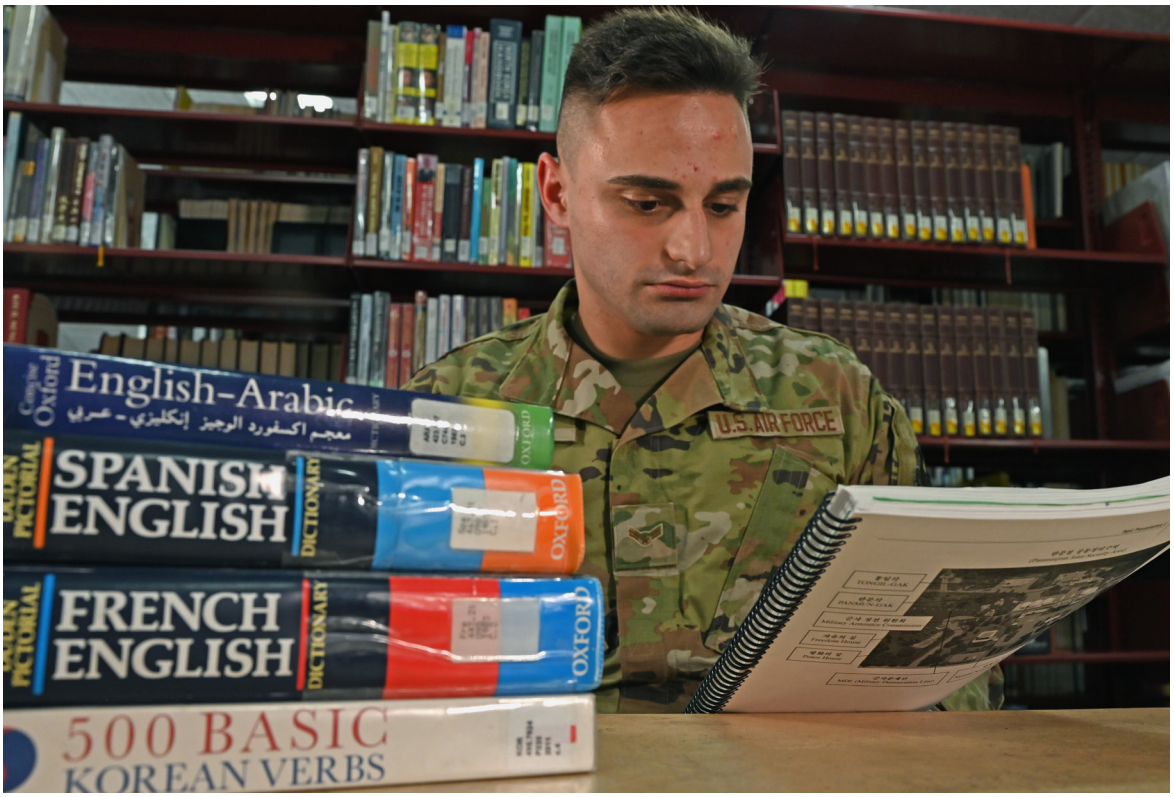
RAND CORPORATION

Research Report

MATTHEW WALSH, MARK TOUKAN, THOMAS EDWARD GOODE, ANDREA M. ABLER, SEAN MANN, LEWIS SCHNEIDER

Exploring the Use of Computational Cognitive Models to Personalize Training



For more information on this publication, visit **www.rand.org/t/RRA1565-1**.

About RAND

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2023 RAND Corporation
RAND® is a registered trademark.

Library of Congress Cataloging-in-Publication Data is available for this publication.

ISBN: 978-1-9774-1160-0

Cover: U.S. Air Force photo by Senior Airman Michael Bowman.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, please visit www.rand.org/pubs/permissions.

About This Report

The purpose of training and education in the United States Department of the Air Force (DAF) is to develop and sustain mission-critical knowledge, skills, and abilities (KSAs) among airmen, guardians, and civilians. The DAF must deliver effective training and education to fully use its human capital, provide warfighting assets to combatant commanders, and maintain asymmetric advantage over competitors. Yet training and education is costly. A recent budget request included more than $2 billion for training and education, and recent guidance has highlighted that the United States Air Force (USAF) must transform all facets of training and education. By doing so, the DAF can field a highly capable force in an affordable manner.

This report focuses on computational cognitive models, a class of training technologies with transformative potential. Computational cognitive models emulate psychological processes like knowledge acquisition and retention. These models have been used to develop empirically grounded training curricula and deliver personalized training in diverse domains such as math education, language learning, and medicine. The primary benefits of using these models to deliver personalized training are enhanced learning gains and reduced training time.

This report describes the feasibility of applying computational cognitive models to the acquisition and sustainment of mission-critical KSAs, with emphasis on second-language learning. We find that cognitive models can be integrated with training curricula in a variety of ways, and that each of these potential courses of action presents different levels of benefits along with different technical and logistical challenges.

The research reported here was commissioned by the Air Force Research Laboratory (AFRL) and conducted within the Workforce, Development, and Health Program of RAND Project AIR FORCE as part of an FY 2021 project "Exploring the Use of Computational Cognitive Models to Enable Personalized Training."

RAND Project AIR FORCE

RAND Project AIR FORCE (PAF), a division of the RAND Corporation, is the Department of the Air Force's (DAF's) federally funded research and development center for studies and analyses, supporting both the United States Air Force and the United States Space Force. PAF provides the DAF with independent analyses of policy alternatives affecting the development, employment, combat readiness, and support of current and future air, space, and cyber forces. Research is conducted in four programs: Strategy and Doctrine; Force Modernization and Employment; Resource Management; and Workforce, Development, and Health. The research reported here was prepared under contract FA7014-16-D-1000.

Additional information about PAF is available on our website: www.rand.org/paf/

This report documents work originally shared with the DAF on September 29, 2022. The draft report, dated February 2023, was reviewed by formal peer reviewers and DAF subject-matter experts.

Acknowledgments

We thank Timothy Sakulich, Executive Director of the Air Force Research Laboratory (AFRL). We thank Tiffany Jastrzembski at Air Force Research Laboratory, Airman Systems Directorate (AFRL/RH). This research benefited greatly from her input and support. Additionally, we are grateful to Julie Cantwell of the Defense Language Institute (DLI) for sharing her insights about linguist training, and to MSgt Jason Jones.

We thank Jair Aguirre and Philip Pavlik, Jr., for their constructive comments on an earlier version of this report. Finally, we thank the many RAND colleagues who helped with this work, principally—but not exclusively—Nelson Lim, Miriam Matthews, Melissa Baumann, and Claude Setodji.

Summary

Issue

- Training in the United States Department of the Air Force (DAF) is essential to develop, upskill, and sustain the abilities of airmen, guardians, and civilians.
- The DAF might leverage training and education technologies developed in industry, academia, and the government to maximize training effectiveness.
- Computational cognitive models of knowledge acquisition and retention are one potentially transformative training technology.
- These models can be used to develop empirically grounded curricula and to deliver personalized training.
- The DAF should develop and evaluate multiple courses of action (COAs) to find high-leverage applications of computational cognitive models to knowledge acquisition and retention.
- Second-language learning is one such application.

Approach

To evaluate the feasibility of applying cognitive models of knowledge acquisition and retention to DAF training and education, we conducted a market analysis and literature review of adaptive training technologies being developed by industry, academia, and the government. We then validated a computational cognitive model of knowledge acquisition and retention using a large dataset on second-language learning gathered in a naturalistic setting. Next, we developed statistical methods to calibrate the model to individual students and to groups of students. Finally, we conducted a detailed analysis of the goals, data-capture capabilities, and information technology system affordances of the Air Education and Training Command (AETC) Linguist Next (LN) Modern Standard Arabic Basic course. The synthesis of our analyses revealed that a specific computational cognitive model—the Predictive Performance Equation (PPE)—can be applied to the acquisition and sustainment of mission-critical knowledge. We demonstrate multiple COAs for applying PPE using the detailed case study of the LN Standard Arabic Basic course.

Key Findings

- Adaptive training approaches that use computational cognitive models are relatively underexplored, yet they offer more benefits than other approaches.
- Of the computational cognitive models that have been proposed, the PPE has been demonstrated in many domains and settings.

- Building on previous work that has demonstrated PPE in real-world settings, we find that it also provides a valid account of second-language learning in representative settings.
- Of the elements covered in the LN Standard Arabic Basic course, PPE is most applicable to the acquisition and retention of task-critical vocabulary.
- PPE can be used to deliver empirically grounded recommendations for when to introduce and rehearse task-critical vocabulary in the LN curriculum at the classroom level.
- Student performance measures captured in the LN Standard Arabic Basic course provide information about mastery, yet they are not intended for personalized training.
- One promising COA is to deliver rehearsal activities and assess mastery of task-critical vocabulary using a separate software application designed for cognitive model–based personalized training.

Recommendations

We recommend the following:

- *The DAF should leverage cognitively inspired technologies to augment training.* These technologies are based on decades of learning-science research and have been shown to yield measurable benefits in terms of level of mastery and speed of learning.
- *The DAF should use computational cognitive models of knowledge acquisition and retention to deliver empirically based, personalized training.* Cognitive models are underexplored relative to other approaches to delivering adaptive training, yet they might be more suitable and might offer significant benefits beyond those methods. Furthermore, cognitive models provide a way to trace knowledge retention across long periods and during periods of disuse.
- *Air Force Research Laboratory (AFRL) should work with the DAF training enterprise to identify knowledge, skills, and abilities (KSAs) that underlie task performance to permit the application of cognitive models.* Furthermore, AFRL should work with the DAF training enterprise to tag training materials based on the KSAs they involve and to develop quantitative measures of KSA performance. This level of analysis is needed for cognitive model–based, personalized training.
- *AFRL should continue to develop statistical methods to calibrate cognitive models efficiently and effectively for individual students and groups of students.* Different implementations of the method used here, Bayesian hierarchical models, are needed to scale methods to larger numbers of students.
- *AETC LN should use a separate software application specifically designed for personalized training to deliver rehearsal activities of task-critical vocabulary.* The LN curriculum must satisfy several objectives. Thus, learning events cannot be fully tailored for cognitive model–based, personalized training. A promising alternative is to track all learning activities that students complete but to limit personalized training to a separate application that complements the full curriculum.

Contents

About This Report iii
Summary v
Figures and Tables viii
Chapter 1. Introduction 1
Study Context 3
Study Methodology 10
Organization of the Report 12
Chapter 2. Trends in Adaptive Training 14
Market Analysis 15
Literature Review 19
Summary 28
Chapter 3. Validating a Cognitive Model of Knowledge Acquisition 29
Descriptive Analyses of Second Language Acquisition Modeling 29
Application of the Predictive Performance Equation to Student Performance 33
Summary 36
Chapter 4. Case Study Using Arabic Curriculum in Universal Curriculum Assessment Tool 37
Capability 1. Validated Cognitive Model 38
Capability 2. Extract, Process, and Evaluate Curriculum Data in Universal Curriculum and Assessment Tool Modules 38
Capability 3. Extract Student Data in the Universal Curriculum and Assessment Tool to Track Proficiency over Time 41
Capability 4. Recommend Student Activities Using the Model 44
Capability 5. Deliver Activities to Students or Instructors 47
Summary 48
Chapter 5. Conclusion 50
Key Findings 50
Recommendations 51
Appendix. Technical Description of the Predictive Performance Equation and Ancillary Analysis 53
Abbreviations 58
References 60

Figures and Tables

Figures

Figure 1.1. Training System Designs Formed by the Intersection of Primary Dimensions 7
Figure 1.2. Expanded System Design Space Created by Including Cognitive Models 8
Figure 1.3. Adaptive Training Cycle 9
Figure 1.4. Study Approach 11
Figure 2.1. Data Sources 15
Figure 2.2. Approaches to Adaptive Training (Graph 1) and Training Medium (Graph 2) 17
Figure 2.3. Families of Cognitive Models of Memory 22
Figure 2.4. Cognitive Models and Tools by Study Domain and Study Setting 25
Figure 2.5. Families of Educational Data Mining Approaches 27
Figure 3.1. Second Language Acquisition Modeling Learning Curves by Question Type and Language 31
Figure 3.2. Second Language Acquisition Modeling Forgetting Curves by Question Type and Language 32
Figure 3.3. Retention Predictions for Eight Individuals from the Second Language Acquisition Modeling Dataset 34
Figure 3.4. Comparison Bayesian Hierarchical and Nonhierarchical Models 35
Figure 4.1. Percentage of Remaining Modules That Vocabulary Words Appear in, by Module of First Appearance 40
Figure 4.2. Count of Activity Types in Linguist Next Standard Arabic Basic Course Modules 42
Figure 4.3. Count of Activity Types in Linguist Next Standard Arabic Basic Assessment Modules 43
Figure 4.4. Predicted Proficiency for an Example Word in the Linguist Next Standard Arabic Basic Course 45
Figure A.1. Cognitive Model Expressed in Graphical Notation 56

Tables

Table 1.1. Opportunities for Personalized Training in United States Department of the Air Force Occupations 10
Table 2.1. Number and Size of Investments in Adaptive Training Technology Systems 16
Table 2.2. Classes of Approaches to Adaptive Training 18
Table 2.3. Article Counts by Application Domain, Setting, and Population 20

Table 2.4. Model- and Theory-Based Tools....24
Table 2.5. Opportunities and Barriers for Cognitive Model–Based Adaptive Training....26
Table 5.1. Summary of Key Results....51
Table A.1. Negative Log-Likelihoods for Models That Allow Parameters to Vary by Question Type and Language....55
Table A.2. Odds Ratios and Confidence Intervals from Logistic Regression Analysis of Second Language Acquisition Modeling....57

Chapter 1. Introduction

The objective of training and education in the United States Department of the Air Force (DAF) is to give airmen, guardians, and civilians the occupational and institutional competencies needed to support DAF warfighting functions.[1] By delivering effective training and education, the DAF can more effectively use its human capital. More to the point, the DAF's success in training and education undergirds its ability to provide warfighting assets to combatant commanders and to maintain asymmetric advantage over competitors. New technologies that accelerate knowledge acquisition and retention can have an outsized impact on maintaining this advantage.

The training and education that is vital to DAF mission success is costly. The DAF fiscal year (FY) 2023 Operations and Maintenance (O&M) Budget request includes $2.16 billion for training and education.[2] This does not capture the full cost of training and education, which also includes personnel and procurement expenses contained across other budget activities. In his call to accelerate change in multiple areas—including training and education—Chief of Staff of the Air Force (CSAF) General Charles Brown wrote that we must "transform the way we learn across all facets of USAF [United States Air Force] education and training curricula."[3] By doing so, the DAF can field a highly capable workforce in an economical manner.

The DAF has taken steps to modernize training and education. For example:

- The Air Force Research Laboratory, Airman Systems Directorate (AFRL/RH) develops human-centered technologies, including technologies to deliver training and assess human performance.
- The Air Education and Training Command Analysis and Innovation Directorate (AETC/A9) explores opportunities to integrate technologies into the DAF learning environment.
- The Technology Directorate (AFWERX) of the Air Force Research Laboratory's (AFRL) partners with academia, industry, and government to develop and transition dual-use capabilities, including training capabilities.
- Numerous smaller-scale efforts are underway to leverage educational technologies in the DAF.

[1] Air Force Instruction 36-2201, *Air Force Training Program*, Secretary of the Air Force, 2010.

[2] Air Force Financial Management & Comptroller, *Air Force President's Budget FY23*, Secretary of the Air Force, 2022.

[3] Charles Q. Brown, Jr., *CSAF Action Orders to Accelerate Change Across the Air Force*, Chief of Staff, U.S. Air Force, December 2020.

The DAF's training and education portfolio encompasses a wide variety of technologies. These include technologies to deliver training in live and virtual environments, sensors to capture task behaviors and monitor physiological and mental states, algorithms to objectively assess performance, and information systems to deliver training globally and provide an integrated record of an individual's training history.

Computational cognitive models of knowledge acquisition and retention are one potentially transformative training technology that the DAF is considering. These models emulate psychological processes that occur as people learn. Computational cognitive models are implemented as quantitative algorithms that estimate how knowledge is acquired and retained. These models can be used to simulate and predict the effects of various training interventions on knowledge acquisition and retention as well as to prospectively identify the interventions that will yield the greatest benefits for groups and individuals. Furthermore, these models can be calibrated to account for individual differences in ability and experience. In this way, computational cognitive models can be used to deliver personalized training that, in turn, can enhance learning gains and reduce training time.[4] Terms of art from this area are defined in the box.

Key Terms Related to Cognitive Modeling

Cognition. Mental processes relating to the acquisition, storage, retrieval, and use of information.

Cognitive Model. An approximation of cognitive processes that occur in humans or animals. These models might be used to predict and understand behaviors.

Computational Cognitive Model. A type of cognitive model that approximates mental processes using mathematical equations and is run using computer simulations. Computational cognitive models often focus on the structural properties of the cognitive system that are common across all individuals.

Parameters. Aspects of the cognitive system that vary across individuals, such as the efficacy and efficiency of learning and memory.

Inputs. An individual's current and historical experiences. When *inputs* and *parameters* are combined, a computational cognitive model might be used to predict an individual's future behavior.

Calibration. The process of using statistical methods to estimate parameters for an individual to account for their past behavior using a computational cognitive model.

Prediction. The process of using a calibrated computational cognitive model to predict the future behaviors of the individuals.

SOURCE: Matthew W. Walsh and Marsha C. Lovett, "The Cognitive Science Approach to Learning and Memory," in Susan E. F. Chipman, ed., *The Oxford Handbook of Cognitive Science*: Part 2, *Cognitive Science Perspectives on Classic Problems in Psychology*, Oxford University Press, 2016, pp. 211–230.

[4] In the remainder of the report, we use the phrase *cognitive models*.

Cognitive models encompass diverse approaches, such as neuron-level modeling, production systems that simulate the interaction between different psychology processes, and mathematical modeling that capture highly replicable behaviors.[5] The models we consider in this report are primarily at the level of mathematical expressions of invariant cognitive phenomena.

AFRL asked RAND Project AIR FORCE to examine the feasibility of using cognitive models to deliver empirically grounded and personalized training to USAF linguists. In this report, we present results of a market analysis and academic literature review of state-of-the-art approaches for adaptive training. One cognitive model, the Predictive Performance Equation (PPE), emerged as especially well suited for adaptive training, given the large number of domains and settings in which it has been used. We validate the model in the context of second-language learning and demonstrate statistical methods to calibrate the model for individual students and groups of students. Finally, we show how the PPE can be integrated into the Air Education and Training Command (AETC) Linguist Next (LN) Modern Standard Arabic Basic Course.

Study Context

This study is positioned at the convergence of learning science, instructional design, adaptive training, and DAF training needs. In this report, we briefly review the history of learning science to situate contemporary work on cognitive models. We then discuss how cognitive models can be incorporated into a variety of instructional designs, ranging from self-guided to instructor-led and from fixed to personalized designs. Within different instructional designs, cognitive models can provide adaptive training that flexibly targets the development of skills based on the unique needs of groups or individuals. Finally, we highlight different DAF training needs that adaptive training technologies might address. Although the remainder of the report focuses on the potential for cognitive models to provide adaptive training for linguists, the rest of this section illustrates the broad applicability of these methods for a variety of training needs and learning environments.

Cognitive Science Approaches to Learning and Memory

The world is astonishingly complex, and human beings are not born knowing how to respond to all—or even most—of the circumstances they face. To deal with these challenges, human beings evolved general learning mechanisms.[6] The key question in learning science is *how* humans acquire knowledge and adapt to environments as effectively as they do. Indeed, despite

[5] James L. McClelland, "The Place of Modeling in Cognitive Science," *Topics in Cognitive Science*, Vol. 1, No. 1, 2009.

[6] Matthew M. Walsh, and Kevin A. Gluck, "Mechanisms for Robust Cognition," *Cognitive Science: A Multidisciplinary Journal*, Vol. 39, No. 6, 2015.

impressive advances in artificial intelligence (AI) and machine learning (ML), no AI system yet exhibits anything like general human intelligence.[7]

Learning refers to the acquisition of knowledge, skills, and abilities (KSAs) needed to enable long-lasting behavioral change, and *memory* refers to the mental records of knowledge and experiences that underlie that change.[8] During the past two centuries, modern scientists have sought to understand how humans and animals learn, and, since the late 1800s, to link behavioral changes to the neural mechanisms that enable them. Scientists have also created computational models that emulate the neural and psychological processes that give rise to learning and memory. Several of these models have inspired contemporary approaches in AI. The following text box briefly summarizes the history of learning-science research.

