Measuring the Power of Learning.®

R&D Connections

Simulations of Thought: The Role of Computational Cognitive Models in Assessment

By Jung Aa Moon, Bridgid Finn, Michelle LaMar, and Irvin R. Katz

No. 26 • September 2018

Introduction

Many cognitive scientists, both inside and outside of Educational Testing Service (ETS), use computational models grounded in theories of cognition to better understand the mental processes of humans interacting with assessments. Researchers and assessment developers at ETS are, for example, interested in knowing if the way in which an assessment task is presented can influence how test takers react and think, and ultimately, how they answer. Such models play an important role in many aspects of assessment, including assessment development (e.g., creating test items), evaluation of validity evidence, and improving the efficiency of the raters who score test responses. Here, we highlight how computational cognitive modeling informs research on student and rater responses.

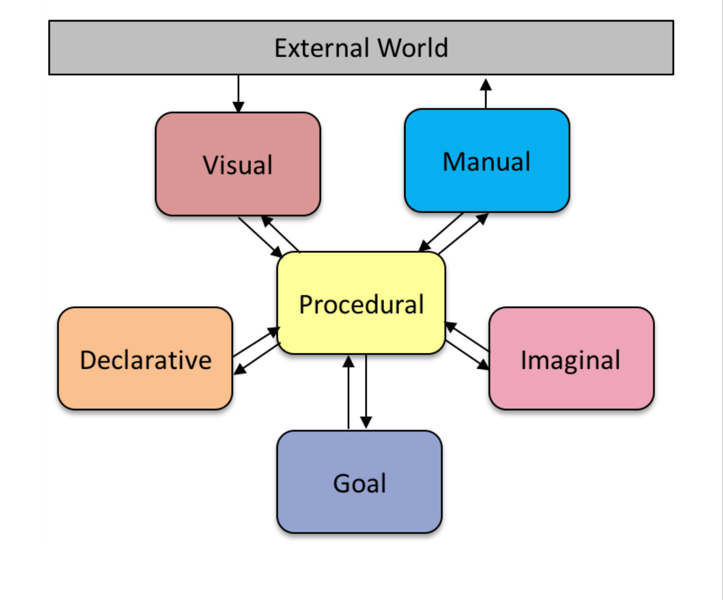
Cognitive models often refer to descriptions of mental processes in words or diagrams while *computational* cognitive models are computer-based models that use mathematical equations or programming languages to simulate mental processes and predict how humans react in specific situations. These models have different purposes and uses from psychometric models, which generally focus on scored outcomes with the goal of evaluating a performance. Not all computer programs designed to simulate human performance rely on computational cognitive models, since many only mimic surface behaviors. Computational cognitive models are grounded in and constrained by the empirical results of studies on human cognition, and so (implicitly) reflect decades of research literature and thousands of experimental studies.

[1] *Editor's note*: The authors work in ETS's Research & Development division. Jung Aa Moon is an Associate Research Scientist, Bridgid Finn is a Senior Research Scientist, Michelle LaMar is a Research Scientist, and Irvin R. Katz is a Senior Research Director. All work in the Cognitive and Technology Sciences Center.

Computational Cognitive Models

Researchers in ETS's R&D division focus on a select few of the numerous computational cognitive models that have been developed around the world. This group of models includes full cognitive architectures with multiple interacting components as well as models that simulate particular mental mechanisms (e.g., learning and forgetting, decision making). The following sections introduce three cognitive models and give examples of how they can be used to improve assessment development, evaluation of validity evidence, and scoring efficiency.

Figure 1. ACT-R cognitive architecture



a. ACT-R Cognitive Architecture

One successful cognitive architecture is the Adaptive Control of Thought – Rational (ACT-R, pronounced akt-ahr). It was developed by John R. Anderson and his colleagues at Carnegie Mellon University (Anderson et al., 2004) and is generally implemented as a programming language for modeling cognitive processes, such as learning and memory, problem solving, and decision making.

ACT-R consists of several modules that each represent specialized mental operations (see Figure 1), for instance, retrieving learned information from memory (declarative module), processing visual signals (visual module), and making motor responses (manual module). Cognition emerges from interactions among the various modules (Anderson, 2007). The specifications for the processes represented by each module, such as the time it takes to complete an operation (e.g., 400 milliseconds for typing a letter) are based on empirical findings from decades of research on human cognition.

