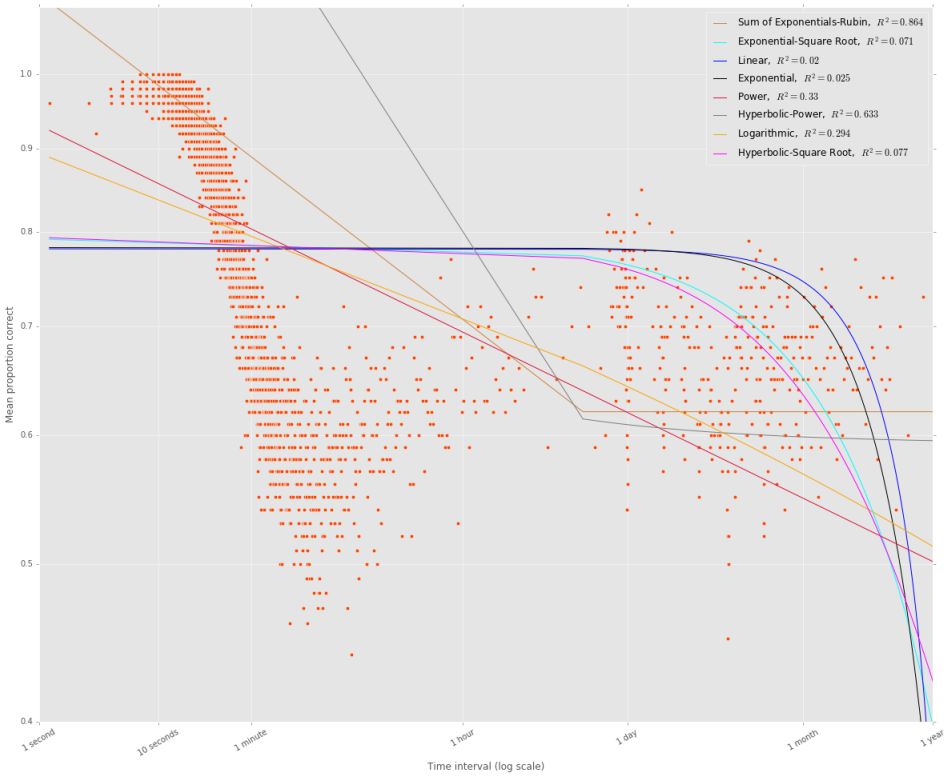


Figure 4.16: Adding and Subtracting Fractions with Like Denominators Word Problems exercise candidate function fits for a bin size of 100

Table 4.5: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users by exercise for short time intervals

Bin size	10	25	50	100	250	500	1000
Exponential	0.137	0.179	0.196	0.201	0.220	0.204	0.176
Exponential-Square Root	0.358	0.469	0.520	0.545	0.576	0.567	0.542
Hyperbolic	0.263	0.345	0.380	0.396	0.423	0.408	0.373
Hyperbolic-Power	0.008	0.220	0.102	0.249	NaN	0.833	NaN
Hyperbolic-Square Root	0.449	0.588	0.652	0.687	0.722	0.720	0.703
Linear	0.087	0.114	0.124	0.127	0.140	0.132	0.120
Logarithmic	0.477	0.626	0.695	0.734	0.769	0.771	0.764
Power	0.510	0.668	0.743	0.786	0.823	0.829	0.827
Sum of Exponentials-Rubin	NaN	NaN	NaN	0.914	0.953	0.963	0.969
Two Trace Model	-1.150	-1.509	0.880	0.934	-1.855	-1.879	0.990

Table 4.6: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users by exercise for long time intervals

Bin size	10	25	50	100	250	500	1000
Exponential	-19.127	-42.215	-71.983	-107.458	-174.729	-233.429	-297.740
Exponential-Square Root	-19.127	-42.215	-71.983	-107.458	-174.729	-233.429	-297.740
Hyperbolic	0.000	-42.191	0.001	-107.181	-173.670	-230.708	-290.830
Hyperbolic-Power	0.004	0.009	0.015	0.022	NaN	NaN	NaN
Hyperbolic-Square Root	0.002	0.004	0.007	0.011	0.012	0.008	0.001
Linear	0.000	0.001	0.001	0.001	0.000	0.000	0.008
Logarithmic	0.004	0.009	0.015	0.022	0.032	0.036	0.033
Power	0.004	0.009	0.014	0.022	0.032	0.036	0.033
Sum of Exponentials-Rubin	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Two Trace Model	0.000	0.000	0.000	0.000	0.000	0.000	0.000

better performance of the previously well performing power and logarithmic functions, although the sum of exponentials and two trace models still out perform them at the highest aggregation bin sizes. By contrast, for the long interval data, no model is able to provide a good fit at any bin size (Table 4.6). This appears to be due to a large shift in the distribution of answers from the latter part of short interval distribution to the first part of the large interval distribution.

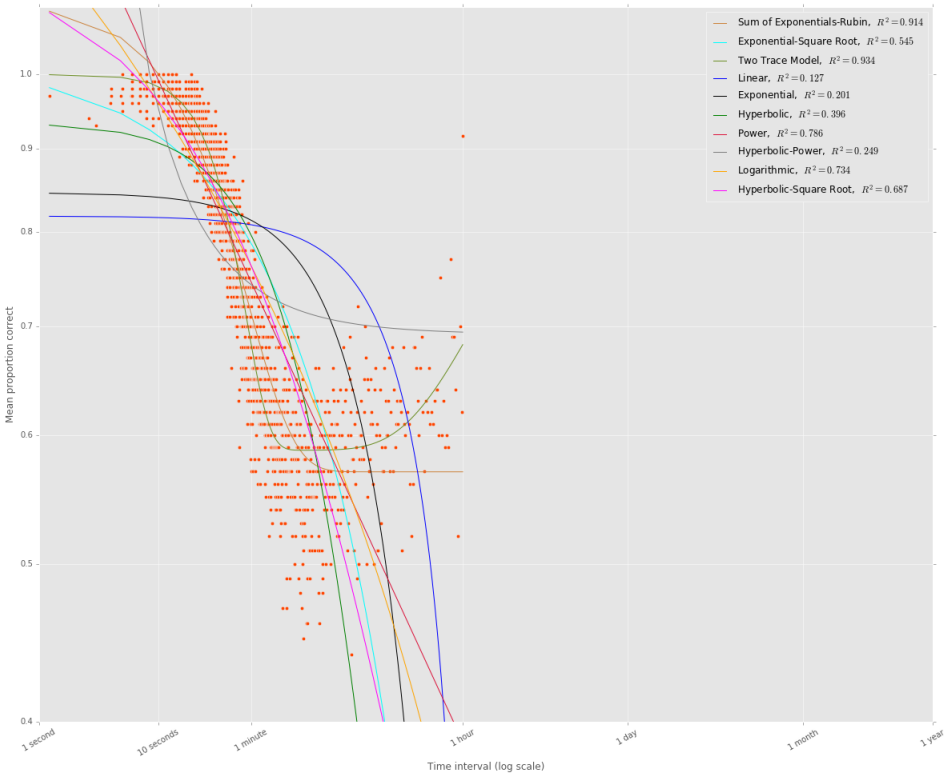


Figure 4.17: Adding and Subtracting Fractions with Like Denominators Word Problems exercise candidate function fits for short time intervals, for a bin size of 100

4.5.3 Signs of Consolidation

There are several potential causes of this effect - other activity by all learners on the Khan Academy site in the interim has caused them to increase their ability to carry out these exercises across this time interval, or some process (such as consolidation) is occurring that serves to actually improve performance across this time interval rather than decrease it. The other alternative is that there is another underlying common cause for poor performance and returning to answer questions at the time intervals that are closer to the one hour mark - which learners who return on a subsequent day lack. In order to consider this possibility, and see if the changes over time are significant (rather than simply artifactual), a rolling Pearson χ^2 test (Pearson, 1900) is conducted on binned subsets of the data at consecutive time points. This compares ten thousand responses clustered around one time point with ten thousand responses clustered at the subsequent time point to determine if there is a statistically significant difference between them. The p value of the Pearson χ^2 test is compared with a Bonferroni corrected (Bonferroni, 1936) significance threshold, $\alpha = \frac{0.05}{N}$ - where N is the number of sequential tests conducted - each distribution that is found to be significantly different from the distribution that preceded it is then compared with all preceding bins until a Pearson χ^2 test is no longer significant at the corrected α. Lastly, the α is corrected further for the additional expanding Pearson χ^2 tests to provide a final corrected threshold for significance due to the large number of tests that the rolling and expanding procedures combined requires.

Examining the results of these tests for the initial correct test-retest pairs, there is a significant increase in performance across the short interval to long interval boundary - i.e. around the time that consolidation takes place (Figure 4.18). It is

important to reiterate that this means that learners, returning to the exercise after having got their first attempt at a question correct were more likely to be successful on their subsequent attempt if they waited on the order of 25 hours, as opposed to closer to approximately 1 hour. This effect remains (but is reduced in significance), even with the inclusion of test-retest trials where learners got the first answer incorrect in the test-retest pair (Figure A.2). This effect remains apparent when examining all consecutive attempts on the exercise (Figure A.3), however, it is now conflated with the results of Khan Academy's own spaced learning algorithm, whereby learners must wait sixteen hours after being flagged for initial mastery of an exercise before returning, so the delayed group in this set have been enriched with learners who are more successful. Finally, if we only examine attempts made before mastery (which for this exercise is set on Khan Academy as being 'get 5 in a row correct') the effect remains, and the confounding increase in performance at the sixteen hour mark is no longer evident in the mean performance graph.

4.5.4 Replication in KA Lite Data

As with the question level data, due to the sparseness of the KA Lite data, the whole corpus of Mathematics exercises were used in order to provide a comparably sized sample to the Khan Academy data. As can be seen in Figure 4.20, the typical length of a sequence of engagements in the KA Lite data on a per exercise basis is almost double Khan Academy (this is due to a slightly more stringent mastery criterion, requiring eight out of the last ten attempts to be correct for mastery, as opposed to the five in a row correct required by Khan Academy for the exercise under analysis).

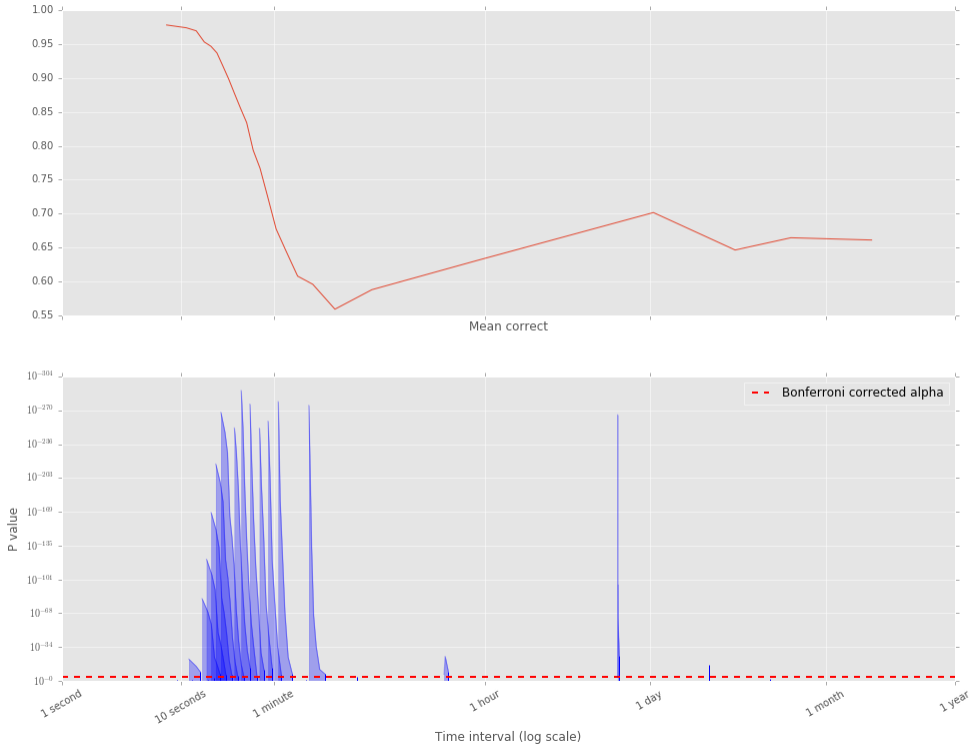


Figure 4.18: Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for initially correct test-retest trials for Adding and Subtracting Fractions with Like Denominators Word Problems

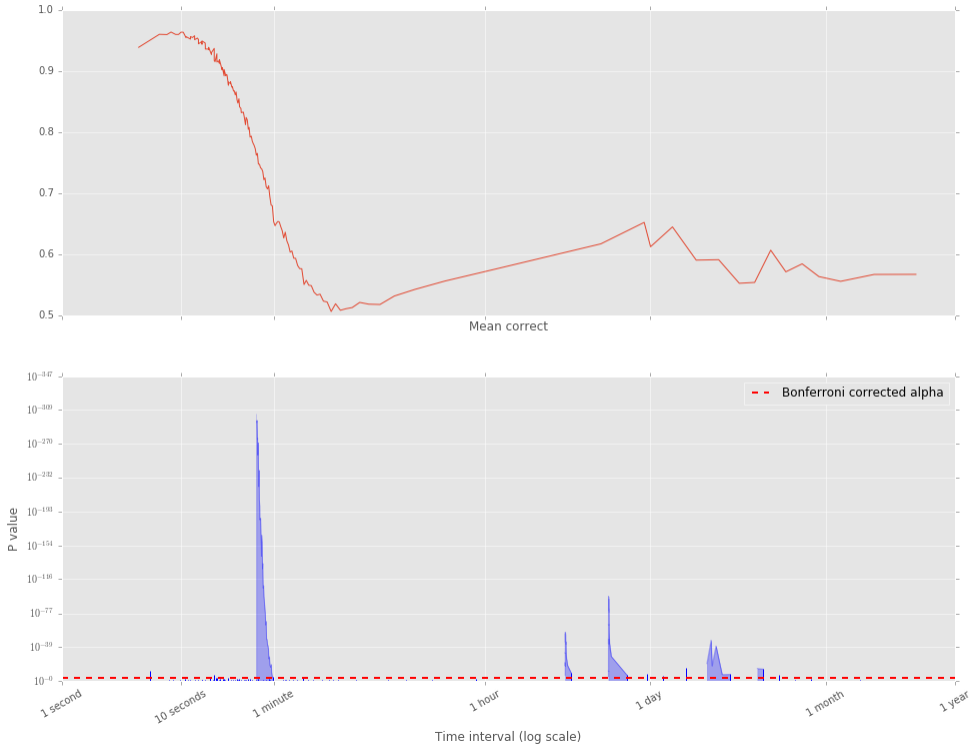


Figure 4.19: Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all trials before mastery has been achieved for Adding and Subtracting Fractions with Like Denominators Word Problems

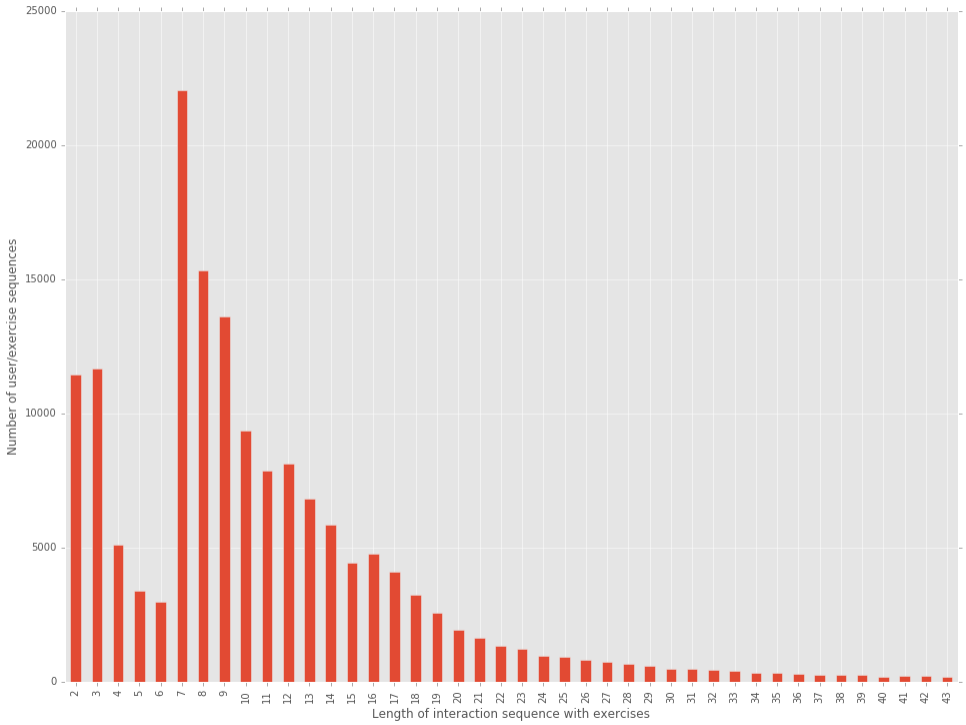


Figure 4.20: Histogram of user/exercise engagements in KA Lite data.