[7] Gary Marcus and Ernest Davis, "Insights for AI from the Human Mind," *Communications of the ACM*, Vol. 64, No. 1, 2021.

[8] John R. Anderson, *Learning and Memory: An Integrated Approach*, 2nd ed., Wiley, 2000.

History of Learning-Science Research

Late 19th century. Hermann Ebbinghaus conducted the first rigorous studies of memory and established that memory performance increased with the amount of practice and decreased with the time elapsed since practice had occurred. At around the same time, Edward Thorndike observed that animals learn to repeat behaviors followed by rewards and to avoid ones followed by punishment.

Early 20th century. Seminal findings on memory and learning gave way to grand theories of behavior. These theories emphasized the roles of learned associations between conditions, actions, and outcomes, and the flexible use of information stored in memory to decide how to act.[a]

Mid-20th century. The advent of computing during the late 20th century led to computational cognitive models that could be implemented in computer code and used to simulate mental processes that give rise to behavior.[b]

Late 20th century. Shortly thereafter, AI emerged as a separate academic discipline.[c] Interestingly, AI researchers have developed methods that learn stimulus-response associations (e.g., model-free reinforcement learning) along with methods that manipulate knowledge in a more flexible manner (e.g., model–based reinforcement learning). This parallels historical and contemporary distinctions present in the study of human learning and memory.

At the same time, the burgeoning field of neuroscience has traced different forms of learning to different brain regions, first by studying patient populations, and then by using functional neuroimaging methods.[d]

Finally, the nascent field of educational data mining (EDM) begin to develop two classes of methods, Bayesian knowledge tracing (BKT) and logistic knowledge tracing (LKT), for tracking student learning.

Early 21st century. The advent of methods to deliver training and education content on computers and to collect measures of student learning and performance has created large datasets.[e] These datasets can be used to develop and refine cognitive models and to explore the implications of those models in naturalistic settings.

[a] Clark Leonard Hull, *Principles of Behavior: An Introduction to Behavior Theory*, Appleton-Century, 1943.
[b] Herbert A. Simon and Allen Newell, "Human Problem Solving: The State of the Theory in 1970," *American Psychologist*, Vol. 26, No. 2, 1971.
[c] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.
[d] Barbara J. Knowlton, Jennifer A. Mangels, and Larry R. Squire, "A Neostriatal Habit Learning System in Humans," *Science*, Vol. 273, No. 5280, 1996; Larry R. Squire, "Memory System of the Brain: A Brief History and Current Perspective," *Neurobiology of Learning and Memory*, Vol. 82, No. 3, 2004.
[e] Ryan S. Baker, and Kalina Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, Vol. 1, No. 1, 2009.

Findings from learning science have had significant practical implications. For example:

- *Desirable difficulties*. The primary goal of instruction is to produce long-term change, yet performance observed during instruction is an unreliable predictor of whether such change has occurred. Research on desirable difficulties has revealed that a variety of strategies (such as overlearning, distributed practice, variability of practice, and retrieval practice) make acquisition more difficult but boost long-term retention.[9]
- *Cognitive load*. Cognitive load is the demand for working memory resources placed on a learner by a task. When cognitive load exceeds an individual's working memory capacity, learning is impaired. Research on cognitive load theory has shown that varying

[9] Nicholas C. Soderstrom, and Robert A. Bjork, "Learning Versus Performance: An Integrative Review," *Perspectives on Psychological Science*, Vol. 10, No. 2, 2015.

the cognitive load according to an individual's level of expertise increases learning gains.[10]

- *Transfer*. Transfer refers to the application of learned KSAs to a new context—for example, rehearsing a task in a training environment and performing it for the first time in an operational environment. Research on transfer has helped to elucidate the perceptual, motor, and cognitive components of tasks that must be represented in simulators to engender transfer to live environments.[11]
- *Cognitive models*. Cognitive models have been developed to emulate human learning and memory. These models can be used to develop training schedules to increase long-term retention. These models can also be used to trace an individual's level of mastery to select and deliver learning events in a personalized manner.[12] A related class of models from EDM that account for the effects of myriad factors on learning have also been used to deliver personalized instruction.[13]

These findings underlie learning-science applications that have enhanced the acquisition and retention of KSAs in myriad domains and for military and nonmilitary populations.[14] In this report, we focus on the class of findings that involve cognitive models.

Instructional Design

A fundamental distinction in educational psychology is between personalized versus instructor-led learning. A second distinction is between self-guided versus fixed learning. Additionally, there are training designs in which these two dimensions intersect. Though we use the terms *student* and *classroom* in our discussion, these designs apply to other types of trainees and training settings. Presented in Figure 1.1, these training designs can be categorized as follows:

- *Fixed curriculum*. This design is most like a teacher delivering classroom instruction. The primary benefits of fixed curriculum are that the teacher can develop an effective sequence of learning events and modify their delivery to meet the needs of the class. Its primary drawback is the limited opportunity to tailor instruction to individuals' needs.

---

[10] John Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science: A Multidisciplinary Journal*, Vol. 12, No. 2, 1988.

[11] Jonathan Borgvall, Martin Castor, Staffan Nählinder, Per-Anders Oskarsson, and Erland Svensson, *Transfer of Training in Military Aviation*, Swedish Defence Research Agency, 2007.

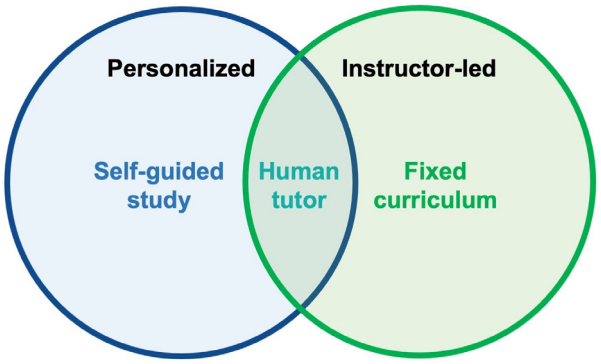
[12] John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray Pelletier, "Cognitive Tutors: Lessons Learned," *Journal of the Learning Sciences*, Vol. 4, No. 2, 1995.

[13] For a review, see Luke G. Eglington and Philip I. Pavlik Jr., "How to Optimize Student Learning Using Student Models That Adapt Rapidly to Individual Differences," *International Journal of Artificial Intelligence in Education*, 2022.

[14] John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham, "Improving Students' Learning with Effective Learning Techniques: Promising Directions from Cognitive and Educational Psychology," *Psychological Science in the Public Interest*, Vol. 14, No. 1, 2013.

- *Human tutor*. One-on-one tutoring is the gold standard in education.[15] The primary benefits are that the teacher can continuously assess student understanding and tailor learning events to meet the student's needs. Its primary drawback is cost.
- *Self-guided study*. This design entails student-guided selection of learning events. The primary benefit of self-guided study is the opportunity for students to review weak concepts. Its primary drawbacks are the reliance on accurate self-assessment and selection of effective learning strategies.[16]

Figure 1.1. Training System Designs Formed by the Intersection of Primary Dimensions



These designs are not mutually exclusive. For example, a teacher might supplement classroom instruction with out-of-class assignments. The teacher might also provide limited one-on-one tutoring. Finally, students might engage in self-directed learning outside the classroom.

Figure 1.2 shows the expanded design space created by incorporating cognitive models into training and education. Models can be incorporated into each of the designs previously discussed. By enabling new affordances, models can enhance each design in the following ways:

- *Model evaluates curriculum.* The cognitive model is used to evaluate the quality of a curriculum, focusing on things like the selection and sequencing of learning events. This might produce a more empirically grounded curriculum.
- *Model recommends classroom activities to instructor*. The cognitive model traces student knowledge and then recommends classroom learning events to reinforce the weakest concepts for the most students.
- *Model recommends student activities to instructor.* The cognitive model also traces student knowledge and then recommends personalized activities for the instructor to deliver to students during one-on-one tutoring.

---

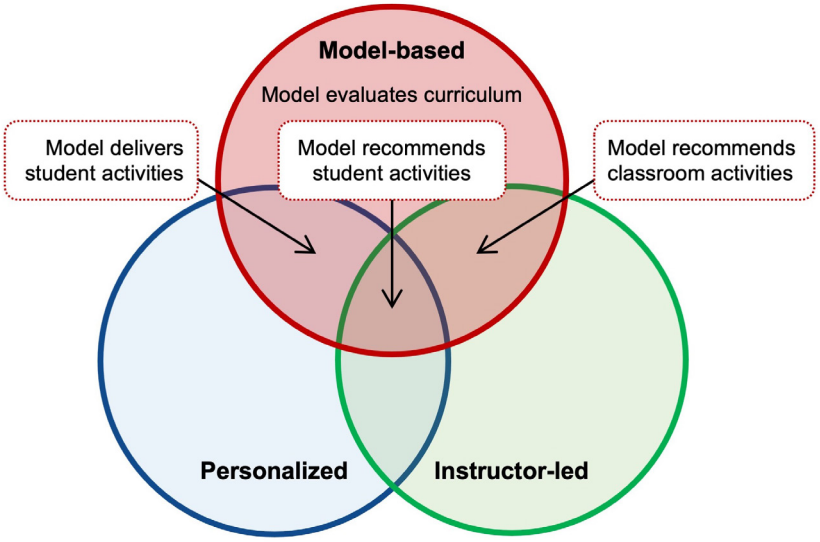
[15] Benjamin S. Bloom, "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," *Educational Researcher*, Vol. 13, No. 6, June–July 1984.

[16] Nate Kornell and Bridgid Finn, "Self-Regulated Learning: An Overview of Theory and Data," in John Dunlosky and Sarah Tauber, eds., *The Oxford Handbook of Metamemory*, Oxford University Press, 2016.

- *Model delivers student activities*. As in the previous two approaches, the cognitive model traces student knowledge. The model then directly delivers personalized activities to students to reinforce the weakest concepts. This is sometimes called a *cognitive tutor*.[17]

To reiterate, these designs are not mutually exclusive. Educational designers can use a cognitive model to develop and evaluate a curriculum in advance. Instructors can then continue to use the model to tailor learning events to meet individual and class needs.

Figure 1.2. Expanded System Design Space Created by Including Cognitive Models



Adaptive Training

The philosophy of adaptive training is to create a flexible learning environment that supports the needs of individuals with different abilities, interests, and backgrounds. In terms of motivation and learning outcomes, the benefits of adaptive training are well established.[18] However, it is difficult to monitor student progress and correctly diagnose the needs of different individuals.

Figure 1.3 shows how cognitive models can enable adaptive training.[19] The figure shows a four-step adaptive training cycle. At the onset of the training cycle, personal information about

---

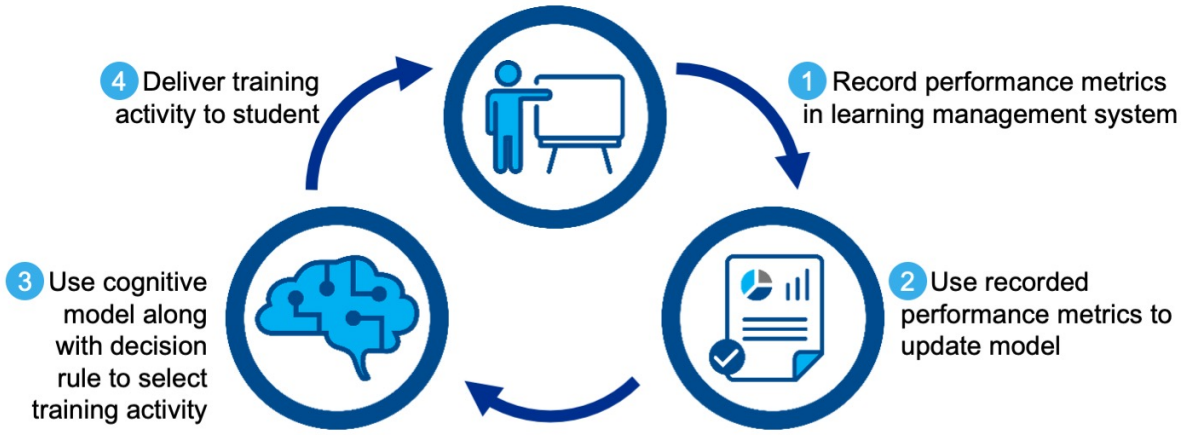
[17] Anderson et al., 1995.

[18] Carla R. Landsberg, Randy S. Astwood, Jr., Wendi L. Van Buskirk, Lisa N. Townsend, Natalie B. Steinhauser, and Alyssa D. Mercado, "Review of Adaptive Training System Techniques," *Military Psychology*, Vol. 24, No. 2, 2012; Christopher D. Wickens, Shaun Hutchins, Thomas Carolan, and John Cumming, "Effectiveness of Part-Task Training and Increasing-Difficulty Training Strategies: A Meta-Analysis Approach," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 55, No. 2, 2012.

[19] Valerie J. Shute and Diego Zapata-Rivera, "Adaptive Technologies," *ETS Research Report*, RR-07-05, 2007.

the learner and performance data, such as response accuracy, are captured and recorded in a learning management system (LMS). These data are analyzed to infer the learner's current level of mastery. The output is sometimes called a *student model* because it is a digital representation of what the student understands. The model is used to select learning activities to correct or strengthen the learner's grasp of domain knowledge. Finally, the selected content is delivered. The adaptive training cycle repeats as additional data from learner interactions are captured. The cognitive model emulates the role of a human instructor by diagnosing student weaknesses and selecting and delivering learning experiences to address those weaknesses.

Figure 1.3. Adaptive Training Cycle



United States Department of the Air Force Training Needs

The primary objectives of adaptive training are to allocate a fixed amount of training resources to optimize learning outcomes or produce equivalent learning outcomes using fewer resources. Compared with one-size-fits-all training, the adaptive cycle shown in Figure 1.3 achieves these objectives by accelerating training for individuals who quickly master the material by avoiding wasted learning experiences and by delivering learning experiences that are tailored to individual needs. By applying adaptive training, the DAF can accrue significant performance gains and savings across the full spectrum of training and education activities. As the examples in Table 1.1 show, adaptive training might address certain training challenges for many DAF occupations.

Table 1.1. Opportunities for Personalized Training in United States Department of the Air Force Occupations

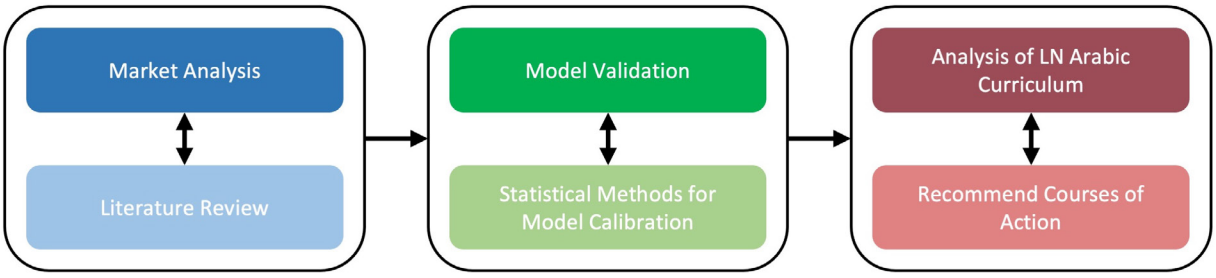
Example	Description	Potential for Cognitive Model–Based Training
Pilots	The Ready Aircrew Program (RAP) specifies the mix of live and virtual training mission types and events that pilots must complete annually to remain Basic Mission Capable or Combat Mission Ready.[a]	Adaptive training, enabled by objective measures of pilot performance, might allow the DAF to allocate limited flying hours more effectively and better assess risks of failing to meet RAP requirements.
Maintainers	The DAF has considered consolidating F-35A maintenance career fields. This could reduce personnel requirements and increase force generation at austere locations. However, aircraft maintainers would need to be proficient in more skills than are now required.[b]	Adaptive training could be used to deliver effective sequences of training events to develop and sustain a larger set of mechanical skills. By recording operational maintenance events, adaptive training could be used to reinforce skills that have likely decayed.
Medical	Comprehensive Medical Readiness Program checklists document knowledge and skills that medical professional must possess and maintain to be qualified to provide in-garrison care and to be ready for combat.[c] Because of low patient volume at many medical treatment facilities, it might be difficult to build and maintain proficiency across the full set of elements.	Adaptive training could be used to deliver effective sequences of training events to develop and sustain a large set of clinical skills. A cognitive model could also be used to determine assignment sequences to build proficiency at critical times during an individual's career (e.g., prior to a deployment).
Linguists	Second-language learning requires the acquisition of more than 1,000 task-critical words.[d] Because many of the words occur with low to moderate frequency in natural language, words must be periodically restudied.	Adaptive training could be used to assess student mastery and tailor the review of task-critical vocabulary.
Multi-capable airmen	The USAF's emerging multi-capable airmen construct calls for equipping airmen with secondary skill sets beyond those needed in their primary occupations. The cost to deliver training to develop these secondary skill sets along with the risk to sustain and assess proficiency during periods of disuse are significant.[e]	Adaptive training could be used to deliver effective sequences of training events that develop and sustain secondary skill sets. A cognitive model could also be used to track proficiency and assess risk during periods of disuse.

[a] Air Force Manual 11-2RC-135, Vol. 1, *Flying Operations*, Secretary of the Air Force, 2019.
[b] Anna Jean Wirth, Thomas Light, Daniel M. Romano, Shane Tierney, Ronald G. McGarvey, Moon Kim, Michael J. Lostumbo, Amanda Nguyen, Paul Emslie, and John G. Drew, *Evaluating Alternative Maintenance Manpower Force Structure Concepts for the F-35A*, RAND Corporation, RR-4433-AF, 2020.
[c] Air Force Instruction 41-106, *Health Services, Air Force Medical Readiness Program*, Secretary of the Air Force, 2020.
[d] Further details provided in Chapter 4.
[e] Air Force Doctrine Note 1-21, *Agile Combat Employment*, Secretary of the Air Force, 2021.

Study Methodology

To evaluate the feasibility of applying cognitive models of knowledge acquisition and retention to DAF training and education, we followed the approach shown in Figure 1.4.

Figure 1.4. Study Approach



Market Analysis

We compiled a database of technology programs that spans a variety of sectors (academia, industry, and government) and research activities (basic research, advanced technology development, and operational system development). The database contains more than 35,000 records. We identified records that pertain to adaptive training technologies and analyzed the attributes and maturity of those technology programs. Our analysis led us to favor cognitive models as an approach for delivering adaptive training.

Literature Review

We reviewed academic literature on cognitive models of knowledge acquisition and retention. Using the collection of 55 publications identified from the past ten years, we analyzed major classes of cognitive models, along with the domains (e.g., language learning, medical skills, and military skills) and settings (e.g., laboratory, classroom, and field study) in which they have been used. Our analysis led us to favor PPE, a cognitive model employed across an exceptionally diverse set of domains and settings.

Quantitative Analysis of Large-Scale Databases Gathered in Naturalistic Settings

We fitted PPE to a dataset on second-language learning gathered on Duolingo, a commercial platform.[20] The dataset provided the opportunity to validate PPE using large samples of individuals performing a knowledge-acquisition task (i.e., second-language learning) in ecologically valid conditions: that is, in a self-paced manner, using personal devices, and at times and in locations chosen by individuals. Evaluating PPE in ecologically valid conditions is an important step beyond laboratory evaluations. Using the construct of technology readiness levels

[20] As described in Chapter 3, we used the second language acquisition modeling (SLAM) dataset from the U.S. language-learning website Duolingo.

(TRLs), this constitutes an advance from component validation in a laboratory environment (TRL 4) to prototype demonstration in a relevant environment (TRL 6).[21]

Develop Statistical Methods to Calibrate PPE to Individuals and Groups of Students

Using objective performance metrics like response accuracy, PPE can be calibrated to account for individual differences in learning and memory. In this way, PPE can make personalized predictions that can be used for adaptive training. In this regard, a key challenge is to combine information from an individual student with information from other students. At the outset of training, little is known about an individual's learning abilities. Information about other students can be used to bootstrap model predictions. As more data are gathered from an individual, PPE can be calibrated to better reflect their distinct abilities. We used an approach called *Bayesian hierarchical modeling* to optimally pool information across individual students.

Case Study Focused on Linguist Training

To ground our evaluation of the feasibility of cognitive model–based personalized training, we focused on applying PPE to the LN Standard Arabic Basic course. We catalogued learning activities delivered via the course's LMS, the Universal Curriculum and Assessment Tool (UCAT), and student response data captured by UCAT. We also catalogued learning activities delivered outside UCAT.