Researchers use ACT-R to build cognitive models that represent perceptual, cognitive, and motor steps involved in performing tasks. Performance characteristics (e.g., correctness of a response, completion time) and behaviors (e.g., solution strategies, sequence of visual attention, mouse clicks) predicted by a model can be compared with characteristics and behaviors observed in human participants. Such a comparison of model predictions to data from observations of human behavior provides a quantitative measure of how closely the model represents human cognition.

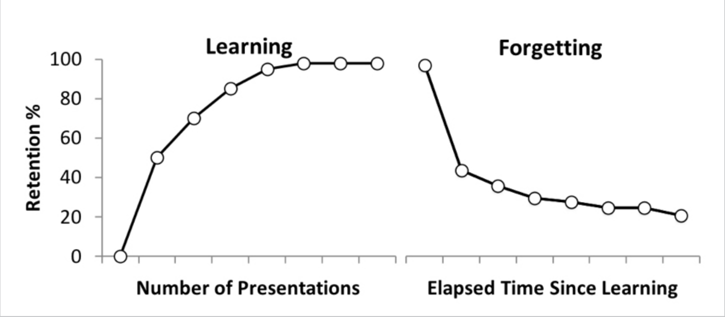
b. Markov Decision Process Cognitive Model

The Markov decision process (MDP) is a cognitive model that predicts sequential decision making in open-ended environments, such as navigating various steps to solve a problem. MDP models are built to simulate human decision making as a function of a person's goals and beliefs about the world (Baker, Saxe, & Tenenbaum, 2011). The models are designed to account for the current state of the environment as understood by the decision maker, as well as that person's predictions of how various actions can change the state of the environment. When designing the models, cognitive scientists can also include a parameter that reflects the probability of choosing an action by mistake.

c. Models of Learning and Forgetting

In 1885 the German psychologist Hermann Ebbinghaus (1885/2013) described what became known as the forgetting curve, which illustrates how humans forget what they have learned. He studied learning and forgetting based on simple word lists, but more recent research has replicated his findings for complex behavior, including that of students learning mathematics (Anderson & Schunn, 2000), adults learning cardiopulmonary resuscitation (CPR; Jastrzembski et al., 2017), and Air Force personnel practicing aircraft reconnaissance (Jastrzembski, Gluck, & Gunzelmann, 2006). Figure 2 shows the basic shape of learning and forgetting functions.

Figure 2. Typical learning and forgetting curves



The learning curve is generally described as an exponential rise to a limit, with increasing gains in performance early on, but diminishing gains with practice. Forgetting follows an exponential decay function, with the most loss occurring early and the forgetting tapering off as more time passes. This function models the forgetting that occurs if a person does not practice what has been learned. The above description is a simplification of the learning and forgetting phenomena. Cognitive scientists use more precise equations and take into account other phenomena (e.g., type of practice, the spacing of practice over time, individual differences) when modeling performance.

Using Models of Human Cognition in Assessment at ETS

Computational cognitive models can, as we mentioned previously, be used to simulate the mental processes that students use when responding to assessment items, as well as those that raters use when scoring test responses. The links between assessment items, student responses to items, and the interpretations (scoring) of those responses, all involve implicit assumptions about the associated cognitive processes. Assessment designers assume that students will interpret an item prompt in a certain way, understand the available response options, and base their responses on knowledge and skills that the test is designed to measure. Computational cognitive models can clarify these assumptions, and help researchers identify where they lack support. Such models of student and rater cognitive processes can be used for assessment design, supporting validity, and improving automated and human scoring. Below are three examples of how researchers at ETS are applying the cognitive models described above.

a. Item Format Design

Computer-administered assessments offer new ways to present items and collect responses, which have expanded the choices of item formats available to assessment developers. These new item formats may affect test takers' thinking processes and test results. Computational cognitive models may be used to study the effects of different item formats on student behavior, which in turn can inform decisions on test design. A recent ETS research study (Moon, Keehner, & Katz, 2018) sought to find out if test takers behave differently when interacting with mathematics items (Figure 3 on next page) that have the same item content, but are presented in three different formats:

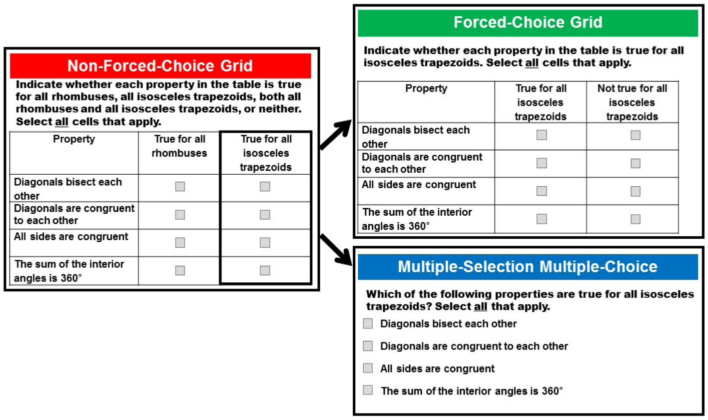
1) Non-forced-choice options (i.e., select all that apply) presented in a grid,
2) Forced-choice options (True/False) in a grid,
3) Multiple-selection multiple-choice (i.e., select all that apply).

We found that participants who were given True/False options made "True" judgments more frequently than those who did not face True/False options. They also made more frequent "True" judgments when items were presented in a grid format than when they were not.

To understand how distinct item formats produced different response patterns, we built a cognitive model (Moon, Katz, & Keehner, 2017) in ACT-R (overview of computation model shown in Figure 4), in which a simulated test taker reads the stem of a given item and then selects one of the available response options. The model simulates attempts to solve the item by retrieving relevant knowledge learned in the past (e.g., "do diagonals of rhombuses bisect each other?"). If this solution attempt is successful, the model uses the retrieved information to make a True/False judgment. If the attempt fails, the model will either make a guess, skip ahead to the next option, or make another attempt.

In this case, our computational cognitive model provided insights into how different item formats can affect test takers' choices when they are unsure of their answer. For example, by providing a "do not know" option, differences in performance due to the item format were mitigated, potentially strengthening the measurement. Earlier work at ETS (Nhouyvanisvong & Katz, 1998) demonstrated how an ACT-R model generated testable predictions about factors affecting item difficulty.

Figure 3. Target item formats



Non-Forced-Choice Grid

Indicate whether each property in the table is true for all rhombuses, all isosceles trapezoids, both all rhombuses and all isosceles trapezoids, or neither. Select all cells that apply.

Property	True for all rhombuses	True for all isosceles trapezoids
Diagonals bisect each other	□	□
Diagonals are congruent to each other	□	□
All sides are congruent	□	□
The sum of the interior angles is 360°	□	□

Forced-Choice Grid

Indicate whether each property in the table is true for all isosceles trapezoids. Select all cells that apply.

Property	True for all isosceles trapezoids	Not true for all isosceles trapezoids
Diagonals bisect each other	□	□
Diagonals are congruent to each other	□	□
All sides are congruent	□	□
The sum of the interior angles is 360°	□	□

Multiple-Selection Multiple-Choice

Which of the following properties are true for all isosceles trapezoids? Select all that apply.