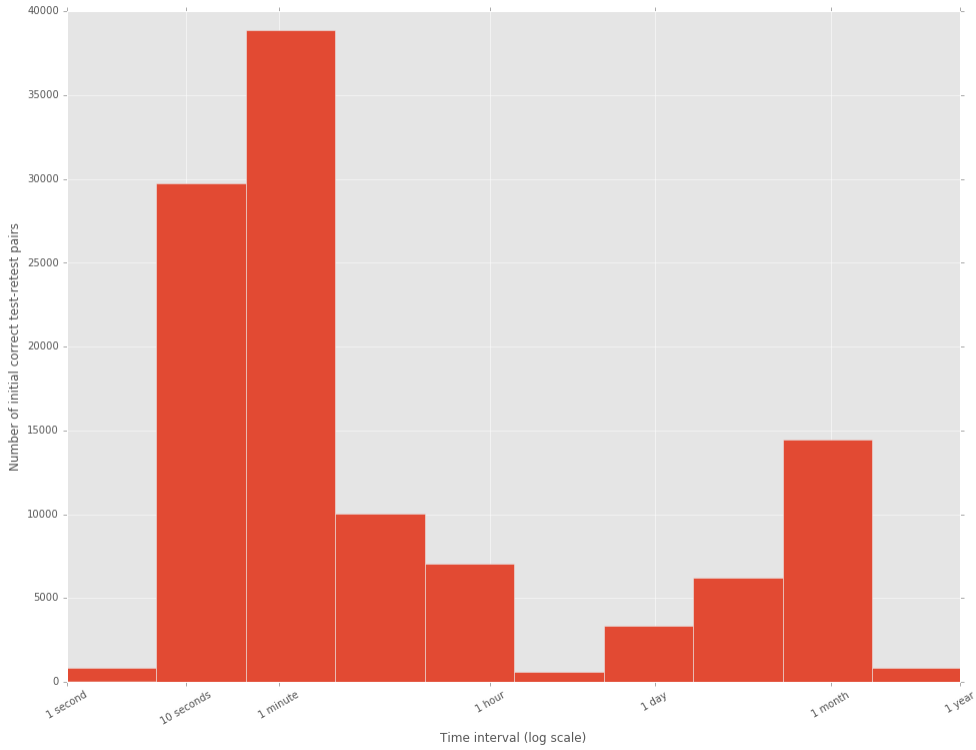


Figure 4.21: Distribution of KA Lite exercise test-retest time intervals

Similarly to the Khan Academy data, the grouping by exercise type, as opposed to question produces a much larger number of short intervaled test-retest pairs, as shown in Figure 4.21, with approximately 3.23 times as many short interval (less than an hour) data points than long interval data points.

Retention Curves

Model fits for the KA Lite exercise grouped data suffer from the same issue as the Khan Academy data - model fits (shown in Table 4.7) that perform well are driven by a good fit at short time intervals, and then fitting to a mean value at longer intervals. However, unlike the KA Lite data, it seems likely from the distribution in

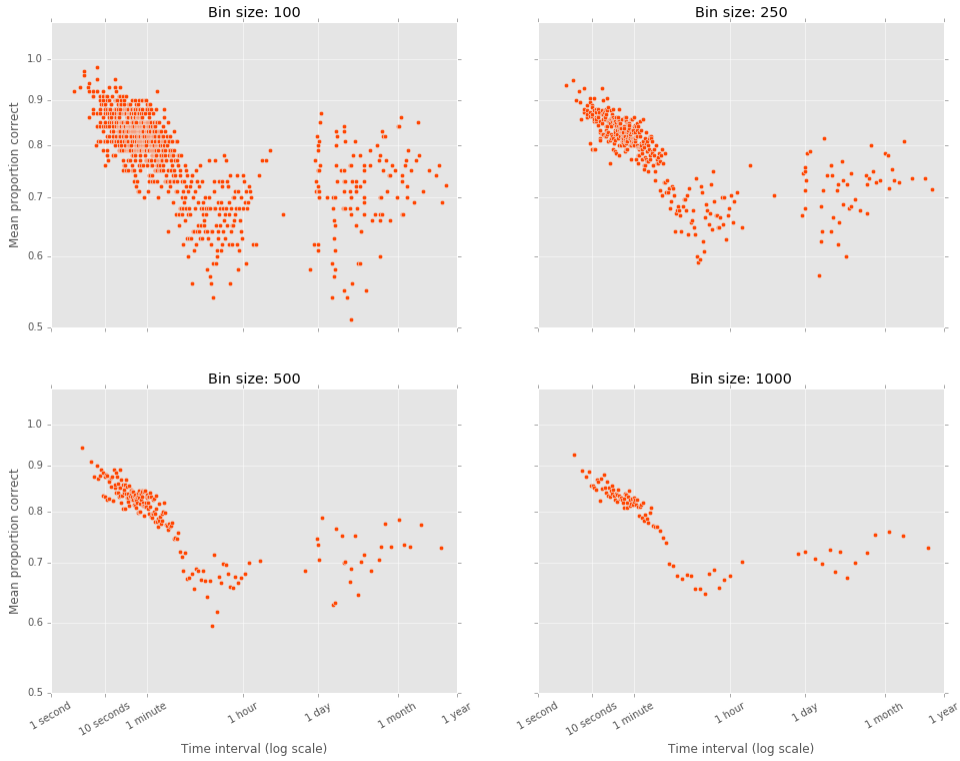


Figure 4.22: Scatter plots of aggregated KA Lite exercise data points at bin sizes from 100 to 1000

Figure 4.22 that any thing akin to the potential consolidation effect identified in the Khan Academy data is not driving these results, as there is high variability in the mean proportion correct occurring at longer time intervals.

Parallel to the Khan Academy data, conducting model fits for the short (Table 4.8) and long (Table 4.9) time intervals only confirms this assumption that the model fits are being driven by the short time interval data. Good fits are again recovered for power and logarithm fits in the short intervals, while no good fits are found for the longer time interval.

Table 4.7: R^2 values for different bin sizes for aggregating KA Lite data across all users by exercise

Bin size	10	25	50	100	250	500	1000
Exponential	0.004	0.007	0.010	0.013	0.016	0.017	0.021
Exponential-Square Root	0.018	0.033	0.046	0.058	0.071	0.076	0.077
Hyperbolic	0.004	0.008	0.010	0.014	0.017	0.018	0.022
Hyperbolic-Power	0.162	0.301	0.425	0.537	0.667	0.000	0.000
Hyperbolic-Square Root	0.019	0.035	0.049	0.062	0.076	0.081	0.082
Linear	0.004	0.007	0.009	0.012	0.016	0.016	0.020
Logarithmic	0.108	0.200	0.283	0.357	0.442	0.479	0.507
Power	0.114	0.211	0.298	0.376	0.466	0.505	0.535
Sum of Exponentials-Rubin	0.188	0.350	0.494	NaN	0.774	0.841	0.904
Two Trace Model	0.182	0.338	-5.403	-6.824	-8.473	0.814	-9.953

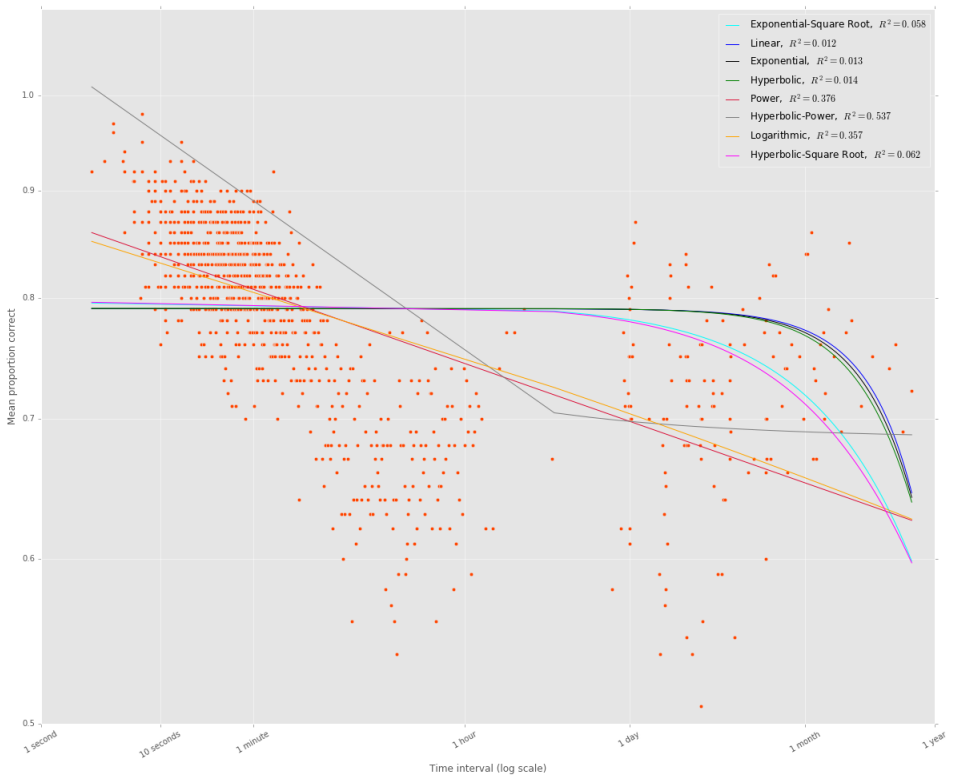


Figure 4.23: KA Lite exercise candidate function fits for a bin size of 100

Table 4.8: R^2 values for different bin sizes for aggregating KA Lite data across all users by exercise for short time intervals

Bin size	10	25	50	100	250	500	1000
Exponential	0.115	0.215	0.307	0.394	0.459	0.480	0.535
Exponential-Square Root	0.168	0.313	0.446	0.576	0.681	0.725	0.784
Hyperbolic	0.124	0.231	0.330	0.424	0.495	0.518	0.574
Hyperbolic-Power	NaN	NaN	NaN	NaN	0.000	0.000	0.000
Hyperbolic-Square Root	0.175	0.326	0.465	0.600	0.711	0.758	0.818
Linear	0.107	0.200	0.285	0.365	0.426	0.444	0.498
Logarithmic	0.192	0.357	0.508	0.659	0.786	0.850	0.904
Power	0.190	0.355	0.505	0.654	0.781	0.846	0.899
Sum of Exponentials-Rubin	0.203	0.372	0.529	0.686	0.820	0.890	0.942
Two Trace Model	0.195	0.363	-5.409	-7.014	0.800	0.867	0.000

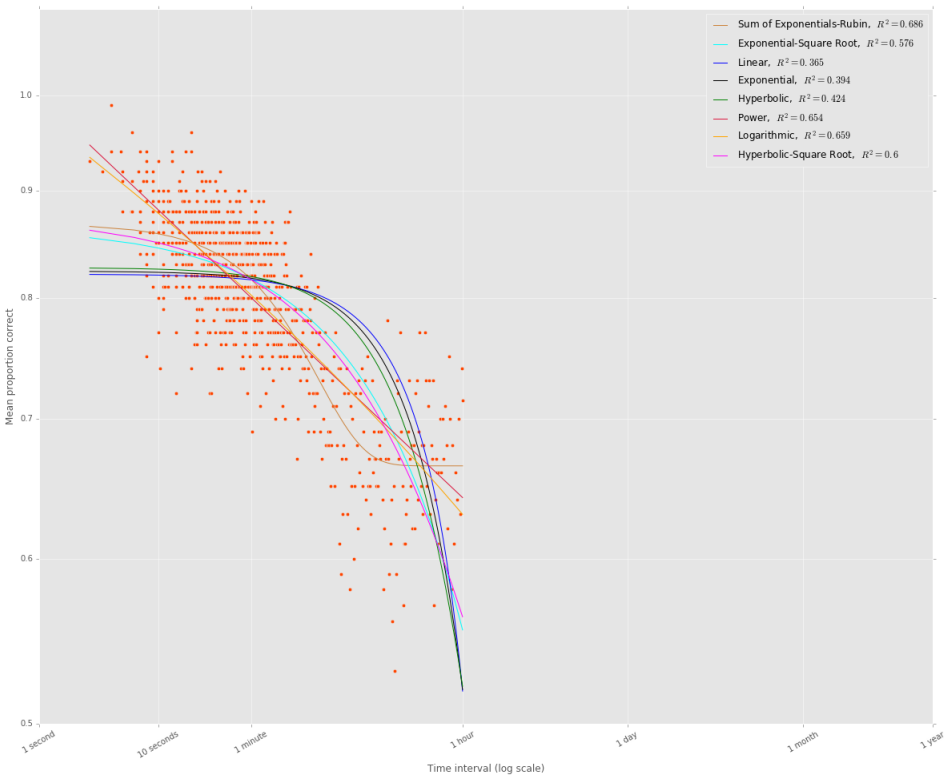


Figure 4.24: KA Lite exercise candidate function fits for short time intervals, for a bin size of 100

Table 4.9: R^2 values for different bin sizes for aggregating KA Lite data across all users by exercise for long time intervals

Bin size	10	25	50	100	250	500	1000
Exponential	-18.841	-37.643	-61.446	-101.057	-189.864	-293.078	-383.496
Exponential-Square Root	-18.841	-37.643	-61.446	-101.057	-189.864	-293.078	-383.496
Hyperbolic	0.003	0.006	0.011	0.017	0.024	0.056	0.119
Hyperbolic-Power	0.007	0.014	0.024	0.038	NaN	NaN	NaN
Hyperbolic-Square Root	0.006	0.013	0.021	0.033	0.055	-283.452	0.164
Linear	0.003	0.007	0.012	0.018	0.026	0.059	0.123
Logarithmic	0.006	0.012	0.019	0.032	0.062	0.097	0.144
Power	0.006	0.012	0.020	0.032	0.062	0.098	0.146
Sum of Exponentials-Rubin	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Two Trace Model	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Signs of Consolidation

Similarly to the Khan Academy data, therefore, we are driven to ask if there may be something else that is driving these effects, with no apparent evidence of systematic forgetting at longer time scales. Carrying out the same analysis to investigate the possibility of some sort of consolidation effect (Figure 4.25), we see no such differential at a timescale that would be explained by consolidation. Instead, the apparent uptick in performance appears to happen at a time scale of tens of days. As suggested above, one potential confound that makes the KA Lite data different to the Khan Academy data is that a large proportion of the KA Lite data is aggregated from the use of the software in formal school systems, meaning that the learning that is happening when engaging in the KA Lite system is supplemented by other learning opportunities that would not be registered in the learning data. In fact, 45% of the data in the KA Lite corpus comes from a single project to use KA Lite in low income schools in India.