From our analysis of the course goals, data sources, and information technology (IT) system affordances, we propose four courses of action (COAs) that involve using PPE to (1) develop an empirically grounded curricula that adheres to the principles of knowledge acquisition and retention, (2) predict student retention of task-critical vocabulary during periods of disuse, (3) recommend activities to instructors to reinforce the weakest language concepts, and (4) automatically deliver activities to students to reinforce the weakest language concepts.

Organization of the Report

The remainder of this report is organized as follows:

- Chapter 2 describes data sources used for the market analysis and literature review and presents findings about trends in adaptive and cognitive model–based training.
- Chapter 3 describes the large-scale database used to validate PPE along with the statistical methods developed to calibrate the model.
- Chapter 4 describes the case study focused on integrating PPE with the LN Standard Arabic Basic course.
- Chapter 5 summarizes our findings and makes recommendations.

[21] TRLs range from 1 (basic principles observed and reported) to 9 (actual system proven through successful mission operations).

- The appendix provides a technical description of PPE and of model fitting and evaluation methods. We also provide additional supporting analyses.

Chapter 2. Trends in Adaptive Training

Students receiving one-on-one tutoring reach dramatically higher levels of academic achievement than those in conventional classroom settings. Educational psychologist Benjamin Bloom observed that the average student who received one-on-one tutoring outperformed more than 95 percent of students in a conventional classroom.[22] This finding has since been replicated; one-on-one tutoring is consistently found to be one of the most effective methods of instruction.[23]

The benefits of one-on-one tutoring arise from numerous factors, including how content is adapted based on the learner's needs. Recently, there has been increased demand for technologies to improve training effectiveness, including technologies to adapt the content being delivered to better suit the learner. Collectively, these are called *adaptive training technologies*.[24]

In this chapter, we present a market analysis on adaptive training technologies spanning technology readiness (i.e., from basic research to operational system development) and sectors (i.e., academia, industry, and government). The purpose of the market analysis is to understand the landscape of adaptive training technologies being developed, including but not limited to computational cognitive models. Approaches that leverage cognitive models make up a subset of adaptive training technologies. We find that cognitive model–based approaches have not been sufficiently explored, yet they offer advantages beyond methods that use ML or heuristic approaches to adapt training.

Following the market analysis, we present a literature review focused on cognitive model–based approaches to adaptive training. The purpose of the literature review is to understand the origins of various cognitively inspired approaches and to survey the strength of evidence for each. For the adaptive training approaches and associated tools that we identify, we document the settings in which they have been demonstrated. The fact that the PPE model has been studied across multiple domains relevant to military operations and across field study, classroom, and laboratory settings make it an attractive option for delivering cognitive model–based adaptive training.

---

[22] Bloom, 1984.

[23] John Hattie, John Biggs, and Nola Purdie, "Effects of Learning Skills Interventions on Student Learning: A Meta-Analysis," *Review of Educational Research*, Vol. 66, No. 2, 1996.

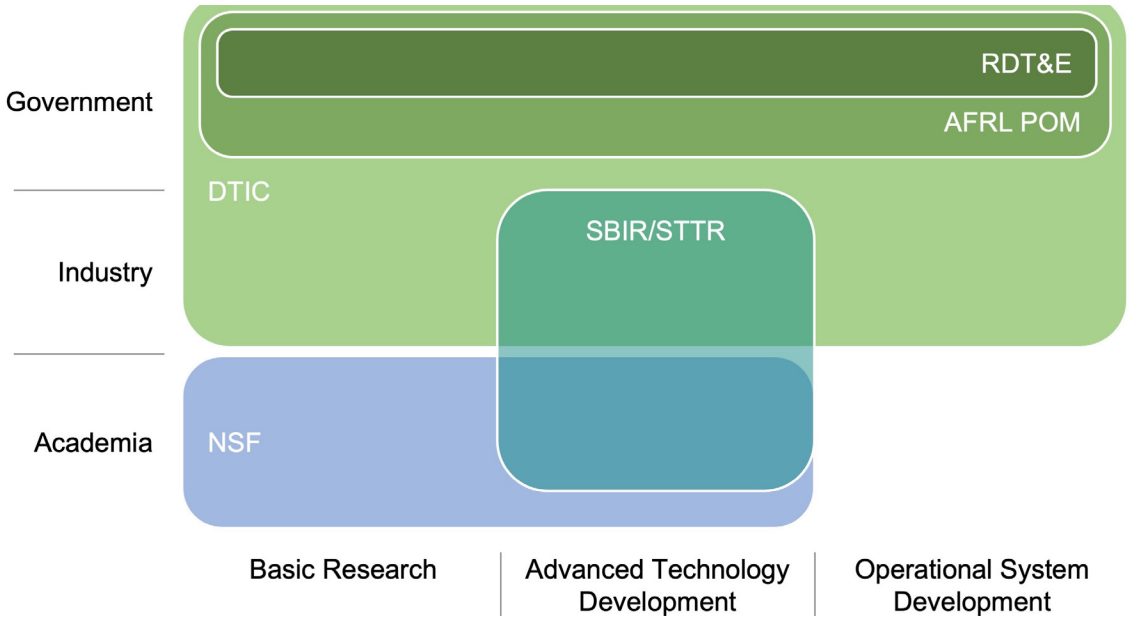
[24] Ok-choon Park and Jung Lee, "Adaptive Instructional Systems," in David H. Jonassen and Marcy Driscoll, eds., *Handbook of Research for Educational Communications and Technology*, Routledge, 2004.

Market Analysis

Methods

To characterize the landscape of work on adaptive training technologies, we gathered descriptions of research and development programs from the sources shown in Figure 2.1. Technologies described in these sources exist across a variety of basic research, applied research, system development, and sustainment activities. Additionally, these technologies originate from government, industry, and academic sectors. The regions shown in Figure 2.1 correspond to the research and development activities and sectors that each source encompasses. For example, the National Science Foundation (NSF) primarily funds academic research to demonstrate scientific principles and develop proofs-of-concept (i.e., basic research and advanced technology development). The Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs fund academic and industry research to develop proofs-of-concept. The Defense Technical Information Center (DTIC) contains reports on industry and government work that span all research and development activities. Finally, the Research, Development, Test and Evaluation (RDT&E) Budget Submission and the AFRL Program Objective Memorandum (POM) submission pertain to government work that spans all research activities.

Figure 2.1. Data Sources



NOTE: RDT&E Defense Budget Materials (2,169 records); AFRL POM submissions (619 records); SBIR and STTR awards (33,084 records); DTIC abstracts (28,775 records); NSF awards (35,519 records).

To process data from these sources, we retained records with variations of the search terms "training," "adaptive" or "personalized," and "cognitive model" or "student model." These three sets of terms encompass progressively narrower bands of training technologies.

Results

Table 2.1 shows the total number of records and monetary amounts of research activities related to adaptive and cognitive model–based training from the five sources.[25] Adaptive training technologies made up less than one percent of total records. Cognitive model–based approaches, in turn, made up only about ten percent of all records on adaptive training technologies.[26]

Table 2.1. Number and Size of Investments in Adaptive Training Technology Systems

		Adaptive Training		Cognitive Model–Based Training	
Source	Total Records	Count	Amount ($ millions)	Count	Amount ($ millions)
AFRL	619	5	–	2	–
RDT&E	2,169	3	51.3	1	22.4
SBIR/STTR	33,084	157	70.2	10	3.4
NSF	30,291	58	39.3	12	6.0
DTIC	28,737	38	–	1	–

NOTE: Data in the Cognitive Model–Based Training columns are subsets of the records contained in the matching Adaptive Training columns.

We performed a thematic analysis to group adaptive training technology programs according to how the technology adapts to meet learner needs (see graph 1 in Figure 2.2). Many records citing adaptive training technologies failed to specify how the system adapts. AI- and ML-based, heuristic-based, and cognitive model–based approaches were the most common methods

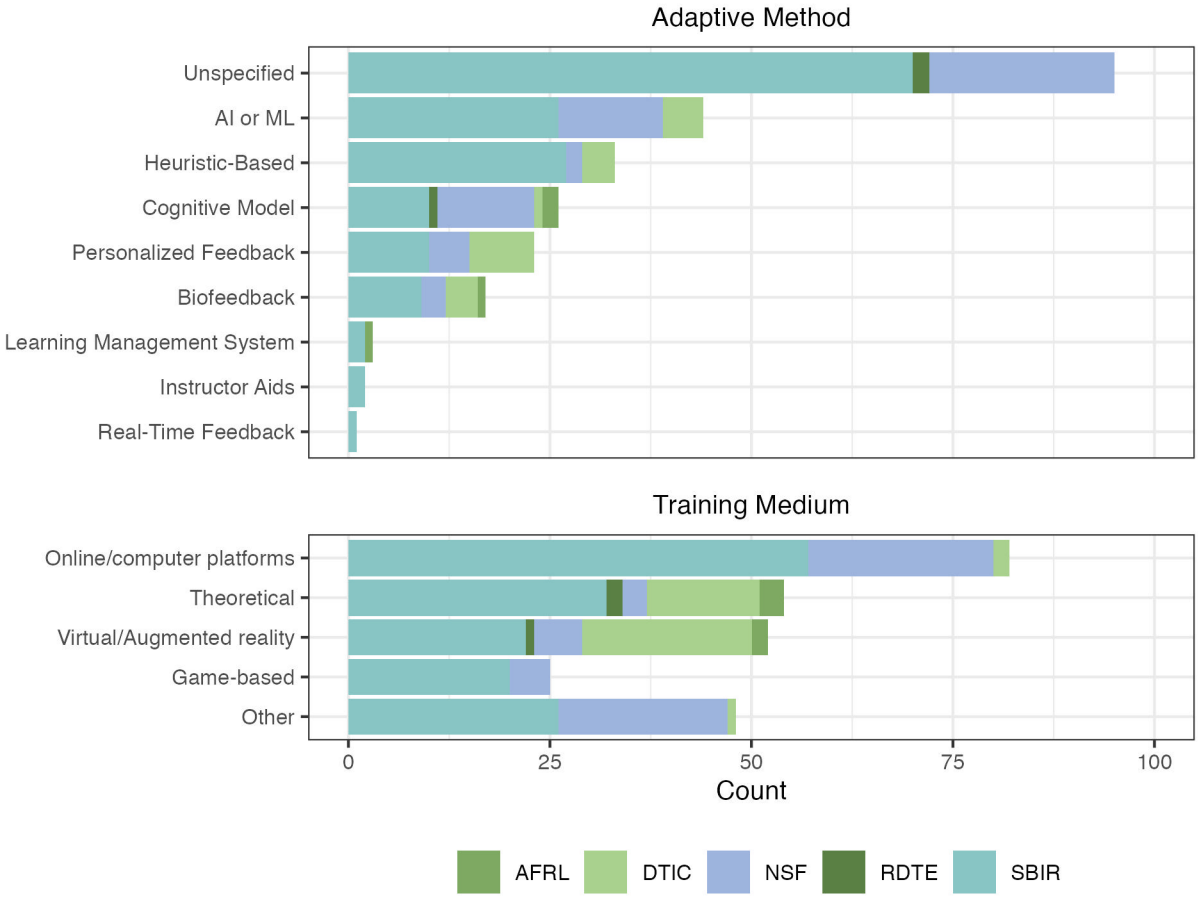
[25] This does not include, for example, O&M and purchases.

[26] The descriptions of FY 2022 AFRL research programs included five on adaptive training, two of which used cognitive models. All were led by the Airman Systems Directorate (AFRL/RH). The descriptions of U.S. Department of Defense (DoD) projects from the FY 2023 RDT&E budget submission included three on adaptive training. Two were led by the Army, and the third, which included cognitive models, was led by the DAF. The DAF submission covers the portion of work being performed by AFRL/RH. Of the SBIR/STTR awards granted in calendar years (CYs) 2016–2021, 157 involved adaptive training, ten of which used cognitive models. DoD and the NSF funded the largest number of SBIR/STTRs on adaptive training, followed by the Department of Health and Human Services, the Department of Education, and the National Aeronautics and Space Administration. Of the NSF awards granted in CYs 2016–2021, 58 involved adaptive training, 12 of which used cognitive models. These were concentrated in the Computer and Information Science and Engineering and the Engineering directorates. Finally, DTIC records contained 38 descriptions of adaptive training technologies, only one of which involved cognitive models.

described among the remaining records.[27] Of note, AI- and ML-based (combined) and cognitive model–based approaches made up most of the NSF records at 37.1 percent and 34.3 percent, respectively. Conversely, heuristic-based approaches (e.g., staircase procedures that incrementally adjusted difficulty or procedures that repeated content after the student responded incorrectly) made up most SBIR and STTR records, at 31.0 percent.

We also grouped adaptive training technology programs based on the training medium (see graph 2 in Figure 2.2). Of note, the majority of NSF and SBIR records involved adaptive training using online and computer platforms (39.7 percent and 36.3 percent, respectively), whereas the majority of DTIC records involved adaptive training using virtual or augmented reality (55.3 percent). Only SBIR and NSF records involve game-based adaptive training.

Figure 2.2. Approaches to Adaptive Training (Graph 1) and Training Medium (Graph 2)



[27] Records varied greatly in terms of the amount of detail they provide. Although some described the manner of adaptation at length, others simply stated that a technology was adaptive or personalized.

Table 2.2 displays the three main approaches to adaptive training from the market analysis—cognitive model–based, AI- or ML-based, and heuristic-based—and the main differences between them. Cognitive model–based approaches use cognitive models to convert training-related variables into performance predictions. The underlying equations are based on psychology research and provide a precise mapping between inputs and performance predictions. ML-based approaches use ML to learn black-box models to map training-related variables to performance predictions. Unlike the case of cognitive models, these black-box models are developed from data and are not based on psychological theory. Finally, heuristic-based approaches use simple rules to adapt training in a principled manner based on an individual's recent performance.

We evaluated these approaches along six dimensions. *Accuracy* is the ability of the approach to correctly predict learner performance. *Generalizability* is the ability to apply the model to new tasks and environments. *Completeness* is the extent to which the approach accounts for all factors that affect knowledge acquisition and retention. *Interpretability* is whether the model can be explained to an end-user. *Computational efficiency* refers to the computing demands to implement the approach. *Data efficiency* is the amount of performance data needed to apply the approach.

As shown in Table 2.2, cognitive model–based approaches perform well across all dimensions. ML-based approaches have relatively lower generalizability because they are typically developed for a specific task, context, and group of individuals. They might also learn "black box" models that increase predictive accuracy but trade off interpretability. Finally, ML-based approaches require a large amount of data to learn predictive models. Heuristic-based approaches have lower generalizability because they are typically tied to task-specific performance measures. These approaches are also incomplete because they do not represent all the factors that might affect performance. Finally, although not shown in the table, heuristic-based approaches do not predict current and future proficiency. Rather, they use rules to prescribe what to deliver next based only on how the learner has performed in the past.

Table 2.2. Classes of Approaches to Adaptive Training

Approach	Accuracy	Generalizability	Completeness	Interpretability	Computational Efficiency	Data Efficiency
Cognitive model–based	High	High	Medium	High	Medium	Medium
ML-based	High	Low	Medium	Low	Medium	Low
Heuristic-based	Medium	Low	Low	High	High	High

Summary

Across research programs and sectors, only a small percentage of work involves adaptive training. Furthermore, cognitive model–based approaches make up a small fraction (10 percent) of those projects. AI- and ML-based approaches (combined) and heuristic-based approaches are the other most common techniques cited in adaptive training research programs and made up 17 and 13 precent of adaptive training records, respectively.

The key takeaway from this analysis is that cognitive model–based approaches are relatively underexplored, yet they might offer benefits beyond ML- and heuristic-based approaches to adaptive training. Furthermore, adaptive training might be delivered in a variety of mediums, including online and computer platforms, virtual and augmented reality, and games.

Literature Review

Methods

To further characterize the current state of adaptive learning, we reviewed the academic literature on how cognitive models are being used for adaptive training.[28] We collected relevant articles through online database searches and by using material from previous related research. We limited our search to English-language articles published during the past ten years (2011 to 2021) in peer-reviewed journals or conference proceedings.[29]

Our search returned 55 relevant articles. We grouped articles based on application domain (language, military tasks, medical tasks, unspecified, and other), setting (laboratory, classroom, field study, theoretical, and other), and population (civilian or military). Table 2.3 shows article counts by group.

[28] As a scoping review, this is not intended to cover the field comprehensively but rather to examine the "nature and extent of research evidence." See Maria J. Grant and Andrew Booth, "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies," *Health Information & Libraries Journal*, Vol. 26, No. 2, 2009.

[29] We conducted the search on Google Scholar using combinations of the terms "adaptive training," "personalized training," "language," "artificial intelligence," "machine learning," "Air Force," "Army," "military," "return on investment," and "cognitive model." A project team member reviewed the search results and retained articles on cognitive models and training.

Table 2.3. Article Counts by Application Domain, Setting, and Population

		Application Domain				
Population	Setting	Language (13)	Military Tasks (4)	Medical Tasks (3)	Unspecified (15)	Other (20)
Civilian (37)	Laboratory (14)	5			4	5
	Classroom (6)	3		1		2
	Field Study (7)	3				4
	Theoretical (5)	1			4	
	Other (5)				1	4
Military (18)	Laboratory (2)			1		1
	Classroom (9)	1	4	1		3
	Field Study (1)					1
	Theoretical (6)				6	
	Other (0)					

Results

Our literature review included multiple cognitive models of knowledge acquisition and retention.[30] Figure 2.3 summarizes the models and shows how they are related. These models have been developed, refined, and validated extensively through laboratory studies, so tools that use these models are theoretically grounded.

One class of models has evolved from a common predecessor, Adaptive Control of Thought—Rational (ACT-R).[31] ACT-R uses an activation-based theory of memory. Each time a piece of information is encountered, its representation in memory increases. Activation gradually decreases over time. Information with high activation is more accessible from memory. The General Performance Equation (GPE) and the PPE provide mathematical approximations of how the recency, frequency, and temporal spacing of practice determine activation.[32] Finally, the

[30] Matthew M. Walsh, Kevin A. Gluck, Glenn Gunzelmann, Tiffany S. Jastrzembski, Michael Krusmark, Jay I. Myung, Mark A. Pitt, and Ran Zhou, "Mechanisms Underlying the Spacing Effect in Learning: A Comparison of Three Computational Models," *Journal of Experimental Psychology: General*, Vol. 147, No. 9, 2018.

[31] John R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, 2007.

[32] John R. Anderson and Christopher D. Schunn, "Implications of the ACT-R Learning Theory: No Magic Bullets," in Robert Glaser, ed., *Advances in Instructional Psychology: Educational Design and Cognitive Science*, Vol. 5, Lawrence Erlbaum Associates, 2000; Walsh, Gluck, Gunzelmann, Jastrzembski, Krusmark, Myung, et al., 2018.