□ Diagonals bisect each other

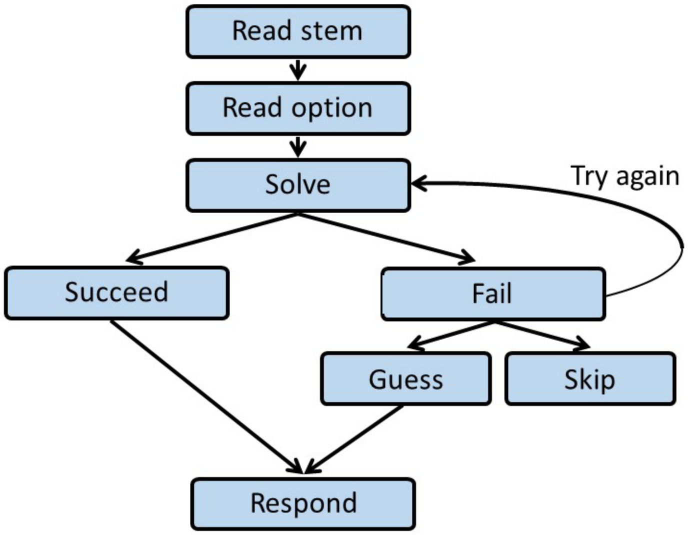
□ Diagonals are congruent to each other

□ All sides are congruent

□ The sum of the interior angles is 360°

Non-forced-choice grid (left) requires selecting all that apply from a list of options presented in a grid. Forced-choice grid (top right) presents True-False choices for each option. Multiple-selection multiple-choice (bottom right) requires selecting all that apply from a list of options. The arrows indicate that the right-hand column ("True for all isosceles trapezoids") of a non-forced-choice grid item was reformatted to a forced-choice item and a multiple-selection multiple-choice item to create content-equivalent items. The same reformatting was applied to the left-hand column ("True for all rhombuses").

Figure 4. Model of test-taker cognition



b. Validity Evidence

Computer-based assessments offer assessment developers new insights into test taker behavior by analyzing *response process data*. Comparisons between predictions based on computational cognitive models and observations of behaviors during an assessment can provide validity evidence about the test takers' response process (Kane & Mislevy, 2017).

ACT-R models can be used to predict the amount of time a student spends on various response steps, including interpreting the item (i.e., reading the prompt), evaluating the options, retrieving needed information from memory, and marking a response. The model's predictions take into account factors such as students' knowledge level, cognitive capacity, and motivation. The student's interactions with an item as reflected in response process data (e.g., number and location of mouse clicks, changes in an onscreen slider) can reveal if the student interacts with the task as expected.

Response process data

Response process data refers to information gathered about how the test takers formulate and mark their response to an item prompt or task. For fixed response items these data can include time spent on the item or changes to the selected response. For interactive tasks, these data can show the steps that the student took to solve the problem and achieve their final answer.

It is also possible to build MDP models that represent one or more expected solution strategies (LaMar, 2018; LaMar, Baker, & Greiff, 2017) for tasks involving multistep problem solving. Each model can predict the probability of a student choosing an action based on the current state of the problem. These predicted probabilities can be used to determine how likely it is that actual student behavior, as reflected in response process data, conforms to the assumptions of the model. If a significant portion of the recorded performances on a task do not fit the results predicted by the MDP models, this could indicate that students are taking an unexpected approach to solving a problem, maybe by misinterpreting the task (LaMar, 2014), or using a shortcut to the solution. Any of these situations might indicate a validity problem for the item or task in question, and may suggest a redesign of the item or task.

c. Automated Identification of Solution Strategies

Research is currently underway at ETS on a computer-based science assessment that records student interactions with a simulated lab experiment (LaMar et al., 2017). The process of scientific inquiry is one of the key constructs for this new assessment. Thus, the steps students take within the simulation when responding to each prompt (e.g., collecting data to test hypotheses) potentially provide rich, critical evidence of students' skills with scientific inquiry processes. ETS's researchers designed a set of MDP models that corresponds to different strategies for inquiry that students might use while working on a task, including both productive and non-productive strategies. Given the recorded response-process data reflecting actions of actual students, different sequences of actions can be associated with strategies by selecting the model that assigns those actions to have a high probability. The goal is to identify the strategy used, as well as where and when strategy shifting occurs. This information enables automated scoring of the inquiry process, which combined with traditional item scores, could produce a more complete measure of students' science ability.

d. Simulation of Human Raters

Computational cognitive modeling is also used to simulate mental processes of people other than test takers. Researchers in ETS R&D have, for example, used models of learning and forgetting to improve our procedures for training people who rate constructed response answers. While automated scoring of constructed responses is improving, most constructed responses are still scored by human raters. The researchers investigated how the time between rating sessions impacted rater performance by fitting learning and forgetting models to rater reliability data from the *GRE*® and *TOEFL iBT*® tests (Finn, Wendler & Arslan, 2018; Finn, Wendler, Ricker-Pedley & Arslan, in press). The resulting data indicated that raters' operational performance was not affected by the number of consecutive days of scoring (i.e., practice or learning), but by gaps between scoring sessions (i.e., skipped days which lead to forgetting). These findings suggest that we should consider the consistency of scoring schedules, and not only the amount of scoring, to maintain the skills of human raters and to avoid unnecessary training.