If we exclude the data from this source (Figure 4.26) we see that while there is still a significant drop at the same rough time point, the gradient of the curve at this point is downwards, rather than upwards. Showing that embedding the learning

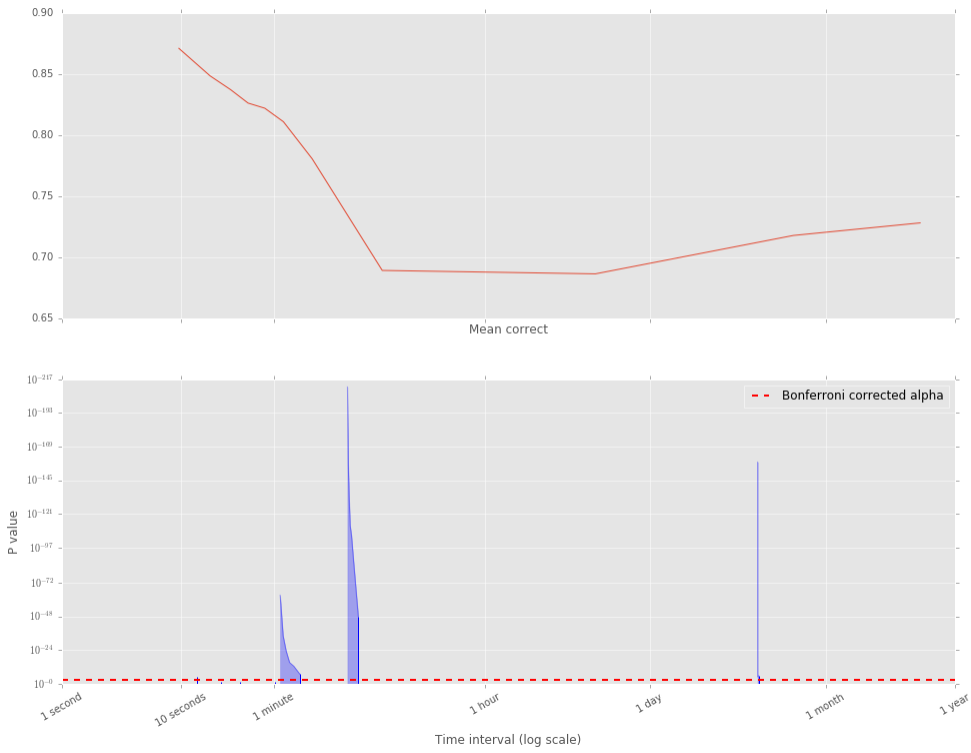


Figure 4.25: Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for initially correct test-retest trials for KA Lite

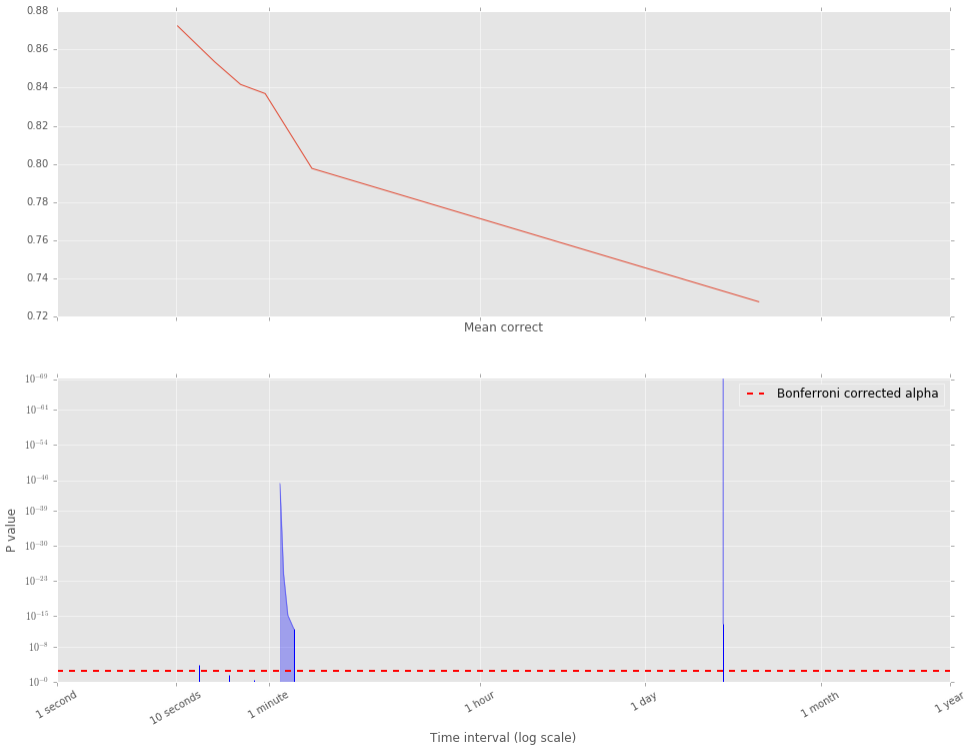


Figure 4.26: Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all trials before mastery has been achieved for KA Lite

program within in a school context may provide a significant amelioration of what otherwise might be quite a drastic drop in performance.

However, within this data, conducting model fits without the data from the highly structured implementation of KA Lite, it seems that we recover some of the forgetting behaviour identified in the item level and declarative fact learning (Table 4.10)

Table 4.10: R^2 values for different bin sizes for aggregating KA Lite data across all users (excluding those in a particular structured school environment) by exercise

Bin size	10	25	50	100	250	500	1000
Exponential	0.006	0.014	0.023	0.040	0.052	0.106	-10.675
Exponential-Square Root	0.018	0.040	0.066	0.102	0.135	0.197	0.266
Hyperbolic	0.007	0.015	0.026	0.043	0.055	-10.591	-11.904
Hyperbolic-Power	0.086	0.190	0.325	0.482	0.672	0.000	0.000
Hyperbolic-Square Root	0.020	0.044	0.073	0.111	0.146	0.209	0.275
Linear	0.006	0.012	0.022	0.038	0.049	0.104	0.176
Logarithmic	0.078	0.172	0.294	0.434	0.600	0.726	0.751
Power	0.081	0.178	0.305	0.451	0.624	0.753	0.778
Sum of Exponentials-Rubin	0.087	0.193	0.331	0.492	0.689	0.817	0.878
Two Trace Model	0.088	0.194	0.332	0.493	0.697	0.830	-16.955

4.6 Summary

The analysis of forgetting in Mathematics learning showed several interesting results. While analyses on a per question basis produced retention curves similar in form to that found in declarative fact learning in the Khan Academy data, exercise level analyses defied this, producing unexpected gains at longer time intervals that might be explicable as some sort of consolidation effect, or may be a result of some exogenous influence that all Khan Academy users are likely to be engaged with (such as formal schooling). This is perhaps reinforced by the analyses on the KA Lite data, where, although not showing the characteristic uptick at the one day mark that the Khan Academy data did, longer time intervals failed to show consistent forgetting behaviour. However, once the confounding impact of the organized, school based program that makes up a large proportion of the KA Lite data was removed, some apparent forgetting behaviour that conformed to power fits is recovered.

Chapter 5

Spacing in Mathematics Learning

From the foregoing, it seems likely that absent any additional and exogenous factors, that there will be forgetting effects over time that are not too dissimilar to those that are to be expected for declarative fact learning. As a consequence, it would seem to follow that if forgetting pertains to these domains, then engaging in spaced retrieval practice, just as with declarative fact learning (Cepeda et al., 2009), should produce better results in learning to engage with Mathematics exercises.

However, as Kang, Lindsey, Mozer, and Pashler (2014) note, there is controversy in the literature about whether such spaced practice should use a fixed or expanding interval for optimal results. While much of the literature has seemed to have favoured a fixed interval spacing regime, it seems that this has been a result of examining experiments on a short time scale (less than a day), while their own work across a the course of a month favoured the expanding interval, not just for final retrieval test performance, but also for producing comparable performance by successful engagement at the intervening testing stages. The reason for this is that while a learner may engage in a fixed interval learning regime, they may either have to use more sessions

to produce comparable performance at some distal retrieval time point, or due to the linear spacing of the intervening practice sessions, then they may have to engage repeatedly at a particular session before successfully retrieving. In either circumstance, the learning they are engaging in is less efficient, and hence not producing optimal use of the learner's time in producing their desired learning outcomes.

5.1 All Time Intervals

As the Khan Academy data set for exercise interaction sequences is dominated by short interval time intervals, it is to be expected that an examination of all the data for the Adding and Subtracting Fractions with Like Demoninators Word Problems exercise would produce a favourable result for learners who engaged in a pattern most similar to a fixed interval spaced study schedule.

As this is not a controlled study, it is necessary to create a measure that summarizes the similarity to either a fixed or expanding interval study schedule. As all spaced study has to be considered relative to the retention interval that a learner will produce final recall at, this was used as the baseline for examining the temporal sequence of study in which a learner engaged.

For fixed intervals, the time between each successive engagement with an exercise was examined, and divided by the retention interval between the penultimate and ultimate attempts at the exercise. This produced a sequence of ratios of the intervening interstimulus intervals compared to the final retention intervals, the standard deviation of this sequence was then taken in order to provide a standardized metric of how similar to an idealized fixed study interval the actual study interval was. In the case that each interval was of identical length, then the standard deviation

would be zero.

For expanding intervals, each interval was divided by the subsequent interval (following the geometrically expanding interval successfully employed by Kang et al. (2014)), and then the standard deviation taken of these quotients. In the case of an idealized geometrically expanding interval, the ratio of each interval to its subsequent would be equal, and the standard deviation would once again be zero.

For the purposes of model fitting, the negative of the above described measures was taken, so that a positive coefficient would be indicative of a positive effect of the associated spacing scheme.

In order to assess the impact of the fixed interval and expanding interval measures, a base model is fitted (Table 5.1), and then one with the fixed interval (Table 5.2) and another with the expanding interval (Table 5.3). While a large proportion of variance is predicted solely from the other variables available for prediction (with the majority coming from the average correct that a learner achieved on the preceding exercise instances), it is evident that both the models that include fixed and expanding interval measures produce moderately better results. However, in the case of the fixed interval, the positive coefficient indicates that the influence of being closer to a fixed interval schedule produces a higher likelihood of success, while the negative coefficient for the expanding interval indicates that having closer to an expanding interval actually has a negative impact on performance. However, from the literature, as described by Kang et al. (2014), this is to be expected, due to the predominance in the data set of retention intervals of less than a day.

Table 5.1: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set using average correct, retention interval, total attempts, and total time as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.262047	0.403280	0.529878	0.645084	0.741571	0.787032	0.820553
Ridge regression test score	0.256648	0.400233	0.527037	0.640301	0.740597	0.787064	0.817872
Ridge regression training score	0.262184	0.403491	0.530204	0.645750	0.742755	0.788704	0.822014
average_correct	0.096209	0.106494	0.120316	0.136417	0.150208	0.158056	0.163829
retention_interval	-0.008988	0.005454	0.013657	0.079941	0.079888	0.104901	0.107525
total_attempts	0.003251	0.026794	0.047723	0.070131	0.089720	0.101094	0.109360
total_times	0.037382	0.014577	-0.001699	-0.077319	-0.085590	-0.115567	-0.120677

Table 5.2: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set using average correct, retention interval, total attempts, total time, and fixed interval fit as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.282541	0.442909	0.579322	0.695379	0.790224	0.832413	0.861780
Ridge regression test score	0.277227	0.441126	0.577139	0.691617	0.788712	0.832625	0.861832
Ridge regression training score	0.282710	0.443237	0.579816	0.696093	0.791429	0.833939	0.864467
average_correct	0.095026	0.103234	0.115594	0.129220	0.141604	0.147955	0.154870
fixed_interval_std	0.029506	0.033473	0.034116	0.032728	0.031049	0.029675	0.028377
retention_interval	-0.092092	-0.111960	-0.108775	-0.072807	-0.080555	-0.009304	0.019563
total_attempts	-0.006389	0.011708	0.029895	0.049131	0.067266	0.077610	0.087814
total_times	0.117618	0.130504	0.120281	0.076722	0.077205	0.001665	-0.030248

Table 5.3: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions, fitted against a holdout data set using average correct, retention interval, total attempts, total time, and expanding interval fit as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.327936	0.504910	0.639692	0.741835	0.814139	0.842048	0.859643
Ridge regression test score	0.323064	0.502303	0.632450	0.733260	0.809719	0.840009	0.859249
Ridge regression training score	0.328094	0.505197	0.640115	0.742437	0.814878	0.843531	0.862464
average_correct	0.090058	0.091683	0.096584	0.103167	0.111370	0.118433	0.126066
expanding_interval_std	-0.056070	-0.063061	-0.065951	-0.065197	-0.060527	-0.055373	-0.049534
retention_interval	-0.057906	-0.079682	-0.108239	-0.111251	-0.158502	-0.106538	-0.031585
total_attempts	-0.026332	-0.024471	-0.020000	-0.011346	0.002537	0.016033	0.030863
total_times	0.081751	0.104089	0.131870	0.132532	0.175688	0.119332	0.040051

Table 5.4: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions for retention intervals greater than 24 hours, fitted against a holdout data set using average correct, retention interval, total attempts, and total time as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.157203	0.243700	0.348715	0.507490	0.676109	0.780426	0.848663
Ridge regression test score	0.153651	0.250261	0.364966	0.515644	0.688814	0.795498	0.861799
Ridge regression training score	0.157383	0.244230	0.349433	0.508543	0.678034	0.782897	0.850782
average_correct	0.059186	0.050547	0.048291	0.048981	0.049290	0.048117	0.044416
retention_interval	0.002236	0.017869	0.033524	0.005765	0.010003	0.002898	0.023409
total_attempts	-0.003274	-0.002881	-0.000232	0.003183	0.008597	0.009717	0.012242
total_times	0.005682	-0.008181	-0.023931	0.002533	-0.004278	0.002950	-0.019919

5.2 Long Intervals

For model fitting at only long intervals, two things are apparent - firstly, that the base model (Table 5.4) performs considerably better than in the short interval case. Secondly, that the very modest improvements in model performance when the fixed interval (Table 5.5) and expanding interval (Table 5.6) measures are introduced are produced by a reversal of the coefficients discovered in the broader fits carried out with the short and long retention interval data. If this result is reliable, this accords with the findings of Kang et al. (2014), that while the literature reports favourable results for the fixed intervals, in tests of longer retention intervals, expanding intervals are favoured.