Pavlik and Anderson (P&A) model describes how factors such as desirable difficulty improve the durability of information in memory.[33]

The second class of models have evolved from the Search of Associative Memory (SAM) model.[34] SAM attributes practice and forgetting effects to the set of contextual elements associated with information (e.g., the external and internal context in which information was encountered). Because context changes over time, SAM predicts that memory performance will decline during periods of disuse. The multi-scale context model (MCM) and others derived from it borrow elements from SAM, but they also describe how the temporal distribution of practice affects the durability of information in memory.[35]

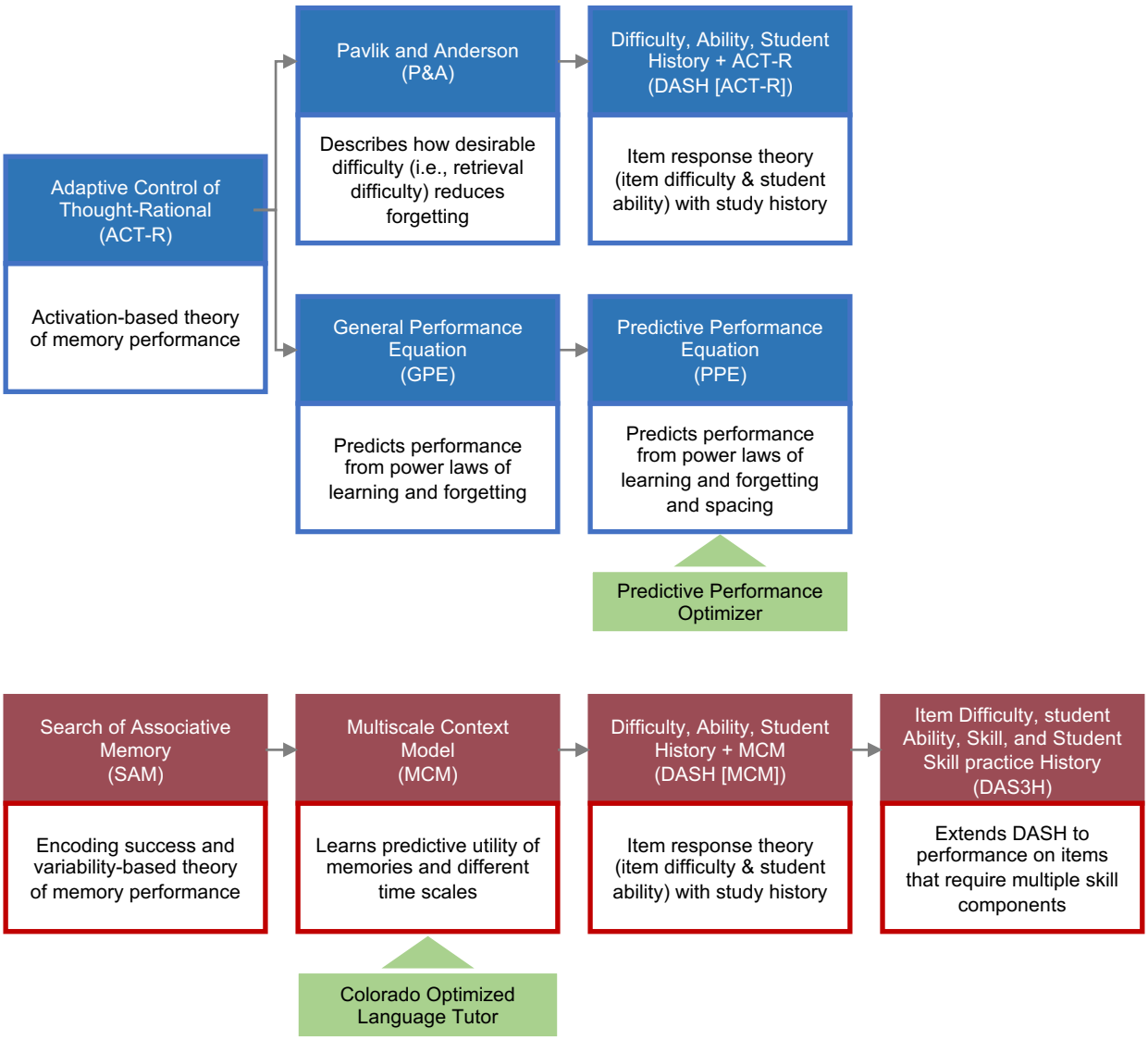
---

[33] Philip I. Pavlik, Jr., and John R. Anderson, "Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect," *Cognitive Science: A Multidisciplinary Journal*, Vol. 29, No. 4, 2010.

[34] Jeroen G. Raaijmakers and Richard M. Shiffrin, "Search of Associative Memory," *Psychological Review*, Vol. 88, No. 2, 1981.

[35] Harold Pashler, Nicholas Cepeda, Robert V. Lindsey, Ed Vul, and Michael C. Mozer, "Predicting the Optimal Spacing of Study: A Multiscale Context Model of Memory," in Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, eds., *Advances in Neural Information Processing Systems*, Vol. 22, Neural Information Processing Systems Foundation, 2009.

Figure 2.3. Families of Cognitive Models of Memory



Some cognitive models shown in Figure 2.3 have been embedded in tools to deliver adaptive training (Table 2.4). These tools include a cognitive model along with a decision method for using the model's outputs to select instructional sequences. For example, the Predictive Performance Optimizer (PPO) allows users to enter performance objectives, and it applies operations research methods to PPE's outputs to select schedules that optimally meet those

objectives.[36] The Colorado Optimized Language Tutor (COLT) combines MCM with a heuristic that selects the item expected to benefit most from the very next practice repetition.[37] General-purpose tools like reinforcement learning have also been applied to instructional sequences, or *policies*, to enhance learning outcomes.[38]

Research has also shown that adaptive training systems benefit from models of *time costs*—that is, the difficulty of retrieving items from memory is associated with the time needed to complete the task. Thus, given a fixed amount of time, delivering practice that targets the hardest items might reduce the total number of items that can be rehearsed. To counter this, other tools use an efficiency metric that captures the rate of gain per unit of time. This metric is applied to model outputs to select items that will produce the greatest rate of gain.[39]

Other tools have been developed that use heuristics derived from learning-science research, rather than cognitive models, to deliver adaptive training. The predecessor to these is the Leitner system, which gradually expands the amount of time between repetitions as memory performance improves. The Personalized Adaptive Scheduling System (PASS) follows the same principle, whereas Adaptive Response-Time-Based Sequencing (ARTS) uses correctness and speed of response to determine which items to present next.[40] Finally, the MEMORIZE algorithm offers a general approach to computing optimal study schedules that can use any memory model.[41]

---

[36] Tiffany S. Jastrzembski, Kevin A. Gluck, and Stuart Rodgers, "Improving Military Readiness: A State-of-the-Art Cognitive Tool to Predict Performance and Optimize Training Effectiveness," *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2009.

[37] Tiffany S. Jastrzembski, Kevin A. Gluck, and Stuart Rodgers, "The Predictive Performance Optimizer: An Adaptive Analysis Cognitive Tool for Performance Prediction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53, No. 21, 2009.

[38] Shayan Doroudi, Vincent Aleven, and Emma Brunskill, "Where's the Reward? A Review of Reinforcement Learning for Instructional Sequencing," *International Journal of Artificial Intelligence in Education*, Vol. 29, No. 4, 2019.

[39] Luke G. Eglington and Philip I. Pavlik, Jr., "Optimizing Practice Scheduling Requires Quantitative Tracking of Individual Item Performance," *npj Science of Learning*, Vol. 5, No. 1, 2020.

[40] Xiaolu Xiong, Yan Wang, and Joseph Barbosa Beck, "Improving Students' Long-Term Retention Performance: A Study on Personalized Retention Schedules," *LAK '15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, March 2015; Everett Mettler, Christine M. Massey, and Philip J. Kellman, "A Comparison of Adaptive and Fixed Schedules of Practice," *Journal of Experimental Psychology: General*, Vol. 145, No. 7, 2016.

[41] Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez, "Enhancing Human Learning via Spaced Repetition Optimization," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 116, No. 10, 2019.

Table 2.4. Model- and Theory-Based Tools

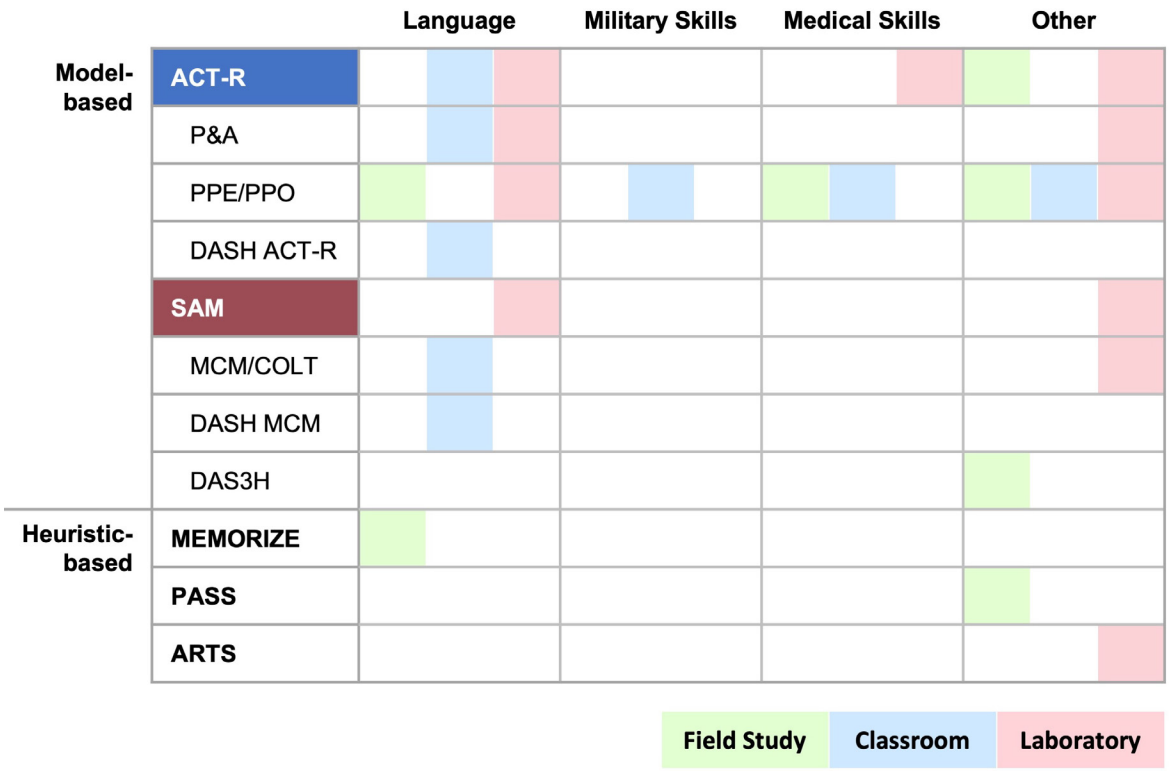
Tool	Class	Description
PPO	Cognitive model–based	Uses PPE to derive optimal study schedules
COLT	Cognitive model–based	Uses MCM to deliver review to maximize learning gains
Leitner System	Heuristic-based	Progressively increases time between repetitions based on memory performance
PASS	Heuristic-based	Progressively increases time between repetitions based on memory performance
ARTS	Heuristic-based	Progressively increases time between repetitions based on memory speed and performance
MEMORIZE	Cognitive model–based	General approach to derive optimal study schedules using cognitive models

Figure 2.4 maps model-based and heuristic-based adaptive training approaches to domains and settings where they have been applied. This visualization reveals three findings:

1. Cognitive models and other theory-based approaches have been commonly applied to second-language learning. Although our search terms were not tailored for military or medical training, cognitive models have been applied in these domains as well.
2. Although most models and tools have only been applied in one domain, PPE has been applied across all the domains we considered.
3. Most models and tools have been demonstrated only in laboratory studies or in classroom settings. PPE has also been evaluated in large-scale field studies. For this reason, and because PPE performs favorably in comparisons with other computational cognitive models, we focus on PPE in the subsequent chapters.[42]

[42] Matthew M. Walsh, Kevin A. Gluck, Glenn Gunzelmann, Tiffany S. Jastrzembski, and Michael Krusmark, "Evaluating the Theoretic Adequacy and Applied Potential of Computational Models of the Spacing Effect," *Cognitive Science: A Multidisciplinary Journal*, Vol. 42, Supp. 3, 2018; Walsh, Gluck, Gunzelmann, Jastrzembski, Krusmark, Myung, et al., 2018.

Figure 2.4. Cognitive Models and Tools by Study Domain and Study Setting



These methods vary in terms of theoretical motivation, computational complexity, and data requirements. However, all have been shown to produce significant memory gains beyond ones achieved using nonadaptive training. Table 2.5 summarizes the most-cited benefits of personalized training from the literature. These include size and durability of learning gains, increased learning efficiency, and associated cost savings.

Notwithstanding these benefits, there are also barriers to implementing these solutions (Table 2.5). The most cited barrier in the academic literature is TRL. In addition, adaptive training might require data sources that are not recorded or are not readily available—for instance, for security reasons. More fundamentally, it might be difficult to assess performance and automatically adapt training for complex, ill-defined tasks. Other cited barriers include upfront costs, standards and interoperability, and the difficulty of demonstrating return on investment.

Table 2.5. Opportunities and Barriers for Cognitive Model–Based Adaptive Training

Class	Title	Description	Frequency (%)
Opportunities	Better learning outcomes	Students achieve higher levels of mastery and retain knowledge longer	59
	Increased personalization	Training can be tailored based on each student's strengths, weaknesses, and needs	44
	Increased learning efficiency	Students achieve the desired level of mastery in less time and using fewer resources	31
	Readiness assessment	Training data can be used to evaluate performance and track proficiency during periods of disuse	18
	Cost savings	Related to learning efficiency, training can be accomplished more efficiently	15
Barriers	State of technology	The algorithms for delivering adaptive training and the enabling infrastructure are not yet mature	32
	Data challenges	Models require a large amount of data, or data that are not captured or made available	20
	Complex environments	It is difficult to assess performance and to select learning events for complex environments	11
	Upfront cost	An upfront investment is needed to field an adaptive training system	7
	Standards and interoperability	The adaptive training system must be interoperable with other information systems in the training ecosystem	5
	Training asset evaluation	It might be difficult to demonstrate the benefits of adaptive training solutions	5

NOTE: Frequency shows the percentage of articles that mentioned the benefit or barrier.

Complementary Approaches from Educational Data Mining

The approaches described in the previous section arose from cognitive science research and reflect an attempt to apply basic scientific principles to training and education. The field of EDM arose from the complementary perspective of leveraging data on student learning to enhance training and education.

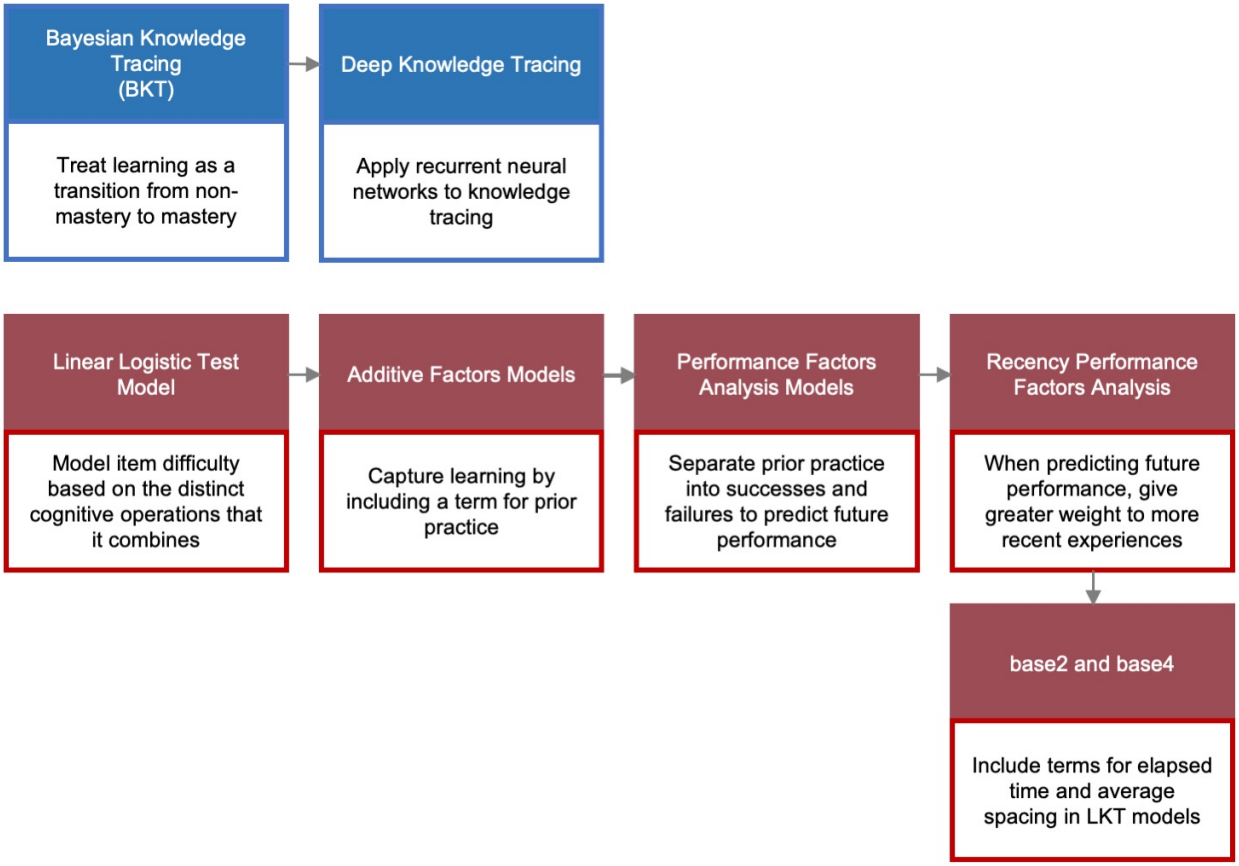
Figure 2.5 summarizes EDM approaches for personalized learning and shows how they are related. One class of approaches build on BKT.[43] These approaches treat learning as a Hidden Markov Model. The goal of BKT is to estimate the probability that, for a particular knowledge component, the student transitions from non-mastery to mastery following each practice repetition. A second class of approaches build on LKT.[44] These approaches combine a variety of

[43] Albert T. Corbett and John R. Anderson, "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," *User Modeling and User-Adapted Interaction*, Vol. 4, No. 4, 1994.

[44] For a review, see Eglington and Pavlik, 2022.

variables (such as item, number of practice repetitions, number of correct and incorrect responses, elapsed time, and spacing) in a general linear model to predict the probability of a correct response.

Figure 2.5. Families of Educational Data Mining Approaches



The distinction between EDM methods and cognitive models is somewhat unclear. For example, LKT methods like base2 and base4 incorporate features that use psychological theory.[45] In turn, cognitive models like PPE, Difficulty, Ability, Student History (DASH) MCM, and item Difficulty, student Ability, Skill, and Student Skill practice History (DAS3H) are expressed as general linear models much like LKT methods.

[45] Philip I. Pavlik, Luke G. Eglington, and Leigh M. Harrell-Williams, "Logistic Knowledge Tracing: A Constrained Framework for Learner Modeling," *IEEE Transactions on Learning Technologies*, Vol. 14, No. 5, 2021.

Summary

This chapter presented a market analysis and literature review. Together, these support three findings. First, cognitive models have not been sufficiently explored but offer advantages beyond other approaches that use ML or heuristics to adapt training. Second, the cognitive model PPE has been demonstrated across a wide variety of application domains and settings. Finally, the potential to use cognitive models to deliver personalized training is especially well established in language learning.

Although it was beyond the scope of this report to evaluate alternative approaches to personalized learning from EDM, many of these methods show considerable promise in basic and applied settings as well. Future work should elucidate conditions that favor each class of methods.

Chapter 3. Validating a Cognitive Model of Knowledge Acquisition

Of the cognitive models considered in the previous chapter, PPE stands out in terms of the number of domains and settings where it has been applied. As discussed in Chapter 1, we choose to focus on linguistic training, specifically the acquisition and retention of foreign language vocabulary. Several cognitive models considered in the previous chapter, including PPE, have been used to track item-level proficiency and to deliver personalized vocabulary reviews in laboratory, classroom, and field studies.

In this chapter, we evaluate PPE using the publicly available second language acquisition modeling (SLAM) dataset from the U.S. language-learning website Duolingo.[46] This dataset was from a diverse set of learners who completed activities in a self-paced manner, on personal devices, at their preferred times and locations. This makes the data highly generalizable. Evaluating PPE under such conditions is an important step because it constitutes an advance from component validation in a laboratory environment (TRL 4) to prototype demonstration in a relevant environment (TRL 6).

This study also presents a statistical approach, Bayesian hierarchical modeling, to calibrate PPE to account for individual differences in rates of learning and forgetting.[47] This approach optimally pools information across individual students and groups of students to estimate model parameters. As more data about the individual become available, model predictions become increasingly personalized. This is especially helpful for selecting content at the outset of training when relatively little is known about a student.

Descriptive Analyses of Second Language Acquisition Modeling

Data Description

SLAM contains more than two million words and answers from more than 6,000 students during their first 30 days of using Duolingo in 2018. Students were categorized according to three language tracks: native French speakers learning English, native Spanish speakers learning English, and native English speakers learning Spanish. Students answered three types of assignments: *translate* required translating a word from one's native language (L1) to the second

[46] Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani, "Second Language Acquisition Modeling," *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, June 2018.

[47] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, *Bayesian Data Analysis*, 2nd ed., Chapman and Hall, 1995.

language (L2); *select* required selecting the correct translation from a word bank; and *listen* required translating a spoken word from L2 to L1. All questions involved a single vocabulary item. SLAM contains student- and item-level information on study history and performance (i.e., response accuracy), as well as metadata descriptors, including question format, language, and part of speech. Our analysis excluded students who were not exposed to at least three unique words three or more times—that is, students who dropped out soon after beginning to use Duolingo. The remaining dataset included 6,447 unique users.