Conclusions

Our understanding of how people think, reason, solve problems, and learn, has improved thanks to the application of computational cognitive models that are based on decades of research in cognitive science. We see great potential in the application of computational cognitive modeling to assessment practice. Research on this topic can influence how we create assessments, how we use them to provide information about test takers, how we design individual test questions and extended assessment tasks, and how we score essays and other responses constructed by test takers. Computational cognitive models offer a way to make assessment development and scoring practices more efficient. Computational models of "simulated test takers" allow us to better understand when an item response format inhibits a test taker's performance, and might suggest ways of revising test questions. Another example comes from models of "simulated raters" that help guide the timing and content of training, practice, and retraining. These models may also help us identify human raters who are more likely to retain their training after breaks of days or weeks.

Future applications of computational cognitive modeling could enhance the efficiency with which we create and administer tests. For example, models of test takers might be used to predict the difficulty of test questions. Armed with this knowledge, assessment developers could tailor questions to particular difficulty levels, rather than discovering item difficulties by administering potential questions to thousands of students in a costly and slow process. Models of student fatigue or anxiety could also suggest how to revise the testing situations (and when to give breaks) to avoid unfairly impacting student scores, and indicate what types of errors someone who is tired or feels anxious might make, as well as what can be done to minimize the effect of such errors.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe*? New York, NY: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111, 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (pp. 1–33). Mahwah, NJ: Erlbaum.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2469–2474). Mahwah, NJ: Erlbaum.

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius, Trans.). *Annals of Neurosciences*, 20, 155–156. http://doi.org/10.5214/ans.0972.7531.200408 (Reprinted from *Memory: A contribution to experimental psychology*, by H. Ebbinghaus, 1885, New York, NY: Dover)

Finn, B., Wendler, C., & Arslan, B. (2018, April). *Applying cognitive theory to the human essay rating process*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Finn, B., Wendler, C., Ricker-Pedley, K., & Arslan, B. (in press). *Does the time between scoring session impact scoring accuracy?* (Research Report). Princeton, NJ: Educational Testing Service.

Jastrzembski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference* (pp. 1498–1508). Arlington, VA: National Training & Simulation Association.

Jastrzembski, T. S., Walsh, M., Krusmark, M., Kardong-Edgren, S., Oermann, M., Dufour, K., ... & Stefanidis, D. (2017). Personalizing training to acquire and sustain competence through use of a cognitive model. In D. D. Schmorrow & C. M. Fidoplastis (Eds.), *Lecture notes in artificial intelligence: Vol. 10285: Augmented cognition. Enhancing cognition and behavior in complex human environments* (pp. 148–161). Cham, Switzerland: Springer.

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 11–24). New York, NY: Routledge.

LaMar, M. M. (2014). *Models for understanding student thinking using data from complex computerized science tasks* (Unpublished doctoral dissertation). University of California, Berkeley.

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika*, 83, 67–88. https://doi.org/10.1007/s11336-017-9570-0

LaMar, M. M., Baker, R. S., & Greiff, S. (2017). Methods for assessing inquiry: Machine-learned and theoretical. In R. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring systems: Volume 5 – Assessment methods* (pp. 137–153). Orlando, FL: U.S. Army Research Laboratory.

Moon, J., Katz, I. R., & Keehner, M. (2017, July). *Effects of question format on test-taker cognition*. Poster presented at the Thirty-Ninth Annual Conference of the Cognitive Science Society, London, UK.

Moon, J., Keehner, M., & Katz, I. R. (2018). *Affordances of item formats and their effects on test-taker cognition under uncertainty*. Manuscript in preparation.

Nhouyvanisvong, A., & Katz, I. R. (1998). The structure of generate-and-test in algebra problem solving. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society* (pp. 758–763). Hillsdale, NJ: Erlbaum.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001
email: **RDWeb@ets.org**

Editor: Hans Sandberg
Copy Editor: Eileen Kerrigan
Layout Design: Sally Acquaviva

Visit ETS Research & Development on the web at **www.ets.org/research**

Follow ETS Research on Twitter®
(@ETSresearch)