5.3 Summary

While it is hard to draw definitive conclusions from the analyses outlined above, it seems relatively clear that the results conform to findings in the literature that over short time periods fixed intervals spacing regimes produce better results. However, this is of less interest in educationally relevant domains, as most learning of interest

Table 5.5: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions for retention intervals greater than 24 hours, fitted against a holdout data set using average correct, retention interval, total attempts, total time, and fixed interval fit as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.162445	0.251262	0.357254	0.514768	0.683639	0.786784	0.853286
Ridge regression test score	0.158495	0.257613	0.372299	0.522078	0.696462	0.801849	0.869001
Ridge regression training score	0.162753	0.251963	0.358281	0.516406	0.686533	0.789985	0.856862
average_correct	0.058922	0.049888	0.047423	0.048467	0.048199	0.046626	0.042894
fixed_interval_std	-0.011958	-0.009611	-0.008074	-0.006116	-0.005313	-0.004545	-0.003638
retention_interval	0.037133	0.063522	0.075867	0.028003	0.045265	0.004928	0.022484
total_attempts	-0.002867	-0.002864	-0.000939	0.001835	0.006929	0.006826	0.009352
total_times	-0.027550	-0.052247	-0.064451	-0.017603	-0.037451	0.003689	-0.016416

Table 5.6: R^2 values for different bin sizes for aggregating Adding and Subtracting Fractions with Like Denominators Word Problems data across all users and questions for retention intervals greater than 24 hours, fitted against a holdout data set using average correct, retention interval, total attempts, total time, and expanding interval fit as regressor variables.

Bin size	10	25	50	100	250	500	1000
R Squared	0.157203	0.243700	0.348834	0.508306	0.677405	0.786778	0.852525
Ridge regression test score	0.153647	0.250252	0.364708	0.515172	0.692178	0.801802	0.869504
Ridge regression training score	0.157413	0.244247	0.349846	0.509632	0.679903	0.789984	0.856774
average_correct	0.058997	0.050636	0.048701	0.049580	0.050012	0.049432	0.045530
expanding_interval_std	-0.001009	0.000426	0.001881	0.002516	0.003211	0.006049	0.005398
retention_interval	0.001992	0.018039	0.034992	0.006304	0.010054	0.003829	0.019161
total_attempts	-0.002934	-0.003054	-0.001135	0.001709	0.006075	0.004400	0.006745
total_times	0.005817	-0.008278	-0.024966	0.002774	-0.002916	0.004964	-0.012439

in education requires retention over longer than a day. In the analyses above, it seems that the use of an expanding interval spacing regime would have better results than a fixed interval regime, simply because the reversal of coefficients between the two implies that the expanding interval has a positive result, whereas the fixed regime has an equally negative impact on performance.

Chapter 6

The Impact of Forgetting of Prerequisite Knowledge on Subsequent Learning

6.1 Introduction

An issue of particular importance and interest in curricular planning and the learning of Mathematics in particular is how prior knowledge impacts subsequent learning. If a student has not used their skill at manipulating fractions for addition and subtraction, will this impact their ability to successfully learn to multiply fractions that require similar skills that underpin such operations? The adoption of spiral curricula in education has been one attempt to guard against the pertaining of this possibility, with learners regularly revisting content on a cyclic basis across the curriculum. Are such curricular structures necessary? Do students need to engage with learning in a spaced way with mathematics in order to ensure future recall to properly underpin

learning of skills that depend on these previously learned Mathematical procedures?

6.2 Prerequisite Skills in Khan Academy

Khan Academy identifies particular skills as prerequisites, or more precisely, many exercises identify one or more other exercises that embody some sort of prerequisite knowledge that should assist a student in engaging with this exercise. However, there is no mechanism in the Khan Academy interface that enforces student engagement through a particular path, although students are given suggested starting paths on the dependency structure through diagnostic testing. As a consequence of this, many student engagements with 'prerequisite exercises' happen subsequent to engagement with initial engagement with the exercise for which it is a supposed prerequisite.

6.3 Impact of Forgetting of Prerequisite Skill Learning

In order to assess the impact of forgetting of prerequisite skill learning on subsequent performance on dependent exercises, test-retest pairs were constructed similarly to previous analyses, except that instead of constructing pairs of first to second engagement on the same exercise, pairs were created of the last engagement on a prerequisite exercise to the first engagement on the exercise in question. As such a test-retest pair was created with a characteristic time interval, and a measure of correct or incorrect recall of the following engagement.



Figure 6.1: Scatter plots of aggregated Adding and Subtracting Fractions with Like Denominators Word Problems exercise data points at bin sizes from 100 to 1000

The form of this engagement can be seen in Figure 6.1, with consequent candidate function fits shown in Table 6.1, with an example visualized in Figure 6.2. What is clear is that any impact of prerequisite learning is strongly attenuated, and while it appears there may be some small effect of forgetting of prerequisite knowledge, it is not a strong predictor of learner success. However, as soon above, it is likely that there are other factors also involved in the maintenance of learner success over time that will not be captured by learner engagement in the Khan Academy website.

Table 6.1: R^2 values for different bin sizes for aggregating Adding And Subtracting Fractions With Like Denominators Word Problems Prerequisites data across all users

Bin size	10	25	50	100	250	500	1000
Exponential	0.000	0.000	0.000	0.000	0.000	0.001	0.008
Exponential-Square Root	0.002	0.004	0.007	0.011	0.016	0.017	0.010
Hyperbolic	0.000	0.000	0.000	0.000	0.000	0.001	0.008
Hyperbolic-Power	0.010	0.001	0.000	0.000	0.000	0.000	0.000
Hyperbolic-Square Root	0.002	0.004	0.007	0.011	0.016	0.017	0.010
Linear	0.000	0.000	0.000	0.000	0.000	0.001	0.008
Logarithmic	0.009	0.020	0.035	0.056	0.092	0.110	0.120
Power	0.009	0.021	0.036	0.057	0.093	0.111	0.121
Sum of Exponentials-Rubin	0.000	0.001	0.000	0.000	NaN	NaN	NaN
Two Trace Model	0.031	0.067	-9.419	-14.990	-25.007	-30.590	-36.933

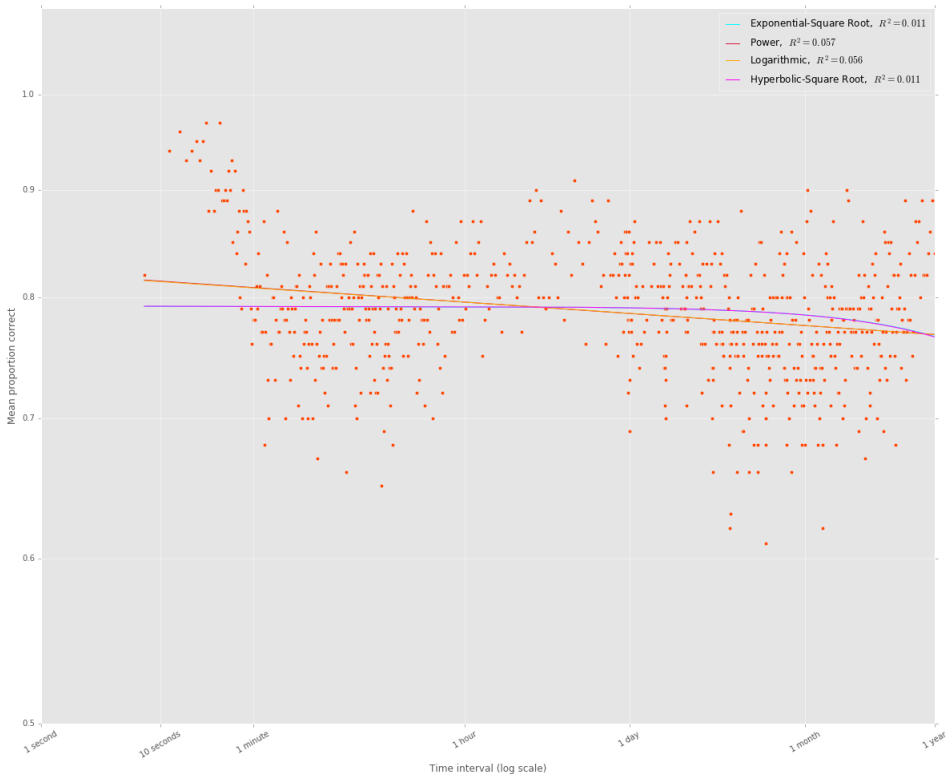


Figure 6.2: Adding and Subtracting Fractions with Like Denominators Word Problems candidate function fits for a bin size of 100

Chapter 7

Implications for Design for Computer Assisted Instruction Systems in Low Resource Contexts

7.1 Introduction

While it is interesting to see laboratory results replicated, corroborated, and nuanced by investigations in real world data sets - practical application of findings both in the laboratory and in ecologically valid replications in learning data sets are important if education is ever to achieve the goal of personalized, effective, efficient learning for all. Of particular interest to myself is the implications for design of the Kolibri platform, a learning environment for providing personalized instruction to learners in some of the most resource constrained environments in the world.

7.2 Design Constraints for Kolibri

There is a huge diversity in the range of resource constraints faced by potential users of Kolibri - one immediate constraint on computer assisted instruction is the availability and cost of hardware. Fortunately, as detailed in Alexandre (2014), over the past several years, the cost of computing devices has fallen rapidly, increasing the wide availability of low cost, low computer power devices that could potentially act as the hardware platform for computer assisted instruction.

Another important design constraint for Kolibri is that it, like the KA Lite platform (Alexandre, 2014; Tibbles & Alexandre, 2014) before it, is designed to be used in entirely offline settings, without any access to the Internet during the time of use. Further, as it is designed to serve a wide range of content from multiple content sources, it lacks the homogeneity of content across many different users that sites like Khan Academy and the PCOCs that allow for large scale data collection to allow the kind of inferences from data, such as those that have been made in this dissertation.

In addition to content being aggregated from multiple sources, the content that is available for use on a particular Kolibri instance is highly likely to have been 'remixed' from its source organizational structure. This process of content curation allows for sources from a wide range (and frequently US education system centric content sources) to be both reordered and aligned to different curricula around the world, and mixed in with content from different sources. Due to this highly prevalent remixing from multiple sources, the limited data available on local devices due to lack of connectivity, and the low computing power of the hardware on which Kolibri is deployed, predictive models would need to be pre-trained to estimate the model parameters, and then distribute a static model based on a particular set of content.

However, the exact set of content that students will be engaging with will vary widely, so precomputing such models from existing data (which will have been collected within a particular set of content) may not be helpful, as the content that is later distributed may have a very different set of content, as predictors based on now 'missing' content will be uninformative.

7.3 Personalized Review

7.3.1 High Data Content Libraries

In content libraries that already have a large amount of accumulated data, such as the Khan Academy content library, it seems possible from the foregoing that parameters and heuristics approximations of predictive models could be estimated in advance of deployment into low resource contexts in order to provide predictions about student performance based on engagement with other content. For example, it may be possible to estimate that because a student performed flawlessly on the prerequisite skill of adding and subtracting fractions with like denominators, then less evidence will be required to satisfy Kolibri that she is competent at adding and subtracting fractions with unlike denominators. Such initial competency heuristics might be similar to the mastery criteria currently employed by Khan Academy where learners are asked to get 'three in a row correct' or 'four correct out of five'. An advantage of these heuristics, trained from a large data set, is that not only do they avoid computationally costly updates to a computational model of student understanding, the requirements for mastery are made clear and transparent to a learner, without the need to carry out even more costly introspection of the model in order to show the user what evidence

would be required in order to satisfy the computer of their competence.

In these high data domains therefore, in spite of the potential for training highly sophisticated and complex computational models of learner competence, what is preferable in the low resource context (and perhaps in others as well) to promote efficient and engaged learning is to use the large data set to calculate suitable heuristics that are human parseable and will allow students to pursue meaningful, intelligible goals in their learning.

Beyond initial estimates of competence, developing good heuristics for spacing review and maintaining student competence over time is evidently very important from the foregoing analyses. The current model of Khan Academy to space review after a sixteen hour period will produce some gains, but it seems unlikely to produce long term retention due to the short period over which it is targeted. However, it seems from the foregoing analyses that a geometrically expanding interval (as suggested by Cepeda et al. (2009) for declarative fact learning) for review could help to improve long term retention and performance.

7.3.2 Low Data Content Libraries

While many content repositories already have a wealth of learning data associated with them, content for previously underserved populations (a primary target group for Kolibri) may either have had limited use, or be completely novel when deployed in the Kolibri ecosystem. As such, little is known about either the characteristic rate of forgetting of these materials, or what level of performane on these materials is suggestive of a learner having achieved competence in a skill or concept.

While these data could be learned over time through aggregation of the learner

data from many instances of the Kolibri software deployed in real world scenarios, the difficulties of aggregating significant quantities of data in such disconnected situations is evident from the limited data aggregated from an estimated 4000000 users of KA Lite. In order to address these issues therefore, allowing content experts to identify prerequisite skills that learners might require to engage with content, as well as initial mastery criteria for assessing learner competence at a skill is an important part of the content creation and curation process.

This is similar to the way that Intelligent Tutoring systems have historically created their models for student learning. Once a skill model has been specified (manually by a human content author), learning data of students engaging with educational software can be used to make inferences about more predictive models (Stamper & Koedinger, 2011). Cen et al. (2006) developed a semi-automated method for using data to improve the skill model created to underlie the different items. Additional approaches that seek to refine the Q-matrix (the mapping of skills underlying assessment items) rely on non-negative matrix factorization to produce a mapping based on observed student data (Desmarais, 2012; Desmarais & Naceur, 2013) - however, all of these approaches require expert human supervision to refine and iterate based on the models of student data. By contrast Matsuda, Furukawa, Bier, and Faloutsos (2015)'s approach provides an automated mechanism for skill model refinement, relying both on student log data, and also bag of words analysis of the content of learning material to determine underlying common skills. In spite of the automated nature of the skill model discovery, human interpretation is still required in order to implement the updated skill models.

Building on these techniques, heuristics could be developed for both recalculat-

ing skill models, and also for enhancements or changes to the dependency structure. In the online content curation interface for Kolibri channel creation using the Bag of Words heuristics of Matsuda et al. (2015) would allow for content creators to be prompted with potential prerequisite skills that they could then assign to a skill, thus allowing easier and quicker prerequisite skill determination within the user interface.

These computer aided human annotations of the knowledge structure could then be used as the basis for a much constrained design space for the knowledge structure. When in active use, Kolibri could then explore variations or perturbations of the design space around the knowledge structure, using a multi-armed bandit approach (Clement, Roy, Oudeyer, & Lopes, 2015) to find the optimal knowledge structure for the group of students using that instantiation of Kolibri, this would give Kolibri the flexibility to adapt its approach based on context, particularly in cases where local curriculum and sequencing of learning may produce a different implicit knowledge structure.

Finally, to deepen the understanding of the time course of retention developed in the foregoing, opportunistic 'random sampling' type data collection can be used to attempt to accrue data that will better inform the individual time course of retention for different content types, and contexts.

7.4 Learning Analytics

7.4.1 Introduction to Learning Analytics

As a field of research, Learning Analytics is relatively new. Drawing on expertise from information visualization, machine learning, human centered design,

and cognitive science, researchers are attempting to discover both useful indicators (Dyckhoff, Zielke, Bültmann, Chatti, & Schroeder, 2012) - signals of learning that can be inferred algorithmically from student behaviour - and effective ways of leveraging these indicators to improve learner outcomes (Ferguson, 2012).

Of particular interest is the use of learner analytics to drive dashboards for teachers. This has received some attention, but much of current practice has ignored the conclusions of Crawford, Schlager, Penuel, and Toyama (2008), whereby the needs of teachers and their role in the classroom as 'clinical professionals' is core to the design of dashboards that provide feedback about student learning. In addition, in order to provide meaningful feedback for teachers, Siemens (2012) recommends that analytics should emphasize 'sense-making, decision-making and action', echoing Norris, Baer, Leonard, Pugliese, and Lefrere (2008) demand that analytics be focused on actions that the learner or teacher can take. This is exemplified by the Course Signals project (Arnold & Pistilli, 2012) at Purdue, where data from a wide range of sources was used to create a predictive model for student failure, and then provide a restricted set of actions with which faculty members could follow up with the student.