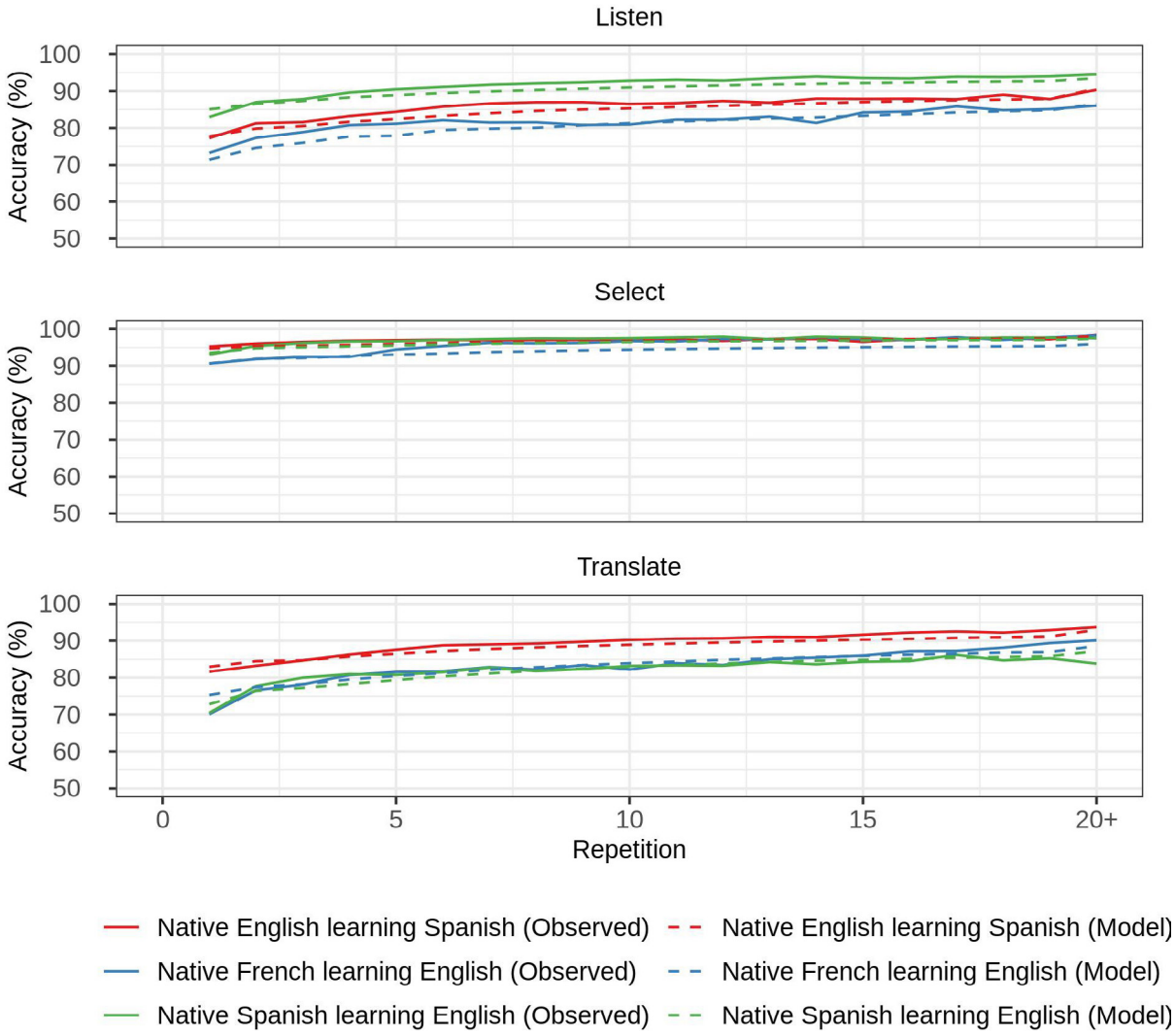
Student Performance

A key finding from learning-science research is that performance improves with the amount of practice. The curves marked "Observed" in Figure 3.1 show how the average user's response accuracy changed based on the number of times they had studied the word. The curves marked "Model" represent the outputs of the cognitive model, PPE, as described later in this chapter.

Observed accuracy increased with the amount of practice; initial practice produced the largest gains, and subsequent practice produced additional gains but at a diminishing rate. Accuracy also varied by question type; it was higher for questions that involved recognizing the correct translation from an answer bank relative to questions that involved generating correct translations from memory. Finally, accuracy varied by language track; it was lower for native French speakers learning English relative to the other tracks.[48]

[48] The regression analysis contained in the Appendix reports the sizes of these effects.

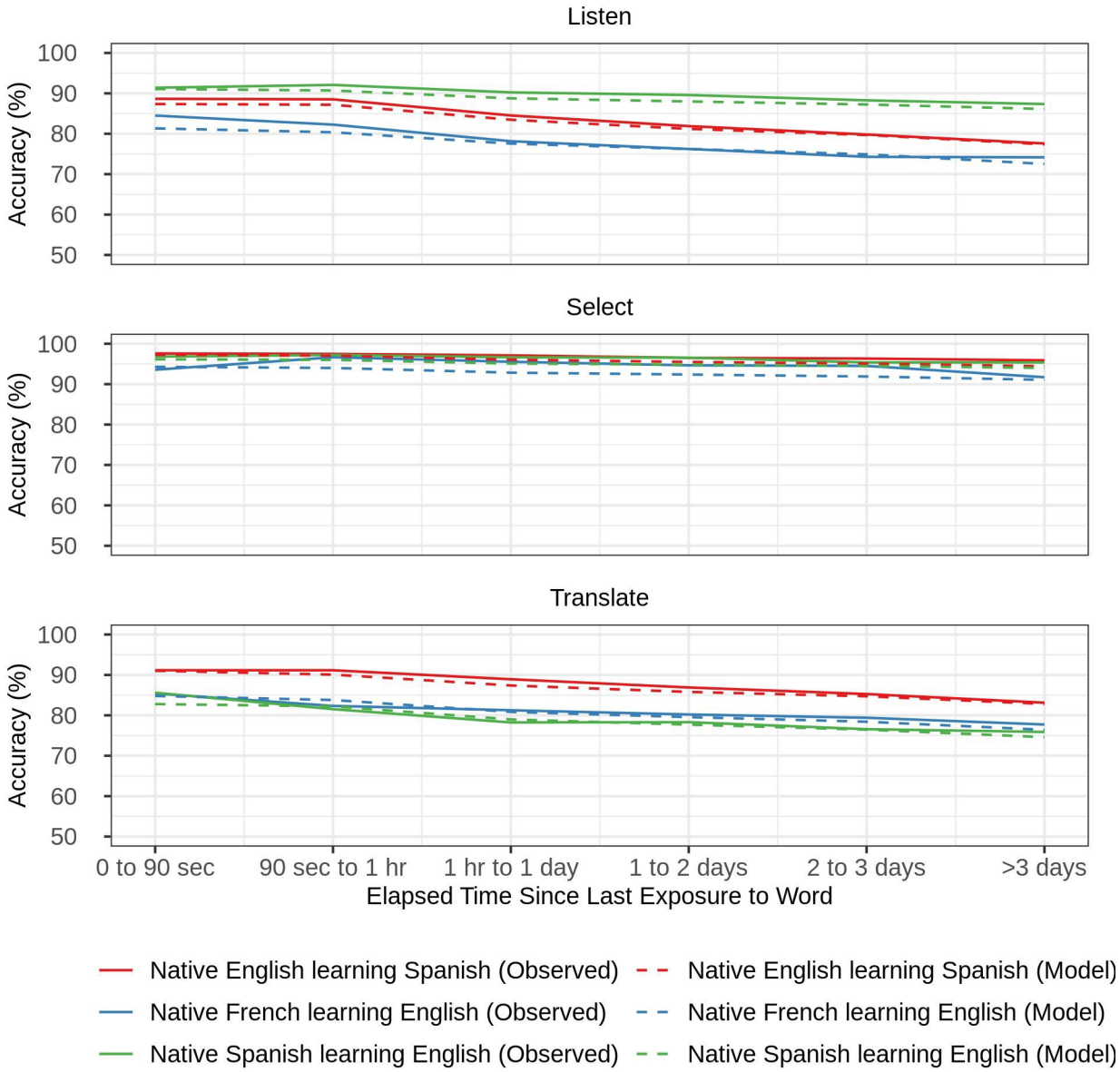
Figure 3.1. Second Language Acquisition Modeling Learning Curves by Question Type and Language



Another key finding from learning-science research is that performance decreases with time elapsed since practice occurred. In SLAM, the elapsed time since previous exposure to a word is a continuously varying value. To visualize the data, we combined elapsed times into bins from minutes to hours to days (displayed in Figure 3.2). The solid lines marked "Observed" show how the average user's response accuracy changed based on the time since they had last studied the

word. Accuracy decreased with the length of the retention interval and varied by question type and language.[49]

Figure 3.2. Second Language Acquisition Modeling Forgetting Curves by Question Type and Language



[49] The regression analysis contained in the Appendix reports the sizes of these effects.

Application of the Predictive Performance Equation to Student Performance

Model Description

PPE is a cognitive model that accounts for the effects of three sets of factors on knowledge acquisition and retention: (1) amount of practice, (2) elapsed time since practice occurred, and (3) the distribution of practice across time. PPE has previously been evaluated alongside other models of knowledge acquisition and retention and has been shown to provide an equally good or better account than other models across many datasets. The Appendix contains a technical description of PPE.

Predicted Performance

We fitted PPE to the SLAM data using maximum likelihood estimates (MLE). This involves finding values for model parameters to maximize the correspondence between model outcomes and observed performance (see the Appendix). The dotted lines marked "Model" in Figures 3.1 and 3.2 shows PPE's average performance by question type, language, number of practice repetitions, and elapsed time since the previous repetition occurred. *Model outputs reproduced the patterns observed in the SLAM dataset*—performance increased with number of repetitions, and it decreased with the time elapsed since the previous repetition. In addition, the model reproduced the observed effect of question type and language.

Evaluation of Model Variants

One parameter in PPE shifts overall level of performance up or down. For the fits displayed in Figures 3.1 and 3.2, we allowed the parameter to vary by question type to capture differences in intrinsic difficulty. For example, retrieving the correct translation from memory is more difficult than selecting it from a word bank. Other parameters in PPE affect rate of learning and forgetting. For the fits displayed in Figures 3.1 and 3.2, we allowed these parameters to vary by language (but not by question type). To determine whether this was appropriate, we fitted variants of the model where parameters corresponding to those effects were held constant or allowed to vary. The results of this analysis are presented in the appendix (Table A.1), and *they provide decisive support for allowing parameters to vary by question type and language.* The implication is that PPE is a valid model for multiple languages, but it must be calibrated separately for each. In addition, question type must be included when fitting PPE.

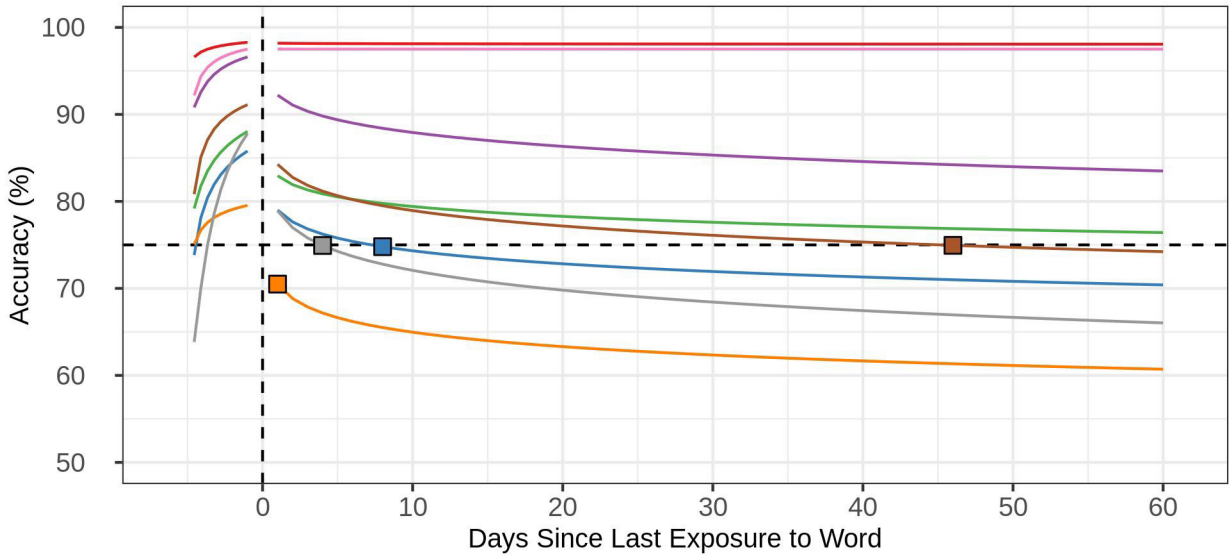
Individual Differences

For the PPE fits displayed in Figures 3.1 and 3.2, a single set of parameter values was estimated for all students. This assumes that the efficiency and effectiveness of cognitive processes, as captured by model parameters, do not vary by individual. To determine whether

this was the case, we refitted PPE at the level of individuals.[50] The results of this analysis are presented in the Appendix, and *they provide decisive support for allowing parameters to vary by individual*. The implication is that PPE should be calibrated separately for each student to enable valid and reliable personalization.

Figure 3.3 illustrates the practical significance of individual differences. The figure shows expected retention across 60 days after first studying a term, where each curve corresponds to one of eight different individuals from the SLAM dataset. Retention levels are predicted using individual-level parameter estimates from PPE. All individuals exceed a hypothetical standard of 75 percent accuracy immediately after studying a term. However, given the vastly different decay rates, some individuals are predicted to drop below the standard within the first ten days, others drop below the standard from ten to 60 days, and still others remain above the standard for more than 60 days.

Figure 3.3. Retention Predictions for Eight Individuals from the Second Language Acquisition Modeling Dataset

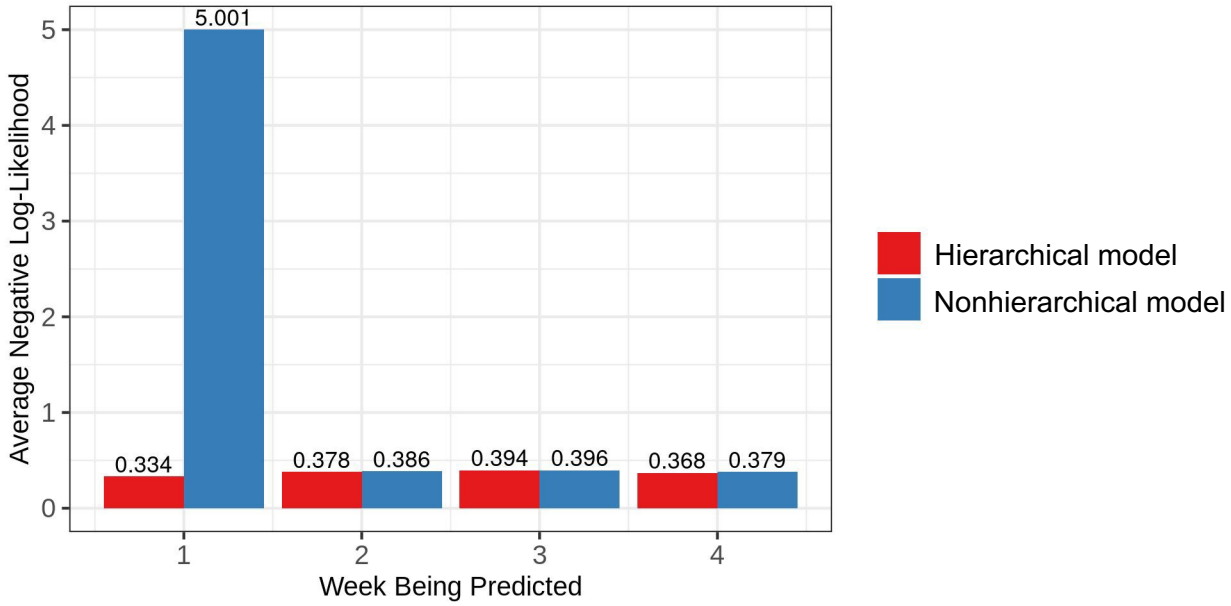


NOTE: Horizontal dotted line shows a 75 percent standard, and filled points show when the model predicts that retention will first fall below that standard. The solid lines represent eight individual retention predictions.

[50] For the 6,445 individuals, the average number of observations per individual equaled 886. Observations corresponded to different words repeated different numbers of times and spread across different retention intervals. The most common question format on average was translate (345 observations), followed by listen (302 observations), and then by select (239 observations). Each language track contained different individuals.

A challenge when fitting a model like PPE to individual students is that little is known about a student when they first begin. Bayesian hierarchical models offer a way to use data from previous students to inform model parameter estimates.[51] To demonstrate the utility of this approach, we fitted a Bayesian hierarchical model using data from a random sample of 30 English-speaking students learning Spanish (see the Appendix). Figure 3.4 shows how well the Bayesian hierarchical model accounts for the performance of a different random sample of students. Higher values on the y-axis (corresponding to the negative log-likelihood of the observed responses given the model's predictions) denote larger model error. For comparison, outcomes from a model that only uses individuals' data are shown as well. At the onset of training, the Bayesian hierarchical model gives a vastly better account of students' performance. As more data are collected about individuals in later weeks, both models do a good job of accounting for student performance. This example demonstrates the value of Bayesian hierarchical models to bootstrap predictions at the outset of training using data from previous students.

Figure 3.4. Comparison Bayesian Hierarchical and Nonhierarchical Models



[51] Specifically, a Bayesian hierarchical model learns about the structure of individual differences, expressed as prior distributions. As more data are gathered about an individual, these data might gradually shift parameter estimates (i.e., the posterior distributions) away from the priors.

Practical Considerations for Using Bayesian Hierarchical Models in the Field

Bayesian hierarchical models have often taken on the order of hours to fit for a group of students. However, once the models have been fitted, the information they provide in the form of prior probabilities for parameter values can be immediately incorporated into parameter estimates for new students. Additionally, they can produce performance predictions in real time. Thus, while there is an upfront cost to using Bayesian hierarchical models, they can improve predictions on a session-by-session basis.[52]

Summary

This chapter evaluated PPE, a cognitive model of knowledge acquisition and retention, using a dataset gathered under naturalistic conditions. The evaluation supports three primary findings. First, PPE offers a valid account of learning and retaining foreign language vocabulary. Second, individuals differ in terms of the effectiveness and efficiency of cognitive processes, as represented by model parameters. PPE might be calibrated to individuals to account for these differences. Finally, Bayesian hierarchical models provide a way to combine information from previous students along with new students to calibrate PPE more reliably.

PPE is an exemplar from a class of methods developed in cognitive science and EDM research. Although PPE provides a compelling account of performance in the Duolingo SLAM dataset, future research should continue to explore task characteristics and environment conditions that favor selecting or combing alternate methods.

[52] Bayesian hierarchical models have also been fitted in real-time to prescribe specific items. For example, see Woojae Kim, Mark A. Pitt, Zhong-Lin Lu, Mark Steyvers, and Jay I. Myung, "A Hierarchical Adaptive Approach to Optimal Experimental Design," *Neural Computation*, Vol. 26, No. 11, 2014.

Chapter 4. Case Study Using Arabic Curriculum in Universal Curriculum Assessment Tool

This chapter presents a case study on using cognitive models to develop an empirically grounded training curriculum and deliver personalized training to DAF linguists. The AETC LN program seeks to accelerate the rate at which airmen can achieve competency in a foreign language by using new and existing methods from learning science along with training technologies such as virtual reality, gamification, and AI. Although the goals of LN and traditional curricula delivered through the Defense Language Institute (DLI) are fundamentally the same, LN seeks to use updated and modern learning environments to "decouple the time component from the competency component to produce skilled language analysts for the mission."[53] The case study reported in this chapter illustrates some of the opportunities and challenges of using cognitive models to meet the training goal. Although the case study involves the LN Standard Arabic Basic course, the steps and findings reported are generalizable.[54]

The LN Standard Arabic Basic course uses the DLI's UCAT, which is a learning platform that allows instructors to develop curricula, deliver learning material, build student assessments, and track student progress over time. We assessed the structure of the LN Standard Arabic Basic course and the data available from UCAT to identify opportunities, potential drawbacks, and requirements for integrating cognitive models at the individual student and classroom levels. The box lists the requirements for model integration with UCAT for the LN Standard Arabic Basic course. We discuss each of these requirements in turn.