Recent work in dashboards for teachers and students has been mostly focused on higher education(Verbert et al., 2014), with a few exceptions (such as work on showing primary school student progress in learning times tables (Ebner & Schön, 2013)). However, Arnold and Pistilli (2012) work on the 'Course Signals' project at Purdue University, shows the power of designing learner analytics with action in mind. The Course Signals project examined historical student data to understand and identify students at risk of failure early on in courses, and then made recommendations of a small subset of preselected actions that course Professors could engage in, in order

to intervene. The use of the Course Signals technology in courses early in students' academic careers significantly impacted retention into later semesters.

In addition to a gap in the literature, a recent report by the Bill & Melinda Gates Foundation (*Making Data Work for Teachers and Students*, 2015) showed that even the commercial provision of learner analytics was not providing K-12 teachers in the US with the kind of ease of access to data that they felt they needed to effectively impact their instruction. The report surveyed 4650 teachers in the US, in grades K-12, predominantly in district controlled public schools. In addition to finding that learner analytics frequently placed significant burdens on teachers, which led to either to neglect of other parts of their teaching role, or to failing to engage effectively with the data, the report uncovered these systematic issues with learner analytics systems as a means to support instruction:

> - **Slow.** Data and the subsequent analysis come too late to be included in an ongoing cycle of teaching and learning. In these cases, data document instead of drive instruction.
> - **Not sufficiently granular**, meaning the information is not detailed enough to drive instruction. For example, data from formative assessments may identify broad standards where students struggled, but not the specific concepts or skill sets where misunderstandings occurred.
> - **Unable to track progress over time.** Single points of data reduce students to a score on narrow measures of student performance rather than revealing broader, ongoing trends that allow teachers to track progress over time.
> - **Inflexible.** Digital data can't be manipulated the way teachers want. In particular, teachers can't drill down to the level of detail needed to drive instruction.
> - **Inaccessible to students.** Many teachers say they want students to take ownership of their own learning and progress, but students don't have timely access to performance data that would allow them to do so.
>
> (*Making Data Work for Teachers and Students*, 2015)

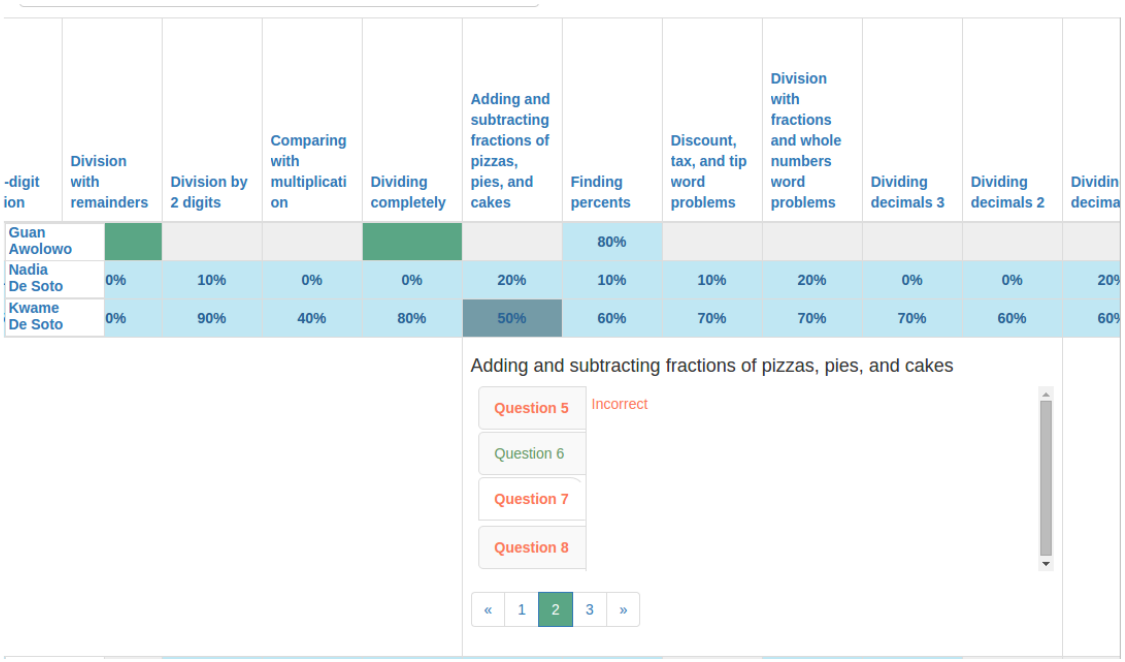


Figure 7.1: KA Lite Coach Reports Per Item Detail View

7.4.2 Learner Analytics in Kolibri

Kolibri's base coach reporting system builds on that of KA Lite (Alexandre, 2014; Tibbles & Alexandre, 2014), which already provides teachers with some capacity to examine student data in more detail (see Figure 7.1, and Figure 7.2). The above findings would be used to enhance this base capability by signalling to teachers in a highly salient way the kinds of information that has been determined to be diagnostic. Further, it would enhance the diagnostic detail of information made available to teachers, and make specific recommendations of action.

In contrast to much work in learning analytics that is concerned with showing current student progress, and tasks completed, the foregoing analyses highlight the importance of giving teachers information that is actionable, especially in terms of promoting student retention of learning over the medium to long term. However, in order to assist the teacher in this way, the system itself must have knowledge of the

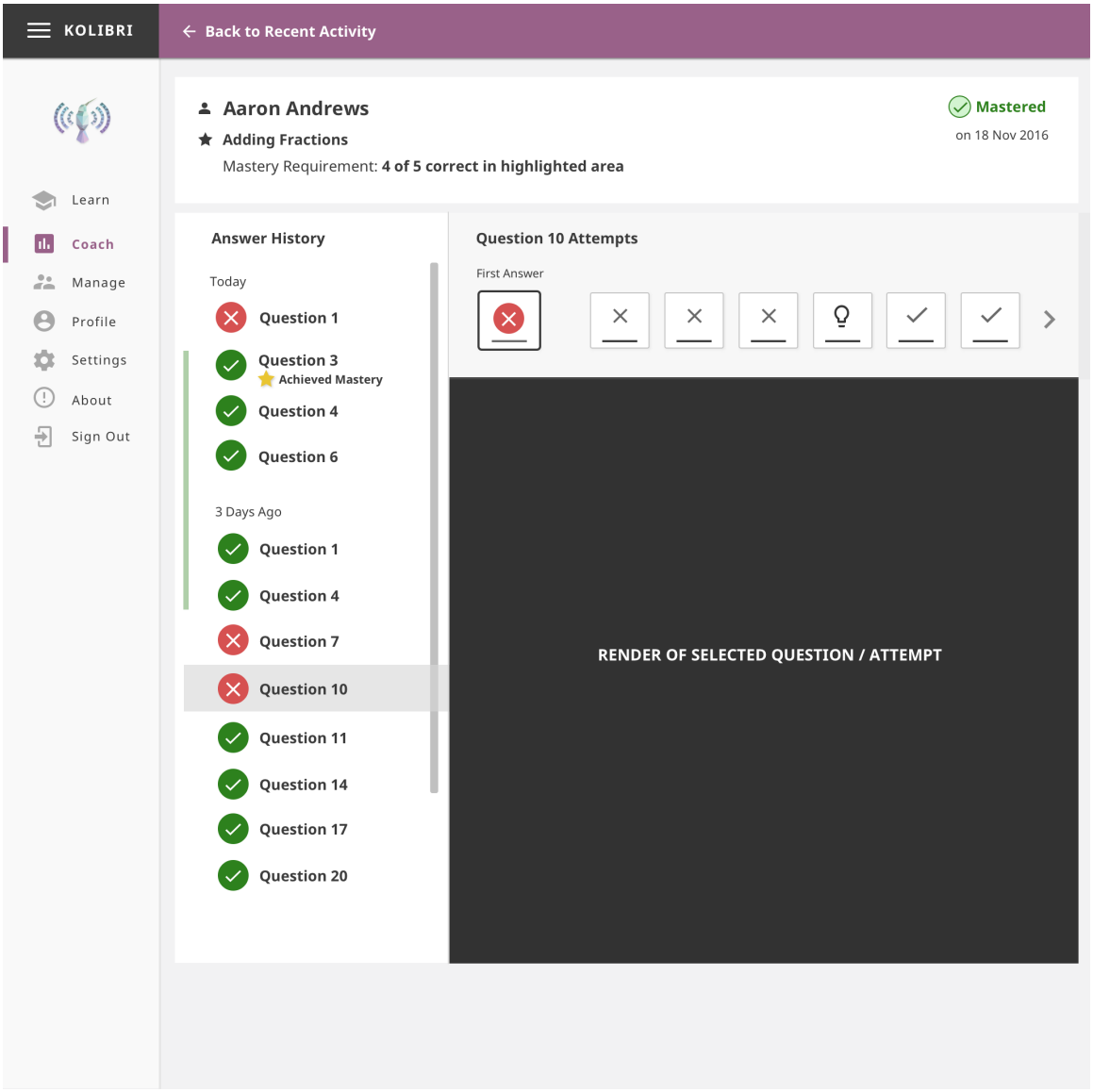


Figure 7.2: Kolibri Coach Reports Learner Detail View

desired time scales for retention (as a spaced review interval would be constructed very differently if the target for learning was required in two or three years, as compared with a month away from initial learning). As such, Kolibri should be able to not only make suggestions to a teacher about how best to support student learning, but also collect information from the teacher in order to understand the goals that the teacher is setting for learners in their classroom.

One way to achieve this by considering teacher needs is to offer the capacity for assignment of curriculum content to students - this adds an additional signal from the teacher that this is a subsection of the content available to a student that it is important for that student to learn, and can also give be used to elicit time frames in which the content is expected to be learned and retained.

Little work has been done to give a teacher an accurate assessment of the current state of students' knowledge across a range of topics (for example when reviewing for standardized tests). Combining the understanding of how a student's learning changes over time, with an understanding of the current learning goals that have been set for that student, would allow the system to prevent potential 'risk areas', where a student's prerequisite knowledge may be in danger of undermining future learning in a topic area. The use of a geometrically expanding interval for review, as suggested above, would also help to ensure that such estimates of learner understanding are updated at appropriate intervals, meaning that if a learner had engaged in learning activities away from Kolibri, their current state of learning could still be more accurately estimated with periodic review.

Just In Time Pedagogy

Actionable, intelligent, predictive learning analytics of this kind have the potential to provide a teacher in an otherwise under resourced classroom information at the correct level of granularity to provide useful interventions to students. However, many teachers find themselves under prepared by teacher training (or in many instances in the target populations for Kolibri, not trained at all). For a long time, the dissemination of good pedagogy (whether stemming from practical experience in the classroom with excellent teachers, or from systematic research) has been a difficult problem (Elmore, 1996). One advantage of the Kolibri platform is the ability to load many different kinds of content, including content that explicitly teaches good pedagogy to teachers - thus allowing the platform to be a resource not just for student learning but also ongoing teacher professional development.

However, by presenting actionable information targeted to particular skills, Kolibri offers another potential solution: a mechanism for making such pedagogic recommendations 'just in time'. By giving content creators the opportunity to add targeted pedagogical information to individual skills, Kolibri could allow for information about different, and very specific pedagogic strategies to be made available to teachers when that information is at its most salient to them (as opposed to during initial teacher training or continuing professional development). For example, if a student is having difficulties with understanding the concept of electrical resistance, the system could provide different conceptual models for the teacher to use in aiding the student's understanding, with additional information and conceptual questions to ask to probe the precise nature of the student's conceptual misunderstandings. This kind of detailed, nuanced, content knowledge driven pedagogy is almost never

directly taught in teacher professional development, due to the time consuming nature of covering the conceptual intricacies of individual items of content knowledge. Even if it were, practiced out of context, with no possibility of review and recall practice, it would likely be forgotten before it could be next used. However, by providing these opportunities in the contextually appropriate moment, the possibility for producing lasting improvement in teacher understanding and performance are increased.

Appendix A

Supplementary Tables and Figures

A.1 Declarative Fact Learning

A.2 Mathematics Learning

Table A.1: R^2 values for different candidate fit functions, aggregated at a bin size of 100, for pure declarative data across all users fitted per question

Seed	xe1d23308	xc00a72ee	x55902ded	xb9e2ec5e	xd5017e1f6c5f530b	xd120f7810a6006d1	x9b8b72b5e7effdad	x9877e517	x431cfbde	x647fd808ac350cd0	x92dba5edd33b481c	xc98f679a3d2afe96
N	3770	2973	2654	2498	2138	2076	1820	1814	1802	1533	1504	1471
Exponential	0.360	0.712	0.746	0.253	-17.256	0.298	-1.400	0.652	0.067	-1.678	-12.724	-2.263
Exponential-Square Root	0.479	0.770	0.623	0.366	0.156	0.362	0.711	0.792	0.018	0.785	0.195	0.622
Hyperbolic	0.367	-439.642	0.731	0.256	0.084	-28.394	-1.376	0.686	-976.386	-1.647	-74.960	0.584
Hyperbolic-Power	0.689	0.000	0.000	0.000	0.720	0.000	0.000	NaN	0.097	NaN	0.546	0.868
Hyperbolic-Square Root	0.489	0.772	0.598	0.376	0.162	0.367	0.702	0.825	0.018	0.799	0.198	0.635
Linear	0.353	0.712	0.757	0.249	0.082	0.298	0.678	0.631	0.067	0.692	0.159	0.581
Logarithmic	0.686	0.633	0.263	0.701	0.537	0.705	0.773	0.921	0.074	0.868	0.357	0.813
Power	0.688	0.628	0.253	0.711	0.556	0.716	0.784	0.905	0.074	0.869	0.369	0.834
Sum of Exponentials-Rubin	0.645	0.015	0.002	0.030	0.000	0.000	0.954	-4.968	0.735	0.956	0.000	0.000
Two Trace Model	-2.901	0.596	0.000	-2.601	-2.555	-4.117	-1.583	-4.968	0.735	-2.226	-6.741	-2.973

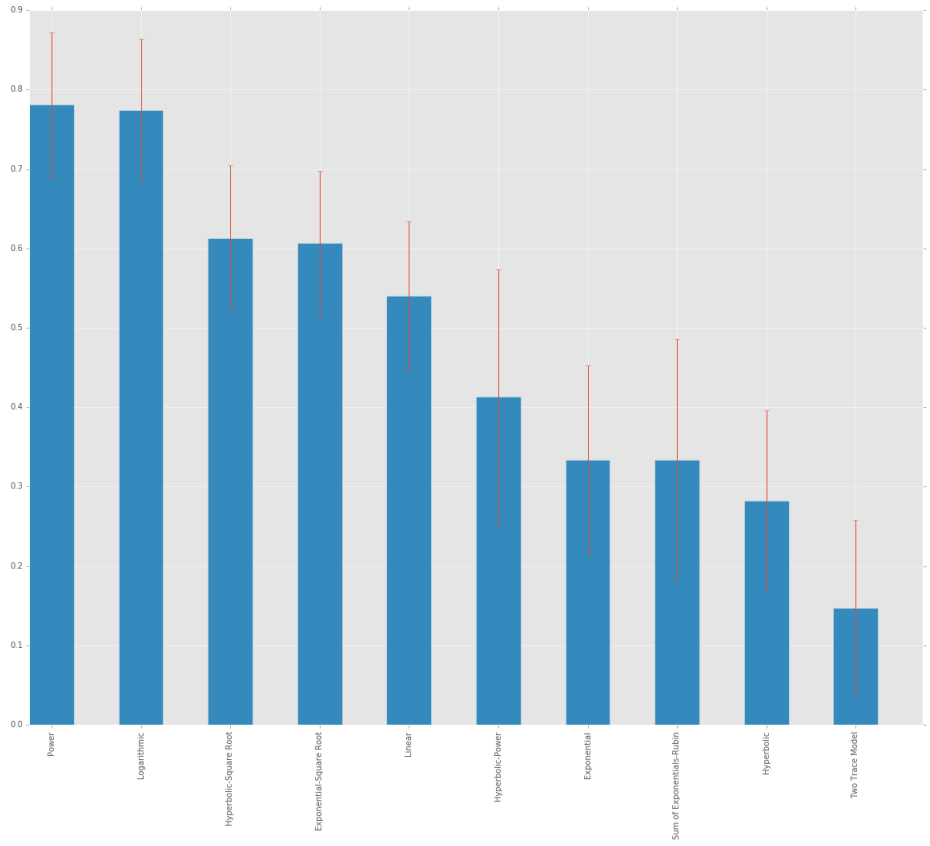


Figure A.1: Declarative fact learning candidate function proportion of varianced explained compared to best fitting model, on a per question basis.