Capabilities Required to Integrate a Cognitive Model with the Universal Curriculum and Assessment Tool

1. Validate cognitive model.
2. Extract, process, and evaluate curriculum data in UCAT modules.
3. Extract student data in UCAT to track proficiency over time.
4. Recommend student activities using the model.
5. Deliver activities to students or instructors.

[53] Tamara Cario, "Air Force Takes a Novel Approach to Cryptolinguist Training," Defense Language Institute Foreign Language Center, October 28, 2019.

[54] The analyses reported in the previous chapter did not include Arabic because the SLAM dataset does not contain an English-to-Arabic track. Future work is needed to confirm that the findings also hold for Arabic.

Capability 1. Validated Cognitive Model

Before integrating a cognitive model with a training design, the model and the statistical methods to calibrate it must be validated. PPE and the corresponding statistical methods for calibration have previously been validated in laboratory settings for the acquisition and retention of psychomotor skills and factual knowledge, including foreign language vocabulary. Chapter 3 shows that PPE also extends to the acquisition and retention of foreign language vocabulary under naturalistic conditions that are representative of language learning using an online platform.

Assessment of Capability

Collectively, these results establish a body of support for applying PPE to the LN Standard Arabic Basic course. There are two caveats, however. First, although PPE accounts for dynamics of English, Spanish, and French language learning, PPE has not yet been validated for Arabic specifically. Although it is reasonable to expect that PPE will also provide a valid account of Arabic language learning, this has not yet been shown. Second, PPE has been applied to one aspect of language learning—the acquisition and retention of task-critical vocabulary. PPE has not yet been validated for other linguistic constructs or for speech production, and so it is most relevant to the vocabulary portions of the LN Standard Arabic Basic course.

Capability 2. Extract, Process, and Evaluate Curriculum Data in Universal Curriculum and Assessment Tool Modules

The 64-week LN Standard Arabic Basic course consists of a mix of classroom and independent study time at a 5:1 ratio. The course is tailored to military students and emphasizes listening, reading, and speaking skills.[55] The course consists of 14 units, each with multiple modules that students move through individually and in classroom activities. These modules are accessible online. Different modules contain different types of activities.

The course consists of basic course modules and assessment modules. Basic course modules contain the bulk of the learning material, which students can access in three ways: (1) individually outside a classroom setting, (2) individually in a classroom setting, and (3) collaboratively in a classroom setting. Assessment modules are assigned by instructors for specific dates and times to determine whether students learned and retained the material. Although students move through assessment modules in a linear fashion, they do not necessarily cover all parts of the basic course modules or progress through basic course modules linearly.

[55] Defense Language Institute Foreign Language Center, "Linguist Next Modern Standard Arabic Program Syllabus 2021," 2021.

Item Frequency and Activity Types Across the Curriculum

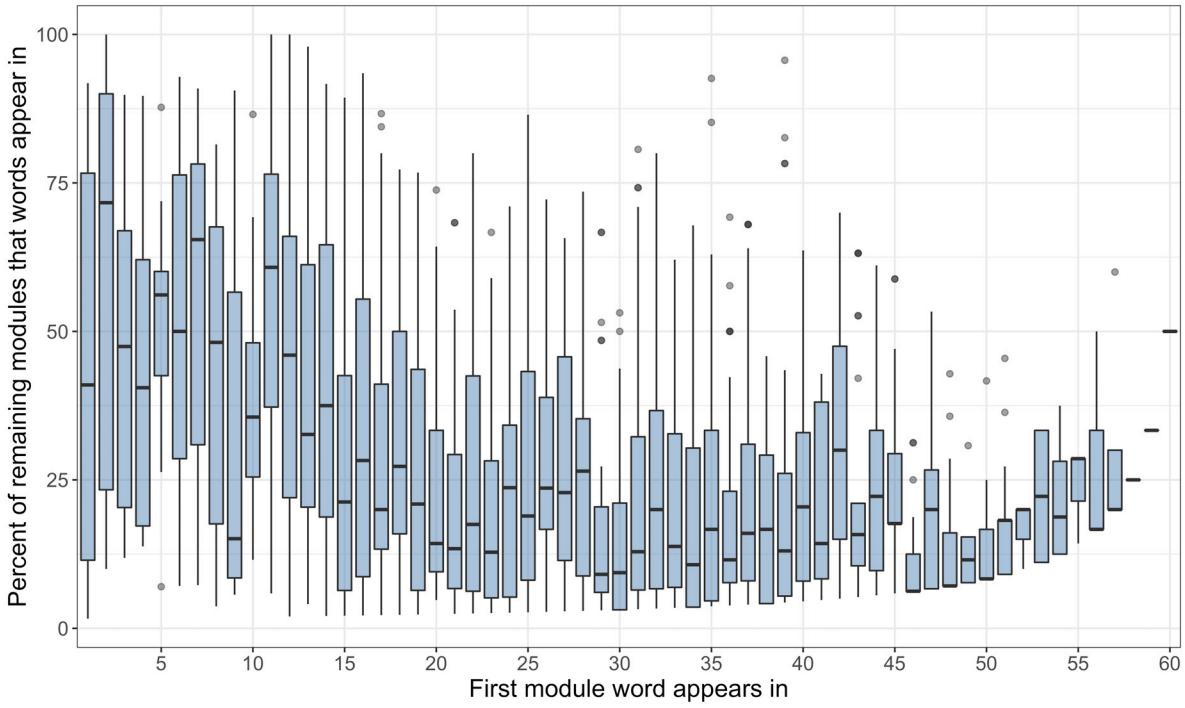
As discussed in Chapter 3, the key determinants of memory performance are number of repetitions, elapsed time since those repetitions occurred, and how the repetitions are spread out over time. Thus, the repetition of words across the curriculum is an important input to PPE and a potential curriculum design parameter that PPE could be used to optimize.

We used data from UCAT to characterize the frequency of words across the 61 basic course modules in the LN Standard Arabic Basic course.[56] Figure 4.1 displays the percentage of the remaining modules that include each term after its first appearance. For example, the first column in Figure 4.1 corresponds with terms that appear in the first module. The boxplot shows the percentage of the remaining 60 modules that these terms appear in, displayed as a range. Note that this figure accounts only for items in the LN Standard Arabic Basic lexical item spreadsheet and not all Arabic words that appear in the curriculum.[57]

[56] This process involved parsing JavaScript Object Notation (JSON) objects for basic course modules pulled from the UCAT service, parsing html within those objects, tokenizing the words within the relevant fields of each module, searching for all variants of each token across modules (e.g., searching for versions with and without short vowels), and creating a dataset of word frequency across all modules. We limited the search to terms identified in UCAT's spreadsheet of lexical items. Note that this list is not exhaustive of words appearing in the LN Standard Arabic basic course. For example, many of the Arabic terms appearing in responses to questions in the assessment modules do not appear in the lexical item spreadsheet. Parsing JSON objects and embedded HTML elements requires extensive manual effort. The LN curriculum should permit access to curriculum data in a fashion that does not require such manual processing.

[57] Words in the lexical item spreadsheet are specifically added by instructors and constitute "task-critical" vocabulary.

Figure 4.1. Percentage of Remaining Modules That Vocabulary Words Appear in, by Module of First Appearance



NOTE: The lower and upper edges of each bar capture the middle 50 percent of observations, the horizontal lines show the mean of the observations, and points show outliers.

Repeating words multiple times across the curriculum and after meaningful delays (i.e., spacing) is useful for enhancing learning and long-term retention. For example, many of the words that appear for the first time within Modules 1 through 10 also appear in 50 percent or more of the remaining modules. This degree of repetition and spacing will produce a more lasting retention of those terms. Conversely, many of the words that appear after Module 10 (along with some of the words that appear in Modules 1 through 10) appear in fewer than 25 percent of the remaining modules. Presenting these terms again in later modules would lead to longer-lasting retention.

Assessment of Capability

UCAT is the primary LMS used in the LN Standard Arabic Basic course, and the activities and vocabulary contained in modules can be extracted from UCAT. Item-level histories can be used as inputs for a PPE model that predicts learning and retention. Alternatively, predictions generated by PPE can be used to optimize when items first appear and when they are repeated.

Notwithstanding this potential, the UCAT data are limited in a variety of ways. First, some of the material that is presented is not machine readable. Modules contain many pieces of audio-visual content (e.g., images, audio, and video clips), a minority of which have associated

transcripts to allow for a straightforward tracking of word exposure. Ideally, the LN Standard Arabic Basic course would ensure text is as machine-readable as possible. However, a full integration of PPE in UCAT would likely require the use of optical character recognition to identify text in images and speech-to-text to identify text in audio passages. Such technologies are available.

Second, we are unable to track extracurricular exposure to words through independent study and consumption of other Arabic-language materials. LN students use other technologies and media to practice outside the classroom, including NetProf, Maya Lingo VR, Memrise, Quizlet, and a variety of other online or physical media such as YouTube videos or flashcards. To the extent that these exposures are delivered in applications that automatically capture exposures and student responses, these data should be captured and combined with UCAT data.

Third and finally, the interface for extracting activities and item exposures from UCAT (i.e., parsing JavaScript Object Notation objects from course modules) is not intended to support research and development efforts. Changes to the interface would be necessary to support efforts to capture data for a cognitive model.

Capability 3. Extract Student Data in the Universal Curriculum and Assessment Tool to Track Proficiency over Time

Students complete formative assessments during most UCAT modules. These come in the form of activities that call for student responses. Students also complete summative assessments at the end of each of the 14 units. These are delivered in modules dedicated specifically to assessment. Assessment modules are separate from the basic course and consist of a mix of multiple-choice, fill-in-the-blank, speaking, short answer, and transcription activities (listed from most- to least-common activity types). Only certain types of activities can be assessed automatically, at least with ease. For example, automatic assessment of multiple-choice questions is straightforward, although short answers are more challenging because a variety of responses to a given prompt might be considered correct.

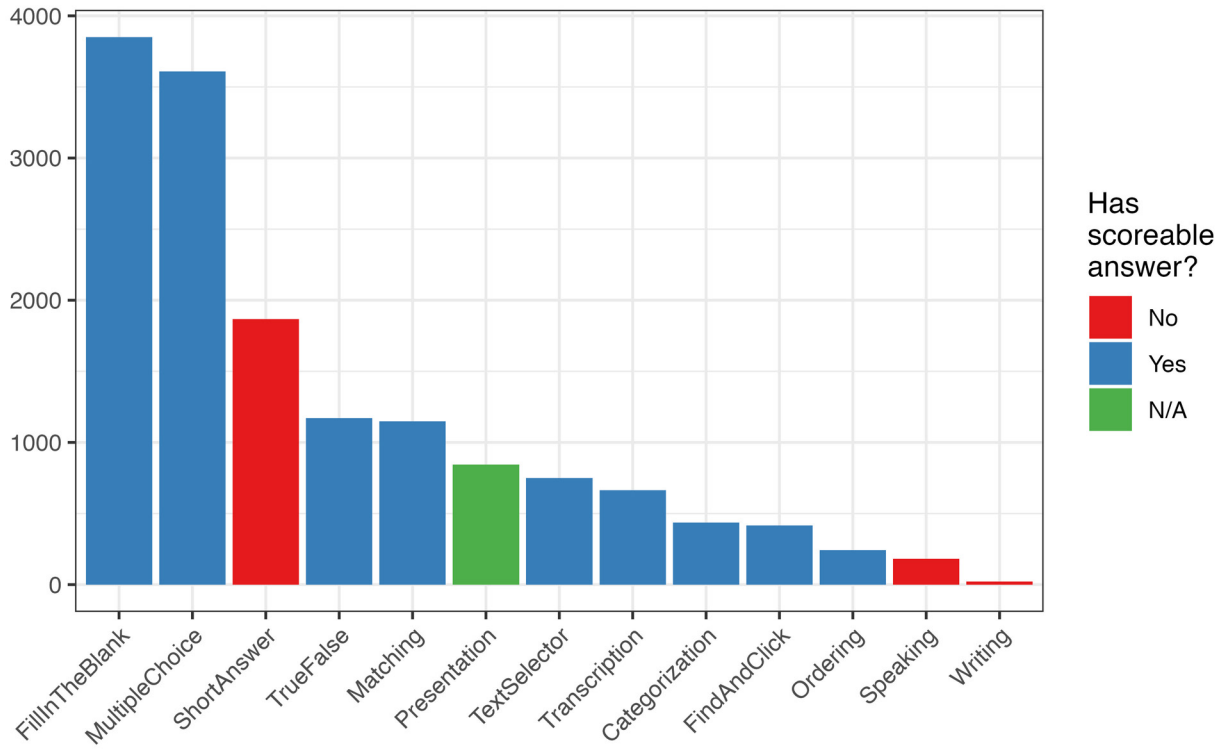
Activity Types and Student Assessment

We used data from UCAT to characterize the distribution of questions across the curriculum. This has important implications for assessing learning and retention, but also for calibrating PPE to individual students and groups of students.

Figures 4.2 and 4.3 display the count of activity types in the LN Standard Arabic Basic course and assessment modules, respectively. The figures also note which types of activities have *scoreable* answers; that is, whether student responses can be automatically assessed. The

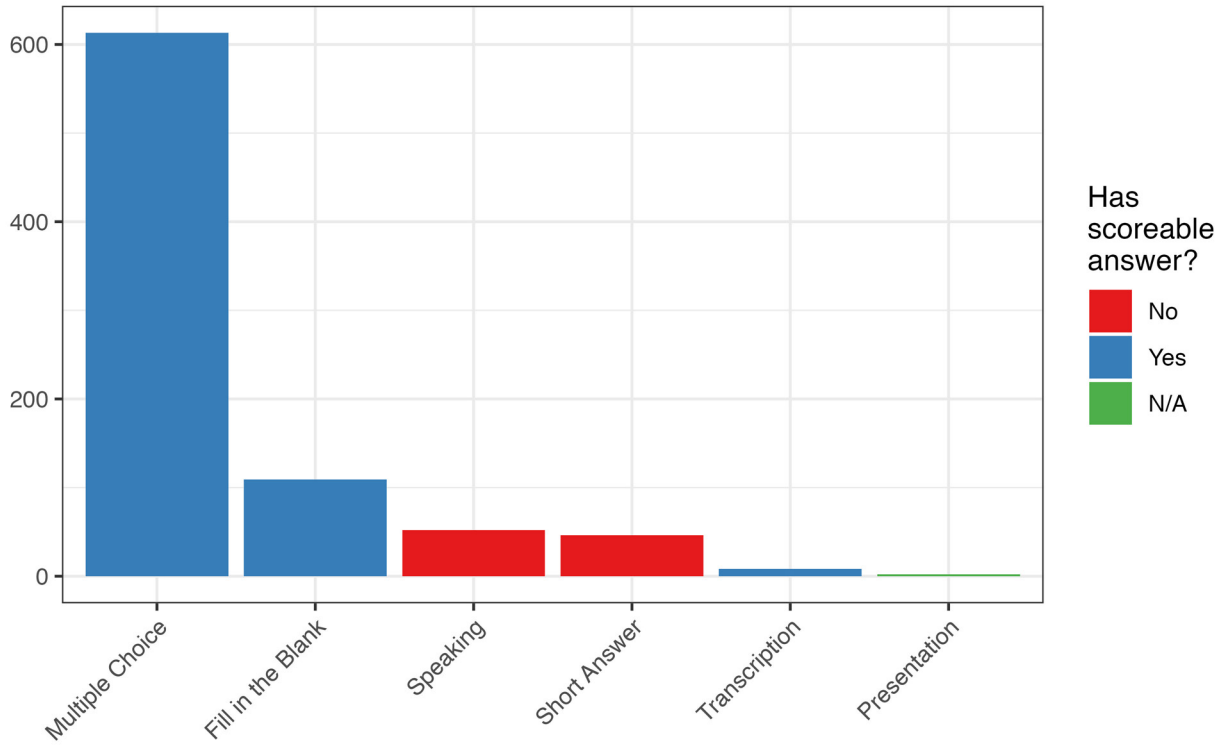
bulk of activities in both sets of modules are automatically assessable. Exceptions are the short answer, speaking, writing, and presentation activities.[58]

Figure 4.2. Count of Activity Types in Linguist Next Standard Arabic Basic Course Modules



[58] Presentation activity types present content without requiring student responses.

Figure 4.3. Count of Activity Types in Linguist Next Standard Arabic Basic Assessment Modules



Only certain activities provide information on student proficiency with a particular word. For example, many of the multiple-choice questions in the assessment modules require students to select a statement or statements in English that best characterize a passage in Arabic. In other words, multiple-choice questions provide a holistic assessment of student performance that is not tied to terms.

The activity counts shown in Figures 4.2 and 4.3 pertain only to the portion of the curriculum delivered through UCAT. For example, activities such as speaking and writing might also be delivered outside UCAT. These activities are not automatically assessable, and they create blind spots in terms of tracking students' complete study history. Nonetheless, they are an important part of the curriculum.

Assessment of Capability

Activities, assessments, and student performance data can be extracted from UCAT and, in most cases, automatically scored. These data include student responses and timestamps, two things needed by the model. In principle, these data can be used to calibrate PPE to individual students and groups of students. However, the basic course curriculum is not currently structured in a manner that easily allows this. For example, most questions provide a holistic assessment of comprehension rather than a measure of student familiarity with specific Arabic terms. Additional item-level assessments are needed to apply the model.

Summative assessments occur infrequently—at the end of the 14 units—and they contain sparse coverage of task-critical vocabulary. Formative assessments contained in activities throughout the course modules provide additional measures of student learning. However, these low-stakes assessments are not delivered in controlled conditions; students might complete them with other students and with supplementary aides. Furthermore, student responses are not part of the grades they receive, and students might not complete all activities.

Finally, formative assessments pertain to material that was just learned. These assessments measure learning rather than retention.

Capability 4. Recommend Student Activities Using the Model

The LN Standard Arabic Basic course contained in UCAT is fixed. Modules, activities, and assessments are determined in advance and, for the most part, delivered uniformly to all students. However, instructors have flexibility to tailor content delivered in the classroom, and students have flexibility to allocate attention as they determine during extracurricular study.

This section reviews options for integrating PPE in the LN Standard Arabic Basic course. The goals of integrating PPE are to (1) deliver a more empirically grounded curriculum, and (2) deliver learning content tailored to the performance of a particular training audience. There are many ways to do this, both in terms of content (i.e., what is to be delivered) and in terms of the audience (i.e., will content be tailored to individuals or to classrooms).