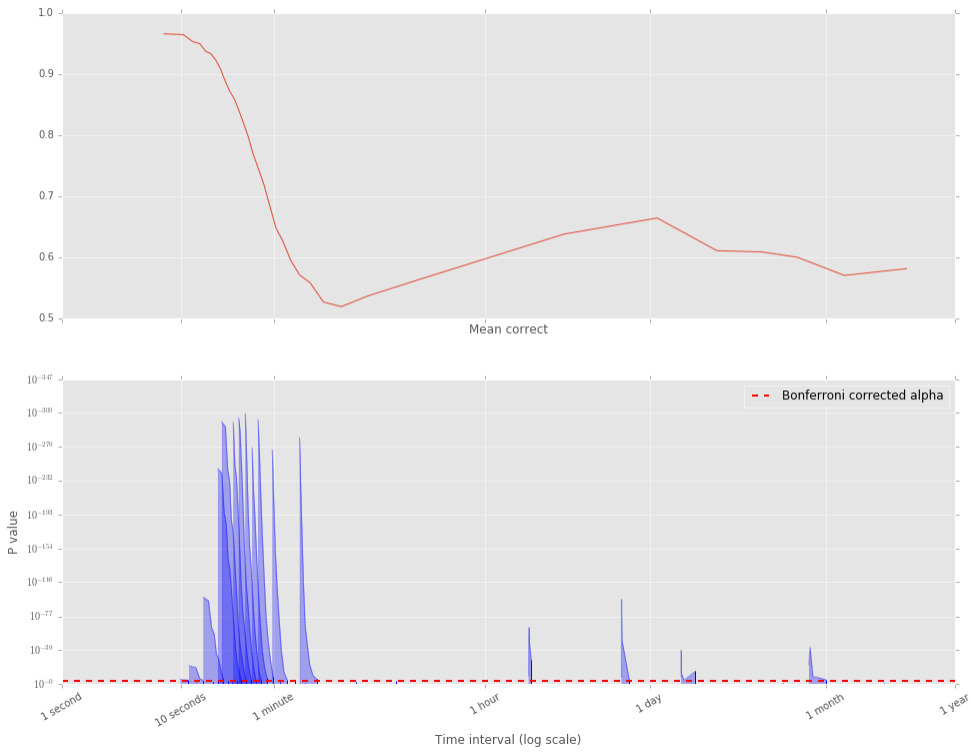


Figure A.2: Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all test-retest trials

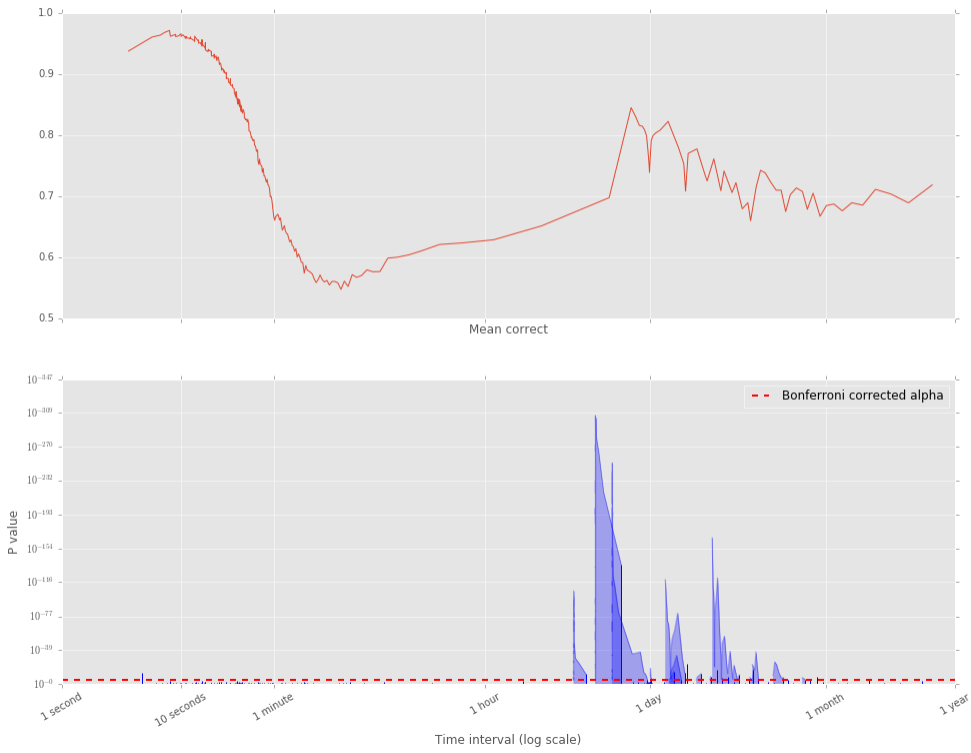


Figure A.3: Graph of changes in mean proportion correct and significance of changes in response distributions between time points, binning at 10000 responses per time point for all trials

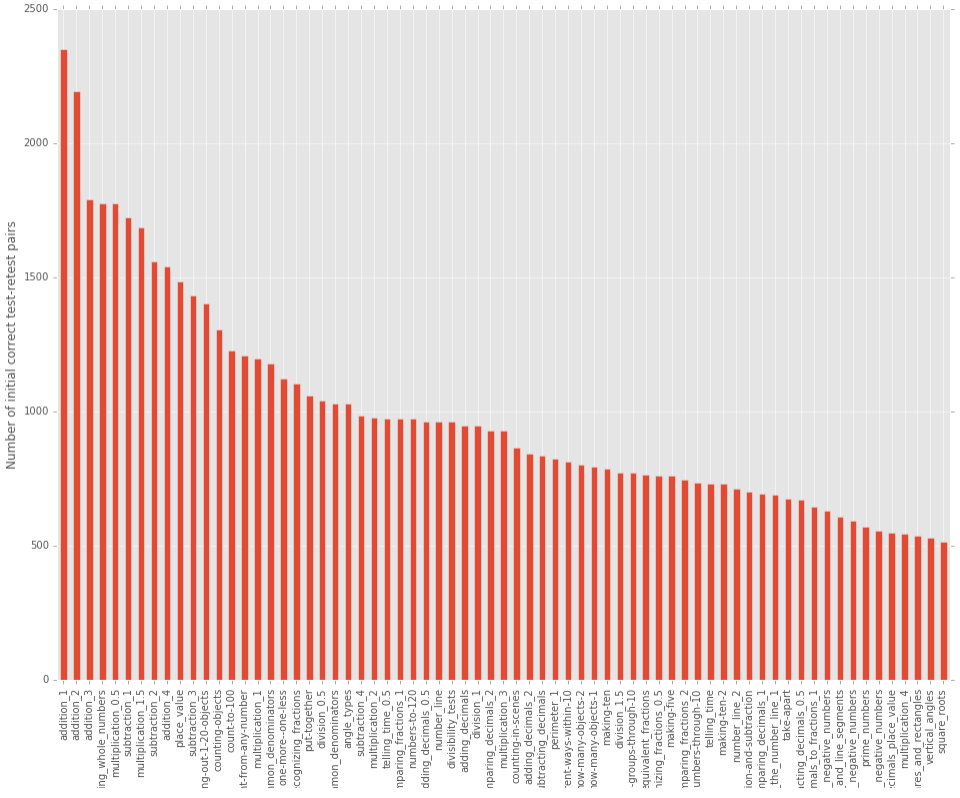


Figure A.4: Histogram of most commonly used exercises in KA Lite data.

Bibliography

Abel, M. & Roediger, H. L. (2016). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format. *Memory & Cognition*, 1–12. doi:10.3758/s13421-016-0641-8

Abott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements*, *11*(1), 159.

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. doi:10.1002/acp.1391

Akers, K. G., Martinez-Canabal, A., Restivo, L., Yiu, A. P., Cristofaro, A. D., Hsiang, H.-L. (, ... Frankland, P. W. (2014). Hippocampal Neurogenesis Regulates Forgetting During Adulthood and Infancy. *Science*, *344*(6184), 598–602. doi:10.1126/science.1248903

Alexandre, J. (2014). *Computer Assisted Learning in a (Dis-)Connected Age: Challenges and Approaches to Digital Education and Equal Access* (PhD Dissertation, University of California, San Diego, San Diego, California).

Altmann, E. M. (2009). Evidence for temporal decay in short-term episodic memory. *Trends in Cognitive Sciences*, *13*(7), 279. doi:10.1016/j.tics.2009.04.001

Ammons, R. B., Farr, R. G., Bloch, E., Neumann, E., Dey, M., Marion, R., & Ammons, C. H. (1958). Long-term retention of perceptual-motor skills. *Journal of Experimental Psychology*, *55*(4), 318–328. doi:http://dx.doi.org/10.1037/h0041893

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, *7*(3), 522–530.

Anderson, R. B. (2001). The power law as an emergent property. *Memory & Cognition*, *29*(7), 1061–1068. doi:10.3758/BF03195767

Anderson, R. B. & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*(5), 724–730. doi:10.3758/BF03211315

Arnold, K. E. & Pistilli, M. D. (2012). Course Signals at Purdue: Using Learning Analytics to Increase Student Success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267–270). LAK '12. New York, NY, USA: ACM. doi:10.1145/2330601.2330666

Atkinson, R. C. (1975). Mnemotechnics in second-language learning. *American psychologist*, *30*(8), 821.

Atkinson, R. C. & Hansen, D. N. (1966). Computer-assisted instruction in initial reading: the Stanford Project. *Reading Research Quarterly*, 5–25.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, *2*, 89–195.

Averell, L. & Heathcote, A. (2009). Long term implicit and explicit memory for briefly studied words. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin TX: Cognitive Science Society. ISBN* (pp. 978–).

Averell, L. & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*. Special Issue on Hierarchical Bayesian Models, *55*(1), 25–35. doi:10.1016/j.jmp.2010.08.009

Backhaus, J. & Junghanns, K. (2006). Daytime naps improve procedural motor memory. *Sleep Medicine*, *7*(6), 508–512. doi:10.1016/j.sleep.2006.04.002

Baddeley, A. D. & Scott, D. (1971). Short term forgetting in the absence of proactive interference. *Quarterly Journal of Experimental Psychology*, *23*(3), 275–283. doi:10.1080/14640746908401822

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, *104*(1), 54–75. doi:http://dx.doi.org/10.1037/0096-3445.104.1.54

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*(3), 296–308. doi:http://dx.doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General*, *113*(1), 1–29. doi:http://dx.doi.org/10.1037/0096-3445.113.1.1

Bahrick, H. P. (1992). Stabilized memory of unrehearsed knowledge. *Journal of Experimental Psychology: General*. 1st ser., *121*, 112–113.

Bahrick, H. P. & Hall, L. K. (1991). Lifetime maintenance of high school mathematics content. *Journal of Experimental Psychology: General*, *120*(1), 20–33. doi:http://dx.doi.org/10.1037/0096-3445.120.1.20

Barr, A., Beard, M., & Atkinson, R. C. (1975). A rationale and description of a CAI program to teach the BASIC programming language. *Instructional Science*, *4*(1), 1–31.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. doi:http://dx.doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., Paepe, A. D., & Langerock, N. (2012). Time causes forgetting from working memory. *Psychonomic Bulletin & Review*, *19*(1), 87–92. doi:10.3758/s13423-011-0192-8

Barrouillet, P., Portrat, S., Vergauwe, E., Diependaele, K., & Camos, V. (2011). Further evidence for temporal decay in working memory: Reply to Lewandowsky and Oberauer (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1302–1317. doi:http://dx.doi.org/10.1037/a0022933

Bean, C. H. (1912). *The curve of forgetting*. Press of the New era printing Company.

Begg, I. & Wickelgren, W. A. (1974). Retention functions for syntactic and lexical vs semantic information in sentence recognition memory. *Memory & Cognition*, *2*(2), 353–359.

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 317–333. doi:http://dx.doi.org/10.1037/a0014873

Bishara, A. J. & Lanzo, L. A. (2015). All of the above: When multiple correct response options enhance the testing effect. *Memory*, *23*(7), 1013–1028. doi:10.1080/09658211.2014.946425

Bjork, E. L. & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 56–64.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA, US: The MIT Press.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 4–16.

Blunt, J. R. & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. doi:http://dx.doi.org/10.1037/a0035934

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber.

Bregman, A. S. (1968). Forgetting curves with semantic, phonetic, graphic, and contiguity cues. *Journal of Experimental Psychology*, *78*(4, Pt.1), 539–546. doi:http://dx.doi.org/10.1037/h0026635

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. doi:http://dx.doi.org/10.1037/0033-295X.114.3.539

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21. doi:10.1080/17470215808416249

Burtt, H. E. & Dobell, E. M. (1925). The curve of forgetting for advertising material. *Journal of Applied Psychology*, *9*(1), 5–21. doi:http://dx.doi.org/10.1037/h0073966

Bush, R. R. & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological review*, *58*(6), 413.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. doi:http://dx.doi.org/10.1037/a0019902

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563.

Carpenter, S. K. (2012). Testing Enhances the Transfer of Learning. *Current Directions in Psychological Science*, *21*(5), 279–283. doi:10.1177/0963721412452728

Carpenter, S. K. & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474–478.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology*, *23*(6), 760–771.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448.

Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis–a general method for cognitive model evaluation and improvement. In *Intelligent tutoring systems* (pp. 164–175). Springer.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, *56*(4), 236–246.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning a temporal ridgeline of optimal retention. *Psychological science*, *19*(11), 1095–1102.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*(1), 49–57. doi:10.1080/09658210903405737

Chan, J. C. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*(2), 153–170.

Chan, J. C., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553.

Chechile, R. A. (2006). Memory hazard functions: A vehicle for theory development and test. *Psychological Review*, *113*(1), 31–56. doi:http://dx.doi.org/10.1037/0033-295X.113.1.31

Chessa, A. G. & Murre, J. M. J. (2007). A Neurocognitive Model of Advertisement Content and Brand Name Recall. *Marketing Science*, *26*(1), 130–141. doi:10.1287/mksc.1060.0212

Chessa, A. G. & Murre, J. M. (2002). A model of learning and forgetting, I: The forgetting curve. *Amsterdam: NeuroMod Technical Report*, 02–01.

Chuang, I. & Ho, A. D. (2016). *HarvardX and MITx: Four Years of Open Online Courses – Fall 2012-Summer 2016* (SSRN Scholarly Paper No. ID 2889436). Social Science Research Network. Rochester, NY.

Clement, B., Roy, D., Oudeyer, P.-Y., & Lopes, M. (2015). Multi-Armed Bandits for Intelligent Tutoring Systems. *JEDM - Journal of Educational Data Mining*, *7*(2), 20–48.

Conway, M. A., Cohen, G., & Stanhope, N. (1991). On the very long-term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *Journal of Experimental Psychology: General*, *120*(4), 395–409. doi:http://dx.doi.org/10.1037/0096-3445.120.4.395

Conway, M. A., Cohen, G., & Stanhope, N. (1992). Why is it that university grades do not predict very-long-term retention?

Conway, M. A., Rubin, D. C., Collins, A. F., Gathercole, S. E., Conway, M. A., & Morris, P. E. (1993). The structure of autobiographical memory. *Theories of memory*, 103–137.

Coursera. (2015). Coursera Impact Revealed: Learner Outcomes in Open Online Courses. Education. Retrieved January 22, 2017, from http://www.slideshare.net/Coursera/coursera-impact-revealed-learner-outcomes-in-open-online-courses

Cowan, N., Elliott, E. M., Saults, J. S., Nugent, L. D., Bomb, P., & Hismjatullina, A. (2006). Rethinking Speed Theories of Cognitive Development. *Psychological Science*.

Crawford, V. M., Schlager, M., Penuel, W. R., & Toyama, Y. (2008). Supporting the art of teaching in a data-rich, high performance learning environment. *Linking data and learning*, 109–129.

Crovitz, H. F. & Schiffman, H. (1974). Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society*, *4*(5), 517–518. doi:10.3758/BF03334277

Cuddy, L. J. & Jacoby, L. L. (1982). When forgetting helps memory: an analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *21*(4), 451–467. doi:10.1016/S0022-5371(82)90727-7

Custers, E. J. F. M. (2010). Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education*, *15*(1), 109–128. doi:10.1007/s10459-008-9101-y

Custers, E. J. F. M. & ten Cate, O. T. J. (2011). Very long-term retention of basic science knowledge in doctors after graduation. *Medical Education*, *45*(4), 422–430. doi:10.1111/j.1365-2923.2010.03889.x

Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in a computer simulation: Worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior*, *23*(4), 1809–1819. doi:10.1016/j.chb.2005.11.001

Darley, C. F. & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*(1), 66–73. doi:http://dx.doi.org/10.1037/h0031836

Déry, N., Goldstein, A., & Becker, S. (2015). A role for adult hippocampal neurogenesis at multiple time scales: A study of recent and remote memory in humans. *Behavioral Neuroscience*, *129*(4), 435–449. doi:http://dx.doi.org/10.1037/bne0000073

Desmarais, M. C. (2012). Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, *13*(2), 30–36.

Desmarais, M. C. & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In *Artificial Intelligence in Education* (pp. 441–450). Springer.

Duke, R. A. & Davis, C. M. (2006). Procedural memory consolidation in the performance of brief keyboard sequences. *Journal of Research in Music Education*, *54*(2), 111–124.

Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers. *Educational Technology & Society*, *15*(3), 58–76.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Teachers College, Columbia University.

Ebner, M. & Schön, M. (2013). Why Learning Analytics in Primary Education Matters. *Bulletin of the Technical Committee on Learning Technology, Karagiannidis, C. & Graf, S (Ed.)* *15*(2), 14–17.

Eglington, L. G. & Kang, S. H. K. (2016). Retrieval Practice Benefits Deductive Inference. *Educational Psychology Review*, 1–14. doi:10.1007/s10648-016-9386-y

Ellis, J. A., Semb, G. B., & Cole, B. (1998). Very Long-Term Memory for Information Taught in School. *Contemporary Educational Psychology*, *23*(4), 419–433. doi:10.1006/ceps.1997.0976

Elmore, R. (1996). Getting to Scale with Good Educational Practice. *Harvard Educational Review*, *66*(1), 1–27. doi:10.17763/haer.66.1.g73266758j348t33

Endres, T. & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. *Educational Psychology*, 1054. doi:10.3389/fpsyg.2015.01054

Fajnsztejn-Pollack, G. (1973). A developmental study of decay rate in long-term memory. *Journal of Experimental Child Psychology*, *16*(2), 225–235. doi:10.1016/0022-0965(73)90163-X

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, *97*(2), 201.

Faus, M. (2014). Khan Academy Mastery Mechanics. Retrieved September 2, 2015, from http://mattfaus.com/2014/07/khan-academy-mastery-mechanics/

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, *19*(3), 243–266. doi:10.1007/s11257-009-9063-7

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5), 304–317.

Finkenbinder, E. O. (1913). The Curve of Forgetting. *The American Journal of Psychology*, *24*(1), 8–32. doi:10.2307/1413271

Fioravanti, M. & Di Cesare, F. (1992). Forgetting curves in long-term memory: Evidence for a multistage model of retention. *Brain and Cognition*, *18*(2), 116–124. doi:10.1016/0278-2626(92)90073-U

Fletcher-Flinn, C. M. & Gravatt, B. (1995). The efficacy of computer assisted instruction (CAI): A meta-analysis. *Journal of educational computing research*, *12*(3), 219–241.

Frankland, P. W. & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews Neuroscience*, *6*(2), 119–130. doi:10.1038/nrn1607

Frankland, P. W., Köhler, S., & Josselyn, S. A. (2013). Hippocampal neurogenesis and forgetting. *Trends in Neurosciences*, *36*(9), 497–503. doi:10.1016/j.tins.2013.05.002

Gais, S. & Born, J. (2004). Declarative memory consolidation: mechanisms acting during human sleep. *Learning & Memory*, *11*(6), 679–685.

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory*, *13*(3), 259–262. doi:10.1101/lm.132106

Gates, A. I. (1917). *Recitation as a factor in memorizing* (PhD Dissertation, Columbia University).

Gehring, R. E., Toglia, M. P., & Kimble, G. A. (1976). Recognition memory for words and pictures at short and long retention intervals. *Memory & Cognition*, *4*(3), 256–260. doi:10.3758/BF03213172

Glasnapp, D. R., Poggio, J. P., & Ory, J. C. (1978). End-of-course and long-term retention outcomes for mastery and nonmastery learning paradigms. *Psychology in the Schools*, *15*(4), 595–603. doi:10.1002/1520-6807(197810)15:4⟨595::AID-PITS2310150426⟩3.0.CO;2-4

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15647–15654. doi:10.1073/pnas.1014269108

Glover, J. A. (1989). The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392.

Glover, J. A. & Krug, D. (1990). The 'testing' effect and restricted retrieval rehearsal. *Psychological Record*, *40*(2), 215.

Gog, T. v., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing After Worked Example Study Does Not Enhance Delayed Problem-Solving Performance Compared to Restudy. *Educational Psychology Review*, *27*(2), 265–289. doi:10.1007/s10648-015-9297-3

Gog, T. v. & Sweller, J. (2015). Not New, but Nearly Forgotten: the Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases. *Educational Psychology Review*, *27*(2), 247–264. doi:10.1007/s10648-015-9310-x

Halamish, V. & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801–812. doi:http://dx.doi.org/10.1037/a0023219

Hinze, S. R. & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290–304. doi:10.1080/09658211.2011.560121

Hirasaki, K. (2014). How many users does Khan Academy have? - Quora. Retrieved September 21, 2015, from https://www.quora.com/How-many-users-does-Khan-Academy-have

Hu, D. (2011). How Khan Academy is using Machine Learning to Assess Student Mastery. Retrieved September 2, 2015, from http://david-hu.com/2011/11/02/how-khan-academy-is-using-machine-learning-to-assess-student-mastery.html

Jenkins, J. G. & Dallenbach, K. M. (1924). Obliviscence during Sleep and Waking. *The American Journal of Psychology*, *35*(4), 605–612. doi:10.2307/1414040

Johnson, C. I. & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621–629. doi:http://dx.doi.org/10.1037/a0015183

Jonge, M. d., Tabbers, H. K., & Rikers, R. M. J. P. (2015). The Effect of Testing on the Retention of Coherent and Incoherent Text Material. *Educational Psychology Review*, *27*(2), 305–315. doi:10.1007/s10648-015-9300-z

Kang, S. H. K., McDermott, K. B., & III, H. L. R. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528–558. doi:10.1080/09541440601056620

Kang, S. H., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic bulletin & review*, 1–7.

Karpicke, J. D. & Aue, W. R. (2015). The Testing Effect Is Alive and Well with Complex Materials. *Educational Psychology Review*, *27*(2), 317–326. doi:10.1007/s10648-015-9309-3

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*, *331*(6018), 772–775. doi:10.1126/science.1199327

Karpicke, J. D. & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162.

King, H. E. (1963a). The retention of sensory experience: I. Intensity. *The Journal of psychology*, *56*(2), 283–290.

King, H. E. (1963b). The retention of sensory experience: II. Frequency. *The Journal of psychology*, *56*(2), 291–298.

Koller, D. (2012). MOOCs on the Move: How Coursera Is Disrupting the Traditional Classroom. Retrieved from http://knowledge.wharton.upenn.edu/article.cfm?articleid=3109

Kornell, N. & Bjork, R. A. (2008). Learning concepts and categories is spacing the "enemy of induction"? *Psychological science*, *19*(6), 585–592.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. doi:10.1016/j.jml.2011.04.002

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and aging*, *25*(2), 498.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. doi:http://dx.doi.org/10.1037/a0037850

Krueger, W. C. F. (1929). The effect of overlearning on retention. *Journal of Experimental Psychology*, *12*(1), 71–78. doi:http://dx.doi.org/10.1037/h0072036

Kulik, C.-L. C. & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in human behavior*, *7*(1), 75–94.

Kuo, T.-M. & Hirshman, E. (1996). Investigations of the Testing Effect. *The American Journal of Psychology*, *109*(3), 451–464. doi:10.2307/1423016

Leahy, W., Hanham, J., & Sweller, J. (2015). High Element Interactivity Information During Problem Solving may Lead to Failure to Obtain the Testing Effect. *Educational Psychology Review*, *27*(2), 291–304. doi:10.1007/s10648-015-9296-4

Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology*, *48*(5), 310–321. doi:10.1016/j.jmp.2004.06.002

Lewandowsky, S., Ecker, U. K., Farrell, S., & Brown, G. D. (2012). Models of cognition and constraints from neuroscience: A case study involving consolidation. *Australian Journal of Psychology*, *64*(1), 37–45. doi:10.1111/j.1742-9536.2011.00042.x

Lewandowsky, S., Geiger, S. M., & Oberauer, K. (2008). Interference-based forgetting in verbal short-term memory. *Journal of Memory and Language*, *59*(2), 200–222. doi:10.1016/j.jml.2008.04.004

Lewandowsky, S., Nimmo, L. M., & Brown, G. D. A. (2008). When temporal isolation benefits memory for serial order. *Journal of Memory and Language*, *58*(2), 415–428. doi:10.1016/j.jml.2006.11.003

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences*, *13*(3), 120–126. doi:10.1016/j.tics.2008.12.003

Lewis, O., Lindsey, R., Pashler, H., & Mozer, M. (2010). Predicting Students' Retention of Facts from Feedback During Training. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological science*, *25*(3), 639–647.

Lindsey, R., Mozer, M. C., Cepeda, N. J., & Pashler, H. (2009). Optimizing memory retention with cognitive models. In *Proceedings of the Ninth International Conference on Cognitive Modeling (ICCM 2009)* (pp. 74–79).

Luh, C. W. (1922). The conditions of retention. *Psychological Monographs*, *31*(3), i–87. doi:http://dx.doi.org/10.1037/h0093177

Mandler, G. & Rabinowitz, J. C. (1981). Appearance and reality: Does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory*, *7*(2), 79–90. doi:http://dx.doi.org/10.1037/0278-7393.7.2.79

Marshall, L. & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in cognitive sciences*, *11*(10), 442–450.

Matsuda, N., Furukawa, T., Bier, N., & Faloutsos, C. (2015). Machine beats experts: Automatic discovery of skill models for data-driven online course refinement. In *Educational Data Mining 2015*.

Mayfield, K. H. & Chase, P. N. (2002). The Effects of Cumulative Practice on Mathematics Problem Solving. *Journal of Applied Behavior Analysis*, *35*(2), 105–123. doi:10.1901/jaba.2002.35-105

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. doi:http://dx.doi.org/10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494–513. doi:10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied*, *21*(4), 370–382. doi:http://dx.doi.org/10.1037/xap0000063

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The Read-Recite-Review Study Strategy: Effective and Portable. *Psychological Science*, *20*(4), 516–522.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in Middle-School Science: Successful Transfer Performance on Classroom Exams. *Applied Cognitive Psychology*, *27*(3), 360–372. doi:10.1002/acp.2914

McGaugh, J. L. (2000). Memory–a Century of Consolidation. *Science*, *287*(5451), 248–251. doi:10.1126/science.287.5451.248

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological review*, *39*(4), 352.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. doi:10.1007/BF02295996

Mednick, S. C., Cai, D. J., Kanady, J., & Drummond, S. P. A. (2008). Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory. *Behavioural Brain Research*, *193*(1), 79–86. doi:10.1016/j.bbr.2008.04.028

Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S., & Wixted, J. (2011). An opportunistic theory of cellular and systems consolidation. *Trends in neurosciences*, *34*(10), 504–514. doi:10.1016/j.tins.2011.06.003

Mednick, S. C., McDevitt, E. A., Walsh, J. K., Wamsley, E., Paulus, M., Kanady, J. C., & Drummond, S. P. A. (2013). The Critical Role of Sleep Spindles in Hippocampal-Dependent Memory: A Pharmacology Study. *Journal of Neuroscience*, *33*(10), 4494–4504. doi:10.1523/JNEUROSCI.3127-12.2013

Mednick, S., Nakayama, K., & Stickgold, R. (2003). Sleep-dependent learning: a nap is as good as a night. *Nature Neuroscience*, *6*(7), 697–698. doi:10.1038/nn1078

Milner, C. E., Fogel, S. M., & Cote, K. A. (2006). Habitual napping moderates motor performance improvements following a short daytime nap. *Biological Psychology*, *73*(2), 141–156. doi:10.1016/j.biopsycho.2006.01.015

Minke, K., A. & Stalling, R., B. (1970). *Long-Term Retention of Conditioned Attitudes.* (Office of Naval Research No. TR-6). Department of Psychology, Hawaii University. Honolulu.

Morin, C., Brown, G. D. A., & Lewandowsky, S. (2010). Temporal isolation effects in recognition and serial recall. *Memory & Cognition*, *38*(7), 849–859. doi:10.3758/MC.38.7.849

Murdock, B. B. (1961). The retention of individual items. *Journal of Experimental Psychology*, *62*(6), 618–625. doi:http://dx.doi.org/10.1037/h0043657

Murre, J. M. J. & Chessa, A. G. (2011). Power laws from individual differences in learning and forgetting: mathematical analyses. *Psychonomic Bulletin & Review*, *18*(3), 592–597. doi:10.3758/s13423-011-0076-y

Murre, J. M. J., Chessa, A. G., & Meeter, M. (2013). A Mathematical Model of Forgetting and Amnesia. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00076

Murre, J. M. J. & Dros, J. (2015). Replication and Analysis of Ebbinghaus' Forgetting Curve. *PLOS ONE*, *10*(7), e0120644. doi:10.1371/journal.pone.0120644

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*(5), 832–840. doi:10.3758/BF03198418

Neath, I. & Brown, G. D. A. (2012). Arguments Against Memory Trace Decay: A SIMPLE Account of Baddeley and Scott. *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00035

Nelson, T. O., Shimamura, A. P., & Leonesio, R. J. (1980). Large effects on long-term retention after standard list learning vs. adjusted learning. *Behavior Research Methods & Instrumentation*, *12*(1), 42–44.