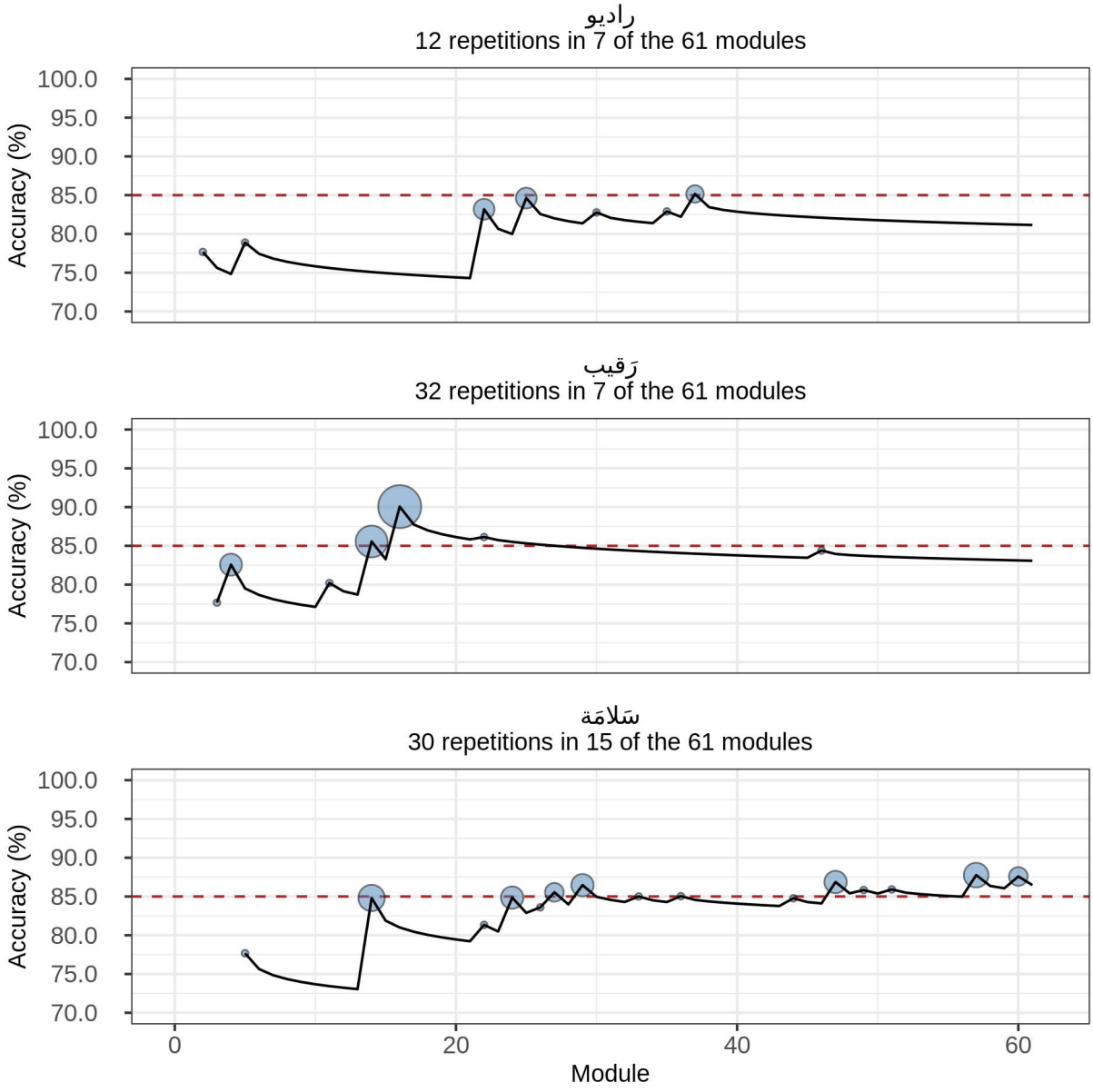
Using the Predictive Performance Equation to Enhance the Curriculum

PPE can be used to recommend practice on individual words without tuning model parameters based on individual performance histories. Figure 4.4 illustrates what the predicted response accuracy looks like for three words in the LN Standard Arabic Basic course that repeat different numbers of times and in different numbers of the 61 modules. Predicted performance increases after each repetition, denoted by the blue bubbles. The greater the number of repetitions, the higher the predicted level of performance. Alternatively, during the time between repetitions, predicted performance gradually decreases.

The top and middle panels show items that appear in the same number of modules but that repeat a different number of times. Increasing the number of repetitions (middle panel) increases the predicted level of performance. The middle and bottom panels show items that repeat a similar number of times but in a different number of modules. Spacing repetitions across more modules (bottom panel) slows the speed of learning but increases retention during later modules.

These item-level profiles can be used to find study schedules to optimize different learning and retention goals. For example, PPE can simulate how performance would be affected by distributing the same number of repetitions across the curriculum in different ways. PPE could then be used to find schedules that accelerate learning, that maximize performance during the course, or that maximize end-of-course performance.

Figure 4.4. Predicted Proficiency for an Example Word in the Linguist Next Standard Arabic Basic Course



NOTE: Bubble size is proportional to the number of word repetitions in a module.

Using the Predictive Performance Equation to Continuously Track Student Knowledge and Retention

Performance assessments provide a measure of the student's knowledge at a point in time. However, because of the dynamics of memory, knowledge is always changing. Thus,

performance assessments provide an incomplete and backward-looking view of a student's level of mastery.

By calibrating PPE using a student's historical performance data, and by providing it with the student's history of exposures to a particular item, PPE can be used to continuously track a student's level of mastery. For example, Figure 4.4 shows how memory performance declines during the time between repetitions. Item-level profiles can be generated for each student and used to continuously trace mastery during the time between assessments.

Using the Predictive Performance Equation to Tailor Content Delivered to Individual Students or to Groups of Students

Aside from reallocating item exposures across the curriculum and continuously tracking student proficiency, PPE can be used to recommend content for student groups or individuals. For example, performance profiles like the one shown in Figure 4.4 can be used to predict when performance will drop below a given threshold, at which time PPE can recommend review. Given a threshold of 85 percent, PPE would first recommend additional practice in early modules for all three items, along with additional practice for the first and second items in subsequent modules when the predicted level of performance drops below 85 percent.

PPE could be used to recommend additional practice of a word based on student exposures to that word and an *individual* student's rates of learning and forgetting, as estimated based on a student's past performance. Alternatively, instructors might wish to adapt learning content at the level of the *classroom* rather than the individual student. As discussed above, content could be delivered based on observed performance, where the model is used to deliver content when some metric—say, the average predicted accuracy of students in a classroom—crosses a threshold. Here, collective metrics would be informed by individual performance histories.

To provide this level of integration, student exposure to words and student performance on activities that require recognition of those words must be tracked. Together, these data would allow for delivery of content based on differences in the rates at which individuals learn and forget. These data would not be available at the outset of training. To bootstrap predictions and recommendations, a statistical method that pools information across individuals, like the one described in the previous chapter, is needed.

Assessment of Capability

Using PPE to enhance the curriculum requires tracking how items are distributed across the curriculum. This is possible in the LN Standard Arabic Basic course. Using PPE to track student proficiency and tailor content delivered to individual students and groups of students further requires capturing performance data to calibrate the model. As discussed, UCAT does capture performance data, but these are not generally of the form needed to calibrate the model.

Capability 5. Deliver Activities to Students or Instructors

Just as the audience for adaptive learning can be narrow or broad, learning content can be narrowly or broadly focused on proficiency goals. In the LN Standard Arabic Basic course, adaptive content could narrowly focus on individual words (e.g., if additional practice is meant to develop proficiency in a single word) or groups of related words (e.g., if vocabulary related to travel requires additional practice).

The most-targeted content delivery strategy would transmit discrete activities (such as multiple choice or fill in the blank) from a database that cue a targeted vocabulary word. This option has the benefit of directly targeting the proficiency goal and is well suited to instances where assessment on specific items is a priority.

Opening the aperture, content could be selected from a database of media that contain the vocabulary item in question. When selecting media for additional practice, it is important to avoid overloading students, so there is a premium on maximizing the value of externally delivered content. Criteria for selecting media include the following:

- *Media length*. Media should be of a length that fits within the time allocated to additional practice and not so long as to discourage completion by students. All else being equal, students might miss target items if media are not consumed in their entirety.
- *Media complexity*. Media should be appropriate to the skill level of students. Potential metrics include mean sentence length, mean word length, or number of words not previously encountered in the curriculum.
- *Occurrences of target word(s) and other related or priority words*. To maximize the value of additional practice for developing proficiency in the target word(s), media could be chosen that simultaneously provides exposure to the target word(s) alongside other related or priority words. For example, a recommendation engine could deliver content that contains terms related to travel if items requiring practice are determined to cluster around that topic. Alternatively, media might be selected to maximize the number of terms from a module that are within a certain distance from some threshold of predicted accuracy. This method would maximize the number of terms requiring additional practice rather than the thematic relatedness of terms. A combination of these methods might be employed, for example, by maximizing the thematic similarity of terms subject to a constraint on maximum predicted accuracy.
- *Word frequency outside the curriculum*. Words that are very frequent in a target language might not be high priority for additional exposure—students will encounter them repeatedly as a matter of course, so lower frequency words have higher value for choosing media.[59]

Content for additional practice could be delivered a few ways, depending on instructor goals and other constraints.

[59] It might be less important to master very low-frequency words, given that they are also less likely to be encountered in natural settings.

- *Content could be delivered to instructors via email as optional material to present or assign to students in upcoming sessions*.
- *Content could be automatically recommended to students based on real-time interactions with activities*. This method is most appropriate where individual student performance is consistently tracked across a course of instruction and requires close integration with a content recommendation engine.
- *Content could be posted online for students to access as optional, additional practice*. This method is again suitable where student performance is tracked across a curriculum, but it places looser demands on integration with a recommendation engine. For example, a recommendation engine could be queried on a daily or weekly basis to select media for additional practice.

Assessment of Capability

UCAT is not configured to vary content in real-time. This constrains how PPE can be integrated with the LN Standard Arabic Basic course. However, the instructor or students can receive tailored recommendations outside UCAT to reinforce linguistic concepts in a targeted manner.

An important enabler for delivering supplemental instruction is a content management system that contains learning events. One such option for an external recommendation engine is Haystack, a content management system for foreign language media that is built on the Apache Solr open-source search platform and is already in limited use at DLI. Haystack can be queried with English and foreign language words and phrases to deliver multimedia content that contain those words and phrases.

Haystack preprocesses uploaded files and extracts content using machine learning, optical character recognition, and other methods. Haystack can be used to select specific kinds of media (e.g., video clips), specific translation engines, media on certain topics (e.g., technology or military), and sections of media that feature certain words (e.g., trimmed video clips). An initial integration with Haystack might employ the model to select words that require additional practice and query the engine for relevant content. The existing capabilities of Haystack provide a backbone for a recommendation engine that can meet the criteria discussed above to maximize the value and relevancy of additional practice.

Summary

The structure of the LN Standard Arabic Basic course and technical design of UCAT pose difficulties for delivering tailored content through UCAT. More fundamentally, the optimal type and presentation of student activities to support the cognitive model might be in tension with the instructional design of the LN Standard Arabic Basic course. For example, small group activities and highly varied response formats might be desirable for instructors, but they pose difficulties for model integration with the LMS as it stands.

The integration of cognitive models still holds promise for reducing the amount of hand-crafting that instructors must do. Integration can be relatively light; it can use the model to deliver content when it expects proficiency to be waning without tracking and tailoring to individual performance. Alternatively, integration can be deep; it can deliver individually tailored content based on systematic tracking of performance and real-time integration with a content delivery engine. Integrating the model using an external recommendation engine can facilitate these different degrees of personalization while avoiding some of the barriers to integration with an existing LMS.

This specific use case points to broader lessons for employing cognitive models for personalized training:

- The capability to track trainee performance throughout a course of instruction opens possibilities for tailoring learning content to the individual, but it also allows instructors to understand which methods produce performance gains and can help drive the business case for personalized training methods.
- Instructor and student feedback is necessary for implementing new learning technologies. Different integrations pose trade-offs and prioritize some goals over others. It is necessary to understand which options fit with instructor plans, fall within time constraints, and engage the training audience. Continuous user feedback should therefore be incorporated throughout the development and deployment of personalized training technologies.
- Relatedly, instructors need to trust model recommendations when employing a personalized training technology. Another benefit of tracking individual performance is to help build this trust by demonstrating the effectiveness of cognitive models for developing proficiency.

Chapter 5. Conclusion

The DAF must deliver effective training and education to develop, upskill, and sustain the KSAs of its workforce. The DAF's success in training and education underlies its ability to deliver mission-capable forces to combatant commanders and maintain asymmetric advantage over competitors. Yet training and education are costly in terms of time, personnel, and resources.

The objective of this research was to examine how the DAF might leverage cognitively inspired technologies—specifically, cognitive models—to transform how it delivers training and education. Cognitive models have been used to deliver personalized training in classroom settings and engender the acquisition and retention of complex medical and military skills in diverse contexts. These applications have established that personalized training can result in higher levels of skill attainment and reduced training times. This report documents how cognitive models can be applied to linguist training, although the findings and are broadly applicable.

Key Findings

Table 5.1 summarizes the key findings of our analyses. These align to three general themes. First, industry, academia, and the government are actively developing a diverse portfolio of cognitively inspired technologies. These technologies are grounded in decades of psychology research, and they have been successfully applied in many domains and settings. The DAF training enterprise can leverage these technologies to produce a more highly skilled workforce in an economical manner.

Second, personalized training can allow individuals to achieve higher levels of proficiency in less time. Cognitive models are a relatively unexplored approach to deliver personalized training, yet they might be more suitable than other approaches that use ML or heuristic-based rules. Among cognitive models, PPE has been demonstrated in an exceptionally wide variety of domains and settings. We find that PPE is a valid model for second-language acquisition in realistic conditions, and that modern statistical techniques can be used to tailor PPE to individual students and groups of students.

Third, PPE is generally applicable to DAF training and education, and linguistic training is one high-leverage use case. The digitization of LN curricula permits use of data analytic tools, including cognitive models such as PPE. Furthermore, the ability to deploy and flexibly modify information technologies (such as a separate language-learning application) within LN curricula creates an opportunity to deliver personalized training. This might be especially effective for learning and rehearsing task-critical vocabulary. Models such as PPE can support both lighter,

supplemental integration, or more in-depth and individually targeted integration with existing curricula, depending on instructional goals and design, as well as technical limitations.

Table 5.1. Summary of Key Results

Chapter	Focus	Key Results
2	Reporting the findings of a market analysis and a review of academic literature on adaptive training technologies	• Adaptive training approaches that use cognitive models are relatively underexplored, yet they offer benefits beyond other more common training approaches. • Of the cognitive models that have been proposed, the PPE has been demonstrated in an exceptionally large number of domains and settings. • Adaptive training approaches, including cognitive models, might be delivered in a variety of mediums, including online and computer platforms, virtual and augmented reality, and games.
3	Validating the PPE cognitive model for second-language learning in realistic conditions	• PPE provides a valid account of second-language learning in real-world conditions. • PPE must be recalibrated for new languages. • A statistical method called Bayesian hierarchical modeling provides a principled way to pool information across individual students and groups of students.
4	Evaluating the feasibility of integrating a cognitive model with the LN Standard Arabic Basic course	• Of the elements covered in the LN Standard Arabic Basic course, PPE is most applicable to the acquisition and retention of task-critical vocabulary. • PPE can be used to deliver empirically grounded recommendations for when to introduce and rehearse task-critical vocabulary in the LN curriculum at the classroom level. • Student performance measures captured in the LN Standard Arabic Basic course provide information about student and classroom proficiency, yet they are not delivered in the manner needed to enable personalized training. • One promising alternative is to deliver rehearsal activities and assess student mastery of task-critical vocabulary using a separate application that complements existing IT systems used in the LN Standard Arabic Basic course but is designed specifically for cognitive model–based, personalized training.

Recommendations

Recommendation 1. The DAF should leverage cognitively inspired technologies to augment training. The learning sciences have yielded empirically grounded principles and approaches to enhance knowledge acquisition and retention. To leverage this body of work, the DAF should do the following:

- *Use service labs, federally funded research and development centers, centers of excellence, and other mechanisms to track advances in the learning sciences.* By doing so, the DAF can identify opportunities to enhance training and education using well-established and emerging technologies. Given the foundational and pervasive nature of training, these opportunities might have outsized impact relative to other technologies.
- *Conduct a campaign of experiments using cognitively inspired technologies, and capture lessons learned in a central repository.* As with other commercial products, the DAF might conduct small-scale experiments to evaluate the feasibility and utility of

transferring training technologies to the force. Given that many training challenges are common across organizations, lessons learned from these experiments should be shared DAF-wide.

Recommendation 2. The DAF should use computational cognitive models of knowledge acquisition and retention to deliver empirically based, personalized training. Cognitive models are underexplored relative to other approaches for delivering adaptive training, yet they might be more suitable and might offer significant benefits beyond other methods. Furthermore, cognitive models can be used to continuously track skill retention during periods of disuse. This is especially valuable for readiness assessment.

- *The DAF should use computational cognitive models to tailor delivery of training in multiple domains including linguist training.* Laboratory, classroom, and field studies have established the feasibility of using cognitive models for second-language learning. These form a basis for the DAF to use cognitive models for linguistic training.
- *The DAF should use computational cognitive models to track proficiency during periods of disuse.* As with training delivery, laboratory, classroom, and field studies form a basis for the DAF to use cognitive models in this way.

Recommendation 3. AFRL should invest in core and surrounding technologies to enable computational cognitive models to be applied to training at scale. In this report, we present a proof-of-concept for applying a computational cognitive model to knowledge acquisition and retention. Numerous technical advances are needed to apply the technology at scale.

- *AFRL should expand the set of factors represented in cognitive models.* PPE accounts for the effects of the amount of practice, elapsed time since practice occurred, and temporal distribution of practice. Other factors, such as stress and fatigue, also affect knowledge acquisition and retention. To enable more-accurate predictions and develop more dimensions of training, PPE should be expanded to account for the effects of additional factors.
- *AFRL should work with the DAF training enterprise to refine methods to identify KSAs that underlie task performance to permit the application of cognitive models.* To apply PPE, the underlying elements of task performance must be identified. Methods to automate and assist task decomposition by human subject matter experts would allow cognitive models to be applied to more tasks. Relatedly, methods to automatically tag training content would also allow cognitive models to be more readily applied to tasks.

Recommendation 4. AETC LN should use a separate software application specifically designed for personalized training to deliver rehearsal activities of task-critical vocabulary. The LN curriculum must satisfy multiple learning, assessment, and social objectives. Thus, learning events cannot be fully tailored to meet requirements for cognitive model–based, personalized training. A promising alternative is to track all learning activities that students complete but to limit personalized training to a separate application that complements the full curriculum.

Appendix. Technical Description of the Predictive Performance Equation and Ancillary Analysis

In this appendix, we provide a technical description of PPE and of model fitting and evaluation methods. We also provide additional supporting analyses.

Model Description

At a high level, PPE takes details related to an individual's study history for a given item. PPE converts those inputs into an activation value that represents the strength of the item in the individual's memory. The activation value determines the probability that the individual will correctly retrieve the item.

In PPE, the effects of practice and elapsed time on activation (M_n) are multiplicative,

$$M_n = N^c \cdot T^{-d}. \quad (Eq. 1)$$

where N is the number of practice repetitions, T is the elapsed time, c is the learning rate, and d is the decay rate. In PPE, activation increases as a power function of amount of practice (N^c), and activation decreases as a power function of elapsed time since practice (T^{-d}).

Elapsed time is calculated as the weighted sum of the time since each of the previous study opportunities for an item,

$$T = \sum_{i=1}^{n} w_i \cdot t_i. \quad (Eq. 2)$$

The weight assigned to each study opportunity decreases exponentially with time,

$$w_i = \frac{t_i^{-x}}{\sum_{j=1}^{n} t_j^{-x}}. \quad (Eq. 3)$$

where the variable x controls the steepness of weighting.[60]

The variable d in Equation 2 accounts for decay. Decay is calculated using the history of lags between successive study opportunities of an item (lag_j),

[60] Using past simulation studies, the value of the learning rate (c) is fixed to 0.1 and the value of x is set to 0.6 (Walsh, Gluck, Gunzelmann, Jastrzembski, Krusmark, Myung, et al., 2018).

$$d_n = b + m \cdot \left(\frac{1}{n-1} \cdot \sum_{j=1}^{n-1} \frac{1}{log(lag_j + e)} \right). \quad (Eq. 4)$$

The key innovation in PPE is the way spacing affects decay rate. This is produced by Equation 5. The quantity inside the summation approaches zero when lags are long, reducing decay rate.[61] Conversely, the quantity inside the summation approaches one when lags are short, increasing decay rate. Spaced practice extends the lags between successive item repetitions, thereby reducing an item's decay rate and increasing retention. The effects of training history are scaled by a decay slope parameter (m) and offset by a decay intercept parameter (b).

The probability of recalling an item (P_n) is a logistic function of its activation,

$$P_n = \frac{1}{1 + exp\left(\frac{\tau - M_n}{s}\right)}. \quad (Eq. 5)$$

Model Fitting Methods

We fitted the model using maximum likelihood estimation. For each question, the model predicts the probability of a correct response. By treating the observed outcome as a binomial variable, one can compute the likelihood of the outcome on question t given the model's prediction, $P(outcome_t \mid predicted_t)$. These probabilities can be multiplied together to find the likelihood of the complete set of outcomes. However, the negative value of the sum of the natural logarithm of the likelihoods, or the negative log-likelihood estimate (–LLE), is typically used instead,

$$-LLE = \sum_{t=1}^{n} log\left(P(outcome_t | predicted_t)\right). \quad (Eq. 6)$$

To fit the model, we found parameter values that minimized the –LLE, thereby maximizing the likelihood of the observed outcomes.

Supporting Analyses

To determine whether it was appropriate to allow the parameter τ to vary by question type and to allow τ and other model parameters to vary by language, we fitted four variants of the

[61] Lag prior to the first presentation of an item (lag_1) is ∞. Consequently, decay after the first presentation of an item is b (Eq. 5).

model where different subsets of parameters were held constant across conditions or allowed to vary (Table A.1). We evaluated model variants using out-of-sample predictions. This involved fitting the models by using combined data from a random split of 90 percent of students and evaluating model predictions using data from the remaining 10 percent of students excluded from the test set.