Nishida, M. & Walker, M. P. (2007). Daytime Naps, Motor Memory Consolidation and Regionally Specific Sleep Spindles. *PLOS ONE*, *2*(4), e341. doi:10.1371/journal.pone.0000341

Norris, D., Baer, L., Leonard, J., Pugliese, L., & Lefrere, P. (2008). Action Analytics: Measuring and Improving Performance that Matters in Higher Education. *Educause Review*, *43*(1), 42.

Oberauer, K. & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*, *115*(3), 544–576. doi:http://dx.doi.org/10.1037/0033-295X.115.3.544

Oberauer, K. & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General*, *142*(2), 380–411. doi:http://dx.doi.org/10.1037/a0029588

Oberauer, K. & Lewandowsky, S. (2014). Further evidence against decay in working memory. *Journal of Memory and Language*, *73*, 15–30. doi:10.1016/j.jml.2014.02.003

ALEKS – Assessment and Learning, K-12, Higher Education, Automated Tutor, Math. (n.d.). Retrieved January 28, 2017, from https://www.aleks.com/

Common Core State Standards Initiative — The Standards — Mathematics. (n.d.). Retrieved March 20, 2012, from http://www.corestandards.org/the-standards/mathematics

einprägen - Wiktionary. (n.d.). Retrieved October 25, 2016, from https://en.wiktionary.org/wiki/einpr%C3%83%C2%A4gen

Home - Carnegie Learning. (n.d.). Retrieved January 28, 2017, from http://www.carnegielearning.com/

Pearson Success Net. (n.d.). Retrieved March 20, 2012, from https://www.pearsonsuccessnet.com

The Khan Academy. (n.d.). Retrieved March 20, 2012, from http://www.khanacademy.org

Table Curve 2D. (1994). San Rafael, CA: Jandel Scientific.

Making Data Work for Teachers and Students. (2015). Bill & Melinda Gates Foundation.

Udacity, Coursera and edX Now Claim Over 24 Million Students (EdSurge News). (2015). Retrieved September 21, 2015, from https://www.edsurge.com/news/2015-09-08-udacity-coursera-and-edx-now-claim-over-24-million-students

Apply for Financial Aid. (2017). Retrieved January 23, 2017, from http://learner.coursera.help/hc/en-us/articles/209819033-Apply-for-Financial-Aid

Osterrieth, P. (1944). Filetest de copie d'une figure complex: Contribution a l'etude de la perception et de la memoire. *Archives de Psychologie*, *30*, 286–356.

Ostrow, K., Heffernan, N., Heffernan, C., & Peterson, Z. (2015). Blocking Vs. Interleaving: Examining Single-Session Effects Within Middle School Math Homework. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (9112, pp. 338–347). Lecture Notes in Computer Science. DOI: 10.1007/978-3-319-19773-9_34. Springer International Publishing.

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology*, *108*(4), 563–575. doi:http://dx.doi.org/10.1037/edu0000074

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language*, *83*, 53–61. doi:10.1016/j.jml.2015.04.001

Pan, S. C. & Rickard, T. C. (2015). Sleep and motor learning: is there room for consolidation? *Psychological bulletin*, *141*(4), 812.

Pashler, H., Cepeda, N., Lindsey, R., Vul, E., & Mozer, M. C. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In *Advances in neural information processing systems* (pp. 1321–1329).

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic bulletin & review*, *14*(2), 187–193.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*(302), 157–175. doi:10.1080/14786440009463897

Peterson, L. & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198. doi:http://dx.doi.org/10.1037/h0049234

Pierce, B. H. & Hawthorne, M. J. (2016). Does the testing effect depend on presentation modality? *Journal of Applied Research in Memory and Cognition*, *5*(1), 52–58. doi:10.1016/j.jarmac.2016.01.001

Portrat, S., Barrouillet, P., & Camos, V. (2008). Time-related decay or interference-based forgetting in working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1561–1564. doi:http://dx.doi.org/10.1037/a0013356

Pyc, M. A. & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. doi:10.1016/j.jml.2009.01.004

Pyc, M. A. & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335–335.

Radosavljevich, P. R. (1907). *Das Behalten und Vergessen bei Kindern und Erwachsenen nach experimentellen Untersuchungen: (Das Fortschreiten des Vergessens mit der Zeit)*. Google-Books-ID: V8CgAAAAMAAJ. O. Nemnich.

Raffel, G. (1934). The effect of recall on forgetting. *Journal of Experimental Psychology*, *17*(6), 828–838. doi:http://dx.doi.org/10.1037/h0075297

Rangel, L. M., Alexander, A. S., Aimone, J. B., Wiles, J., Gage, F. H., Chiba, A. A., & Quinn, L. K. (2014). Temporally selective contextual encoding in the dentate gyrus of the hippocampus. *Nature Communications*, *5*, 3181. doi:10.1038/ncomms4181

Rawson, K. A. (2015). The Status of the Testing Effect for Complex Materials: Still a Winner. *Educational Psychology Review*, *27*(2), 327–331. doi:10.1007/s10648-015-9308-4

Rawson, K. A. & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302. doi:10.1037/a0023956

Rawson, K. A. & Dunlosky, J. (2012). When Is Practice Testing Most Effective for Improving the Durability and Efficiency of Student Learning? *Educational Psychology Review*, *24*(3), 419–435. doi:10.1007/s10648-012-9203-1

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer New York.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Rickard, T. C., Lau, J. S.-H., & Pashler, H. (2008). Spacing and the transition from calculation to retrieval. *Psychonomic Bulletin & Review*, *15*(3), 656–661.

Ridgeway, K., Mozer, M. C., Bowles, A., & Stone, R. (2016). Forgetting of foreign-language skills: A corpus-based analysis of online tutoring software. *Cognitive Science Journal.(Accepted for publication)*.

Rieth, C. A., Cai, D. J., McDevitt, E. A., & Mednick, S. C. (2010). The role of sleep and practice in implicit and explicit motor learning. *Behavioural Brain Research*, *214*(2), 470–474. doi:10.1016/j.bbr.2010.05.052

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological review*, *107*(2), 358.

Robertson, E. M. (2004). Skill Learning: Putting Procedural Consolidation in Context. *Current Biology*, *14*(24), R1061–R1063. doi:10.1016/j.cub.2004.11.048

Roediger, H. L. & DeSoto, K. A. (2014). Forgetting the presidents. *Science*, *346*(6213), 1106–1109. doi:10.1126/science.1259627

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. doi:http://dx.doi.org/10.1037/a0026252

Roediger, H. L. & Karpicke, J. D. (2006a). Test-enhanced learning taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249–255.

Roediger, H. L. & Karpicke, J. D. (2006b). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*(5), 1323–1330. doi:10.3758/s13423-014-0588-3

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900–908. doi:http://dx.doi.org/10.1037/edu0000001

Rohrer, D. & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*(4), 183–186.

Rohrer, D. & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*(5), 406–412.

Rohrer, D. & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*(9), 1209–1224. doi:10.1002/acp.1266

Rohrer, D. & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239. doi:http://dx.doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. doi:http://dx.doi.org/10.1037/a0037559

Rowland, C. A. & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, *23*(3), 403–419. doi:10.1080/09658211.2014.889710

Rubin, D. C. (1982). On the retention function for autobiographical memory. *Journal of Verbal Learning and Verbal Behavior*, *21*(1), 21–38.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1161–1176. doi:http://dx.doi.org/10.1037/0278-7393.25.5.1161

Rubin, D. C. & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760. doi:http://dx.doi.org/10.1037/0033-295X.103.4.734

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, *11*(6), 641–650. doi:10.3758/BF03198289

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2014). How we forget may depend on how we remember. *Trends in Cognitive Sciences*, *18*(1), 26–36. doi:10.1016/j.tics.2013.10.008

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2016). Forgetting Patterns Differentiate Between Two Forms of Memory Representation. *Psychological Science*, *27*(6), 810–820. doi:10.1177/0956797616638307

Schabus, M., Gruber, G., Parapatics, S., Sauter, C., Klosch, G., Anderer, P., ... Zeitlhofer, J. (2004). Sleep spindles and their significance for declarative memory consolidation. *Sleep*, *27*(8), 1479–1485.

Semb, G. B., Ellis, J. A., & Araujo, J. (1993). Long-term memory for knowledge learned in school. *Journal of Educational Psychology*, *85*(2), 305–316. doi:http://dx.doi.org/10.1037/0022-0663.85.2.305

Shah, D. (2016). Monetization Over Massiveness: Breaking Down MOOCs by the Numbers in 2016 (EdSurge News). Retrieved January 22, 2017, from https://www.edsurge.com/news/2016-12-29-monetization-over-massiveness-breaking-down-moocs-by-the-numbers-in-2016

Siemens, G. (2012). Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 4–8). ACM.

Sloman, S. A., Hayman, C. A., Ohta, N., Law, J., & Tulving, E. (1988). Forgetting in primed fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 223.

Smith, M. A., Blunt, J. R., Whiffen, J. W., & Karpicke, J. D. (2016). Does Providing Prompts During Retrieval Practice Improve Learning? *Applied Cognitive Psychology*, *30*(4), 544–553. doi:10.1002/acp.3227

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767.

Southey, R. (1837). *The story of the three bears*. Google-Books-ID: 5JUNAAAAQAAJ.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641.

Squire, L. R. (1989). On the course of forgetting in very long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 241–245. doi:http://dx.doi.org/10.1037/0278-7393.15.2.241

Stafford, T. & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological science*, *25*(2), 511–518.

Stamper, J. C. & Koedinger, K. R. (2011). Human-machine student model discovery and improvement using DataShop. In *Artificial Intelligence in Education* (pp. 353–360). Springer.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, *437*(7063), 1272–1278. doi:10.1038/nature04286

Strong, E. K. (1913). The effect of time-interval upon recognition memory. *Psychological Review*, *20*(5), 339–372. doi:http://dx.doi.org/10.1037/h0072087

Sutterer, D. W. & Awh, E. (2015). Retrieval practice enhances the accessibility but not the quality of memory. *Psychonomic Bulletin & Review*, *23*(3), 831–841. doi:10.3758/s13423-015-0937-x

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press Cambridge.

Taylor, K. & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848. doi:10.1002/acp.1598

Taylor, R. (Ed.). (1980). *The Computer in the School: Tutor, Tool, Tutee*. Teachers College Press New York.

Tibbles, R. (2015). Exploring the Impact of Spacing in Mathematics Learning through Data Mining. In *Proceedings of the 8th International Conference on Educational Data Mining*. Madrid, Spain: Educational Data Mining Society.

Tibbles, R. & Alexandre, J. (2014). KA Lite: Implementation and Research in the Offline Learning Revolution. In *Proceedings of MOOCs4d: Potential at the Bottom of the Pyramid*. University of Pennsylvania, Philadelphia, USA.

Tran, R., Rohrer, D., & Pashler, H. (2014). Retrieval practice: the lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*, *22*(1), 135–140. doi:10.3758/s13423-014-0646-x

Trumbo, M. C., Leiting, K. A., McDaniel, M. A., & Hodge, G. K. (2016). Effects of reinforcement on test-enhanced learning in a large, diverse introductory college psychology course. *Journal of Experimental Psychology: Applied*, *22*(2), 148–160. doi:http://dx.doi.org/10.1037/xap0000082

Tucker, M. A., Hirota, Y., Wamsley, E. J., Lau, H., Chaklader, A., & Fishbein, W. (2006). A daytime nap containing solely non-REM sleep enhances declarative but not procedural memory. *Neurobiology of learning and memory*, *86*(2), 241–247.

Turner, T. E., Macasek, M. A., Nuzzo-Jones, G., Heffernan, N. T., & Koedinger, K. R. (2005). The Assistment Builder: A Rapid Development Tool for ITS. In *AIED* (pp. 929–931).

Turvey, M. T. & Weeks, R. A. (1975). Effects of proactive interference and rehearsal on the primary and secondary components of short-term retention. *Quarterly Journal of Experimental Psychology*, *27*(1), 47–62. doi:10.1080/14640747508400463

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

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Assche, F. V., Parra, G., & Klerkx, J. (2014). Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, *18*(6), 1499–1514. doi:10.1007/s00779-013-0751-2

Vertes, R. P. (2004). Memory Consolidation in Sleep: Dream or Reality. *Neuron*, *44*(1), 135–148. doi:10.1016/j.neuron.2004.08.034

Volante, L. (2004). Teaching to the Test: What Every Educator and Policy-Maker Should Know. *Canadian Journal of Educational Administration and Policy*.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental process*. Cambridge, MA: Harvard University Press.

Walker, M. P., Brakefield, T., Allan Hobson, J., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, *425*(6958), 616–620. doi:10.1038/nature01930

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with Sleep Makes Perfect: Sleep-Dependent Motor Skill Learning. *Neuron*, *35*(1), 205–211. doi:10.1016/S0896-6273(02)00746-8

Waugh, N. C. & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*(2), 89–104. doi:http://dx.doi.org/10.1037/h0021797

Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, *16*(3), 308–316. doi:http://dx.doi.org/10.1037/a0020992

Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied*, *22*(1), 72–84. doi:http://dx.doi.org/10.1037/xap0000071

White, K. G. (2001). Forgetting functions. *Animal Learning & Behavior*, *29*(3), 193–207. doi:10.3758/BF03192887

White, K. G. (2012). Dissociation of short-term forgetting from the passage of time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 255–259. doi:http://dx.doi.org/10.1037/a0025197

Whitten, W. B. & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*(4), 465–478. doi:10.1016/S0022-5371(77)80040-6

Wickelgren, W. A. (1968). Sparing of short-term memory in an amnesic patient: Implications for strength theory of memory. *Neuropsychologia*, *6*(3), 235–244. doi:10.1016/0028-3932(68)90022-5

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology*, *9*(4), 418–455. doi:10.1016/0022-2496(72)90015-6

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition*, *2*(4), 775–780.

Wickelgren, W. A. (1975a). Age and storage dynamics in continuous recognition memory. *Developmental Psychology*, *11*(2), 165–169. doi:http://dx.doi.org/10.1037/h0076457

Wickelgren, W. A. (1975b). Alcoholic intoxication and memory storage dynamics. *Memory & Cognition*, *3*(4), 385–389.

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, *105*(2), 379–386. doi:http://dx.doi.org/10.1037/0033-295X.105.2.379

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*(1), 10–16. doi:10.1111/sjop.12093

Wilhelm, I., Diekelmann, S., & Born, J. (2008). Sleep in children improves memory performance on declarative but not procedural tasks. *Learning & Memory*, *15*(5), 373–377. doi:10.1101/lm.803708

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140–1147. doi:10.3758/s13423-011-0140-7

Wixted, J. T. (2004a). On Common Ground: Jost's (1897) Law of Forgetting and Ribot's (1881) Law of Retrograde Amnesia. *Psychological Review*, *111*(4), 864–879. doi:http://dx.doi.org/10.1037/0033-295X.111.4.864

Wixted, J. T. (2004b). The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology*, *55*(1), 235–269. doi:10.1146/annurev.psych.55.090902.141555

Wixted, J. T. & Ebbesen, E. B. (1991). On the Form of Forgetting. *Psychological Science*, *2*(6), 409–415. doi:10.1111/j.1467-9280.1991.tb00175.x

Wixted, J. T. & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, *25*(5), 731–739. doi:10.3758/BF03211316

Yamashita, H. (2016). Effects of Immediate Recall Trial on One-Year Delayed Recall Performance in Rey Complex Figure Test. *Applied Neuropsychology: Adult*, 1–6. doi:10.1080/23279095.2015.1135441

Zhenghao, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D., & Emanuel, E. J. (2015). Who's Benefiting from MOOCs, and Why. Retrieved October 13, 2016, from https://hbr.org/2015/09/whos-benefiting-from-moocs-and-why