Table A.1 shows –LLE for out-of-sample predictions. Smaller values denote more-accurate predictions. The variant of the model where parameters were allowed to vary by question type and language had the lowest –LLE. The size of the difference amounts to decisive evidence in favor of that model.

Table A.1. Negative Log-Likelihoods for Models That Allow Parameters to Vary by Question Type and Language

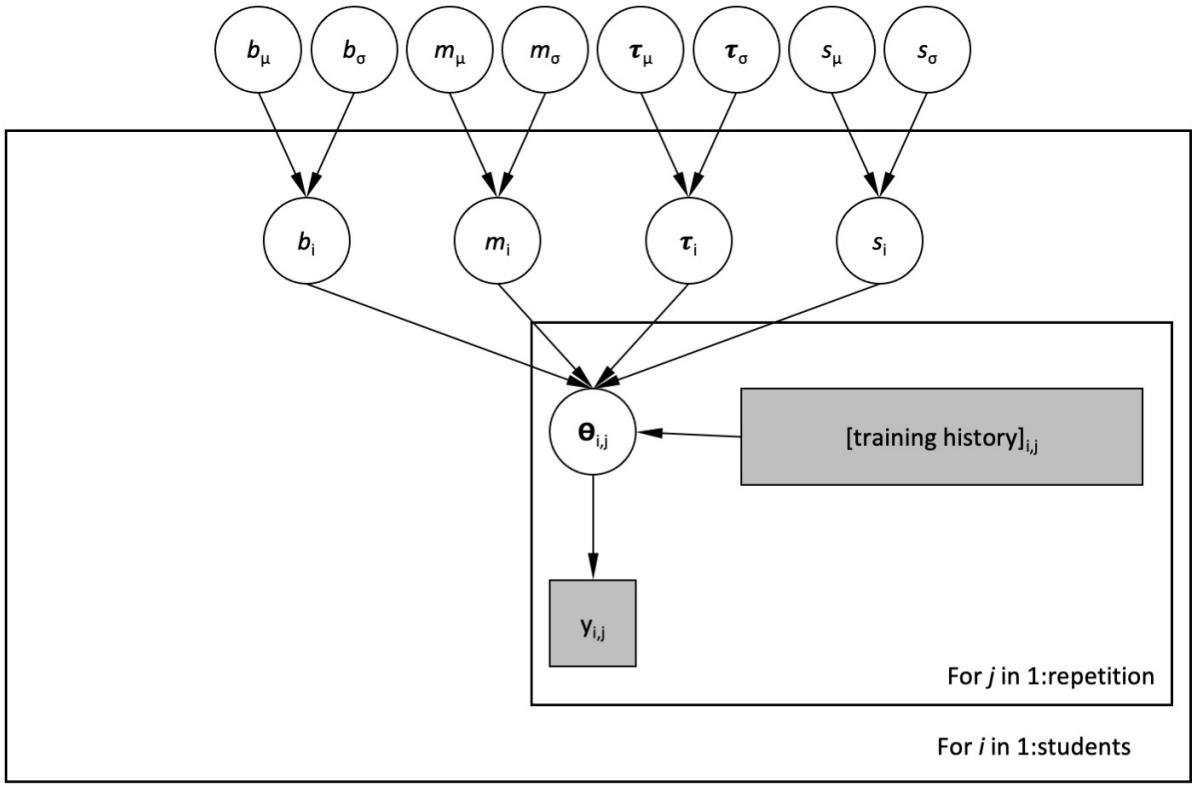
Vary Parameters by Question Type	Vary Parameters by Language	–LLE
No	No	171,698
No	Yes	171,487
Yes	No	166,172
Yes	Yes	164,490

To determine whether it was appropriate to allow parameters to vary by individual, we fitted two variants of the model: one where all parameters were shared across individuals, and another where they were allowed to vary. We fitted the models using combined data from 90 percent of the words that each student studied, and we evaluated model predictions using data from 10 percent of words held out as the test set. The –LLE for the combined and individual-level models were 213,572 and 149,735, respectively. The size of the difference amounts to decisive evidence in favor of the individual-level model.

Bayesian Hierarchical Model

To demonstrate how to calibrate PPE to new students while taking into consideration data from previous students, we implemented PPE as a generative probabilistic model and fitted it using fully Bayesian inference. Figure A.1 expresses the model in graphical notation.

Figure A.1. Cognitive Model Expressed in Graphical Notation



We gave PPE each student's exact study history (i.e., the number and timing of questions for each vocabulary word). This is represented by the *history* node in Figure A.1, which is shaded to denote that the values are observable. The data are the student response accuracies from each question. These are represented by the *y* node in Figure A.1, which is also shaded to denote that the values are observable. PPE assumes that accuracy (the θ node) is determined by training history according to a power law of learning and a power law of forgetting. The model contains four free parameters that are allowed to vary by individual (b, m, τ, and s). Individual differences are constrained by the hyper-parameters shown outside the plates. All unshaded parameter nodes in Figure A.1 denote that the values are not observable and are estimated to maximize the probability of the observed responses. The model was implemented and fitted using Stan.[62]

To simplify matters, we fitted the models using data only from native English speakers learning Spanish. In addition, although we provided the models with students' complete history of exposures, we used response accuracy only from questions that involved translating written

[62] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, Vol. 76, No. 1, 2017.

words from L1 to L2. The model shown in Figure A.1 can be expanded to account for multiple languages and question formats.

Regression Analysis of Second Language Acquisition Modeling Data

To confirm the observations from Figures 3.1 and 3.2, we performed a logistic regression treating item-level response accuracy as the outcome, and treating question type, language, number of practice repetitions, and elapsed time since the previous repetition as predictors. Table A.2 shows the odds ratios (ORs) and 95 percent confidence intervals estimated for the effects.[63] Values less than 1.0 indicate that the factor reduces the probability of a correct response, and values greater than 1.0 indicate that the factor increases the probability of a correct response.

Accuracy was lowest the first time a word was tested (OR = 0.6), increased with the number of repetitions (OR = 1.31), and decreased with the time elapsed since the previous repetition (OR = 0.95). In addition, accuracy varied by question type, language, and part of speech (not shown).[64] This analysis confirms how response accuracy is affected by amount of practice and elapsed time since practice occurred. It also demonstrates that secondary factors, such as language and question type, can further affect performance.

Table A.2. Odds Ratios and Confidence Intervals from Logistic Regression Analysis of Second Language Acquisition Modeling

Factor	Odds Ratio	Lower Bound	Upper Bound
(Intercept)	5.83	5.74	5.92
First presentation	0.60	0.59	0.61
Number of previous repetitions (log)	1.31	1.31	1.32
Elapsed time since previous repetition (log)	0.95	0.95	0.95
Language: French speakers learning English[a]	0.70	0.69	0.70
Language: Spanish speakers learning English[b]	0.99	0.99	1.00
Question type: select[c]	3.58	3.55	3.61
Question type: translate[d]	0.90	0.90	0.91

[a] Relative to base level of native English speakers learning Spanish.
[b] Relative to base level of native English speakers learning Spanish.
[c] Relative to base level of Listen Translation.
[d] Relative to base level of Listen Translation.

[63] Effects are significant at the $p < 0.05$ level if the 95 percent confidence interval does not include 1.0.

[64] Although not shown in Table A.2, accuracy varied across the 17 different parts of speech contained in the dataset.

Abbreviations

ACT-R	Adaptive Control of Thought—Rational
AETC	Air Education and Training Command
AETC/A9	Air Education and Training Command Analysis and Innovation Directorate
AFRL	Air Force Research Laboratory
AFRL/RH	Air Force Research Laboratory, Airman Systems Directorate
AI	artificial intelligence
ARTS	Adaptive Response-Time-Based Sequencing
BKT	Bayesian knowledge tracing
COA	course of action
COLT	Colorado Optimized Language Tutor
CSAF	Chief of Staff of the Air Force
CY	calendar year
DAF	United States Department of the Air Force
DASH	Difficulty, Ability, Student History
DAS3H	item Difficulty, student Ability, Skill, and Student Skill practice History
DLI	Defense Language Institute
DoD	U.S. Department of Defense
DTIC	Defense Technical Information Center
EDM	educational data mining
FY	fiscal year
GPE	General Performance Equation
IT	information technology
KSA	knowledge, skill, and ability
L1	A student's native language
L2	A student's second language
LKT	logistic knowledge tracing
LMS	learning management system
LN	Linguist Next
–LLE	negative log-likelihood estimate
MCM	multi-scale context model
ML	machine learning
MLE	maximum likelihood estimates
NSF	National Science Foundation

O&M	Operations and Maintenance
OR	odds ratio
P&A	Pavlik and Anderson
PASS	Personalized Adaptive Scheduling System
POM	Program Objective Memorandum
PPE	Predictive Performance Equation
PPO	Predictive Performance Optimizer
RAP	Ready Aircrew Program
RDT&E	Research, Development, Test and Evaluation
SAM	Search of Associative Memory
SBIR	Small Business Innovation Research
SLAM	second language acquisition modeling
STTR	Small Business Technology Transfer
TRL	technology readiness level
UCAT	Universal Curriculum and Assessment Tool
USAF	United States Air Force

References

Air Force Doctrine Note 1-21, *Agile Combat Employment*, Secretary of the Air Force, 2021.

Air Force Financial Management & Comptroller, *Air Force President's Budget FY23*, Secretary of the Air Force, 2022.

Air Force Instruction 36-2201, *Air Force Training Program*, Secretary of the Air Force, 2010.

Air Force Instruction 41-106, *Health Services, Air Force Medical Readiness Program*, Secretary of the Air Force, 2020.

Air Force Manual 11-2RC-135, Vol. 1, *Flying Operations*, Secretary of the Air Force, 2019.

Anderson, John R., *Learning and Memory: An Integrated Approach*, 2nd ed., Wiley, 2000.

Anderson, John R., *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, 2007.

Anderson, John R., Albert T. Corbett, Kenneth R. Koedinger, and Ray Pelletier, "Cognitive Tutors: Lessons Learned," *Journal of the Learning Sciences*, Vol. 4, No. 2, 1995.

Anderson, John R., and Christian D. Schunn, "Implications of the ACT-R Learning Theory: No Magic Bullets," in Robert Glaser, ed., *Advances in Instructional Psychology: Educational Design and Cognitive Science*, Vol. 5, Lawrence Erlbaum Associates, 2000.

Baker, Ryan S., and Kalina Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, Vol. 1, No. 1, 2009.

Bloom, Benjamin S., "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," *Educational Researcher*, Vol. 13, No. 6, June–July 1984.

Borgvall, Jonathan, Martin Castor, Staffan Nählinder, Per-Anders Oskarsson, and Erland Svensson, *Transfer of Training in Military Aviation*, Swedish Defence Research Agency, 2007.

Brown, Charles Q., Jr., *CSAF Action Orders to Accelerate Change Across the Air Force*, Chief of Staff, U.S. Air Force, December 2020.

Cario, Tamara, "Air Force Takes a Novel Approach to Cryptolinguist Training," Defense Language Institute Foreign Language Center, October 28, 2019. As of September 1, 2022: https://www.dliflc.edu/air-force-takes-a-novel-approach-to-cryptolinguist-training/

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, Vol. 76, No. 1, 2017.

Cooley, Tom, and Ivar Oswalt, "Operationalizing Artificial Intelligence in Simulation Based Training," *Proceedings of Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2021.

Corbett, Albert T., and John R. Anderson, "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," *User Modeling and User-Adapted Interaction*, Vol. 4, No. 4, 1994.

Defense Language Institute Foreign Language Center, "Linguist Next Modern Standard Arabic Program Syllabus 2021," 2021.

Doroudi, Shayan, Vincent Aleven, and Emma Brunskill, "Where's the Reward? A Review of Reinforcement Learning for Instructional Sequencing," *International Journal of Artificial Intelligence in Education*, Vol. 29, No. 4, 2019.

Dunlosky, John, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham, "Improving Students' Learning with Effective Learning Techniques: Promising Directions from Cognitive and Educational Psychology," *Psychological Science in the Public Interest*, Vol. 14, No. 1, 2013.

Eglington, Luke G., and Philip I. Pavlik, Jr., "Optimizing Practice Scheduling Requires Quantitative Tracking of Individual Item Performance," *npj Science of Learning*, Vol. 5, No. 1, 2020.

Eglington, Luke G., and Philip I. Pavlik, Jr., "How to Optimize Student Learning Using Student Models That Adapt Rapidly to Individual Differences," *International Journal of Artificial Intelligence in Education*, 2022.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, *Bayesian Data Analysis*, 2nd ed., Chapman and Hall, 1995.

Grant, Maria J., and Andrew Booth, "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies," *Health Information & Libraries Journal*, Vol. 26, No. 2, 2009.

Hattie, John, John Biggs, and Nola Purdie, "Effects of Learning Skills Interventions on Student Learning: A Meta-Analysis," *Review of Educational Research*, Vol. 66, No. 2, 1996.

Hull, Clark Leonard, *Principles of Behavior: An Introduction to Behavior Theory*, Appleton-Century, 1943.

Jastrzembski, Tiffany S., Kevin A. Gluck, and Stuart Rodgers, "Improving Military Readiness: A State-of-the-Art Cognitive Tool to Predict Performance and Optimize Training Effectiveness," *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2009.

Jastrzembski, Tiffany S., Kevin A. Gluck, and Stuart Rodgers, "The Predictive Performance Optimizer: An Adaptive Analysis Cognitive Tool for Performance Prediction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53, No. 21, 2009.

Kim, Woojae, Mark A. Pitt, Zhong-Lin Lu, Mark Steyvers, and Jay I. Myung, "A Hierarchical Adaptive Approach to Optimal Experimental Design," *Neural Computation*, Vol. 26, No. 11, 2014.

Knowlton, Barbara J., Jennifer A. Mangels, and Larry R. Squire, "A Neostriatal Habit Learning System in Humans," *Science*, Vol. 273, No. 5280, 1996.

Kornell, Nate, and Bridgid Finn, "Self-Regulated Learning: An Overview of Theory and Data," in John Dunlosky and Sarah K. Tauber, eds., *The Oxford Handbook of Metamemory*, Oxford University Press, 2016.

Landsberg, Carla R., Randy S. Astwood, Jr., Wendi L. Van Buskirk, Lisa N. Townsend, Natalie B. Steinhauser, and Alyssa D. Mercado, "Review of Adaptive Training System Techniques," *Military Psychology*, Vol. 24, No. 2, 2012.

Marcus, Gary, and Ernest Davis, "Insights for AI from the Human Mind," *Communications of the ACM*, Vol. 64, No. 1, 2021.

McClelland, James L., "The Place of Modeling in Cognitive Science," *Topics in Cognitive Science*, Vol. 1, No. 1, 2009.

Mettler, Everett, Christine M. Massey, and Philip J. Kellman, "A Comparison of Adaptive and Fixed Schedules of Practice," *Journal of Experimental Psychology: General*, Vol. 145, No. 7, 2016.

Park, Ok-choon, and Jung Lee, "Adaptive Instructional Systems," in David H. Jonassen and Marcy Driscoll, eds., *Handbook of Research for Educational Communications and Technology*, Routledge, 2004.

Pashler, Harold, Nicholas Cepeda, Robert V. Lindsey, Ed Vul, and Michael C. Mozer, "Predicting the Optimal Spacing of Study: A Multiscale Context Model of Memory," in Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, eds., *Advances in Neural Information Processing Systems*, Vol. 22, Neural Information Processing Systems Foundation, 2009.

Pavlik, Philip I., Jr., and John R. Anderson, "Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect," *Cognitive Science: A Multidisciplinary Journal*, Vol. 29, No. 4, 2010.

Pavlik, Philip I., Luke G. Eglington, and Leigh M. Harrell-Williams, "Logistic Knowledge Tracing: A Constrained Framework for Learner Modeling," *IEEE Transactions on Learning Technologies*, Vol. 14, No. 5, October 2021.

Raaijmakers, Jeroen G., and Richard M. Shiffrin, "Search of Associative Memory," *Psychological Review*, Vol. 88, No. 2, 1981.

Russell, Stuart J., and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

Settles, Burr, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani, "Second Language Acquisition Modeling," *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, June 2018.

Shute, Valerie J., and Diego Zapata-Rivera, "Adaptive Technologies," *ETS Research Report*, RR-07-05, 2007.

Simon, Herbert A., and Allen Newell, "Human Problem Solving: The State of the Theory in 1970," *American Psychologist*, Vol. 26, No. 2, 1971.

Soderstrom, Nicholas C., and Robert A. Bjork, "Learning Versus Performance: An Integrative Review," *Perspectives on Psychological Science*, Vol. 10, No. 2, 2015.

Squire, Larry R., "Memory System of the Brain: A Brief History and Current Perspective," *Neurobiology of Learning and Memory*, Vol. 82, No. 3, 2004.

Sweller, John, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science: A Multidisciplinary Journal*, Vol. 12, No. 2, 1988.

Tabibian, Behzad, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez, "Enhancing Human Learning via Spaced Repetition Optimization," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 116, No. 10, 2019.

Walsh, Matthew M., and Kevin A. Gluck, "Mechanisms for Robust Cognition," *Cognitive Science: A Multidisciplinary Journal*, Vol. 39, No. 6, 2015.

Walsh, Matthew M., Kevin A. Gluck, Glenn Gunzelmann, Tiffany S. Jastrzembski, and Michael Krusmark, "Evaluating the Theoretic Adequacy and Applied Potential of Computational Models of the Spacing Effect," *Cognitive Science: A Multidisciplinary Journal*, Vol. 42, Supp. 3, 2018.

Walsh, Matthew M., Kevin A. Gluck, Glenn Gunzelmann, Tiffany S. Jastrzembski, Michael Krusmark, Jay I. Myung, Mark A. Pitt, and Ran Zhou, "Mechanisms Underlying the Spacing Effect in Learning: A Comparison of Three Computational Models," *Journal of Experimental Psychology: General*, Vol. 147, No. 9, 2018.

Walsh, Matthew W., and Marsha C. Lovett, "The Cognitive Science Approach to Learning and Memory," in Susan E. F. Chipman, ed., *The Oxford Handbook of Cognitive Science*: Part 2*, Cognitive Science Perspectives on Classic Problems in Psychology*, Oxford University Press, 2016.

Wickens, Christopher D., Shaun Hutchins, Thomas Carolan, and John Cumming, "Effectiveness of Part-Task Training and Increasing-Difficulty Training Strategies: A Meta-Analysis Approach," *Human Factors*: *The Journal of the Human Factors and Ergonomics Society*, Vol. 55, No. 2, 2012.

Wirth, Anna Jean, Thomas Light, Daniel M. Romano, Shane Tierney, Ronald G. McGarvey, Moon Kim, Michael J. Lostumbo, Amanda Nguyen, Paul Emslie, and John G. Drew, *Evaluating Alternative Maintenance Manpower Force Structure Concepts for the F-35A*, RAND Corporation, RR-4433-AF, 2020. As of September 1, 2022: https://www.rand.org/pubs/research_reports/RR4433.html

Xiong, Xiaolu, Yan Wang, and Joseph Barbosa Beck, "Improving Students' Long-Term Retention Performance: A Study on Personalized Retention Schedules," *LAK '15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, March 2015.

PROJECT AIR FORCE

The purpose of training and education in the United States Department of the Air Force (DAF) is to develop and sustain mission-critical knowledge, skills, and abilities (KSAs) among airmen, guardians, and civilians. The DAF must deliver effective training and education to fully use its human capital, provide warfighting assets to combatant commanders, and maintain asymmetric advantage over competitors. Yet training and education is costly. A recent budget request included more than $2 billion for training and education, and recent guidance has highlighted that the U.S. Air Force must transform all facets of training and education to field a highly capable force in an affordable manner.

This report focuses on computational cognitive models, a class of training technologies with transformative potential. Computational cognitive models emulate psychological processes like knowledge acquisition and retention. These models have been used to develop empirically grounded training curricula and deliver personalized training in diverse domains. The primary benefits of using these models to deliver personalized training are enhanced learning gains and reduced training time.

This report explores the feasibility of applying computational cognitive models to the acquisition and sustainment of mission-critical KSAs, with emphasis on second-language learning. The authors affirm that cognitive models can be integrated with training curricula in a variety of ways, and each of these potential courses of action presents different levels of benefits along with different technical and logistical challenges.

$29.50

ISBN-10 1-9774-1160-6
ISBN-13 978-1-9774-1160-0

www.rand.org

RR-A1565-1