Research Statement

Noboru Matsuda
July 27, 2021

I am primarily interested in advancing the computational theory for scalable learning technologies and their practical applications for learning and teaching to advance the theory of the sciences of learning. I am particularly interested in artificial intelligence (AI) technologies that assist students to learn, teachers to teach, and researchers to understand how people learn. My scholarly contribution thus spans computer science, cognitive science, and learning science with a primary focus on STEM education.

To expand the theory of computing for scalable learning technologies, I apply the data-driven, iterative design-engineering methodology. To understand students' and teachers' needs and challenges, I utilize user-centered methods commonly used in the field of human-computer interaction and develop technological solutions, which typically leverage AI. To validate the proposed solution, I conduct field evaluation studies with actual students and teachers in authentic classroom settings, which results in collecting a large volume of data. I endeavor to analyze those data using both (traditional) statistical and advanced data mining methods to investigate broad research questions regarding how to advance the theory of computing for education.

To advance the theory of learning and teaching, I test specific hypotheses on how people learn and what makes teaching more effective using the techniques noted above. Consequently, the empirical investigations that I conduct are often randomized controlled trials. The fact that I am a two-time awardee of the Institute of Education Science (IES) of the US Department of Education, an extremely competitive funding program, evidences my intellectual strengths in this line of research.

1. Research Accomplishments

While the technologies that I use for my research are generally domain independent, my primary research focus is on mathematics education, which was my major in my bachelor's and master's degrees. This section describes my research accomplishments over the last 15 years.

1.1 Studying a Computational Theory of Learning

I am probably most well-known for my work on SimStudent (www.SimStudent.org) where we study a computational model of learning. At the behavioral level, SimStudent is s *synthetic tutee* in the form of an interactive machine learning agent that learns cognitive skills to solve complex problems through worked examples and guided problem solving. Technologically, it is a realization of programming by demonstration with a novel combination of multiple machine-learning and AI techniques.

The SimStudent technology has been applied in three major research projects (as detailed below): (1) *Intelligent authoring aid* where I study how to build a computer tutor by tutoring the synthetic tutee; (2) *Student modeling* where I study how students learn by simulating their learning processes using the synthetic tutee; (3) *Learning by teaching* where I study how and when students learn complex problem-solving skills by teaching the synthetic tutee.

I launched the SimStudent project when I joined Carnegie Mellon University as a postdoctoral research fellow in 2005. Since then, I have led and cultivated the project into a multimillion-dollar research initiative. I have been awarded five major federal grants in this line of research totaling over $5M for which I have served as Principal Investigator on all but one projects (I was a postdoc on the first grant). To date, the SimStudent project has yielded eight journal papers (including highly impactful journals such as *Artificial Intelligence* and the *Journal of Educational Psychology*) and 36 peer-reviewed conference papers.

Intelligent Authoring Aid with a Synthetic Peer

The goal of this project is to develop an innovative technology for authoring an intelligent tutoring system (ITS) by interactively *tutoring* SimStudent, the synthetic peer, how to solve target problems that the ITS teaches to students (Matsuda, Cohen, & Koedinger, 2005; Matsuda, Cohen, Sewall, Lacerda, & Koedinger, 2008). In this context, SimStudent functions as a building block of an existing suite of

authoring tools called Cognitive Tutor Authoring Tools (CTAT). Cognitive tutor is a particular type of ITS that facilitates mastery learning for problem-solving skills. Using CTAT, an author first creates a tutoring interface for a target cognitive tutor. Then, the author tutors SimStudent using the tutoring interface just created. SimStudent learns a set of skills to perform the target task represented as production rules that become an expert model of the cognitive tutor.

Empirical studies where the effectiveness of authoring cognitive tutors using SimStudent demonstrated that *the pedagogical agent technology has the potential for a substantial impact on facilitating the authoring of cognitive tutors, which in turn has a significant influence on the rapid and broad dissemination of cognitive tutors that provide adaptive tutoring for students* (Matsuda, Cohen, & Koedinger, 2015). Another empirical study demonstrated that *the synthetic peer can facilitate the cognitive task analysis, which is an essential step which building cognitive tutors, where a failure of the synthetic peer performing the target task hints a flaw in the cognitive factors provided to the synthetic peer, e.g., background knowledge and the configuration of the tutoring interface* (Matsuda, in press).

I was a co-PI on this project that was supported by NSF, Advanced Learning Technologies (ALT): "Building Cognitive Tutors with Programming by Demonstration: When Simulated Students help Cognitive Modeling and Educational Studies". September 15, 2005 to August 31, 2009. I am a PI on the more recent grant project funded by the Institute of Education Studies at the US Department of Education: Developing an online learning environment for learning algebra by teaching a synthetic peer. September 1, 2018 to August 31, 2023.

Student Modeling by Simulating Student Learning with a Pedagogical Agent

Using the SimStudent technology, researchers can conduct tightly controlled simulation studies to explore both domain-specific and domain-general theories of learning. As an example of advancing a domain-specific theory of learning, I studied how students make induction errors in Algebra that lead to learning incorrect skills (Matsuda, Epstein, Cohen, & Koedinger, 2008; Matsuda, Lee, Cohen, & Koedinger, 2009). I hypothesized that induction errors occur due to shallow, perceptually-grounded prerequisite knowledge that leads students to applying algebraic operations to algebraic symbols without taking their meanings into account; e.g., reading the '3' in '3*x*' as a number before a letter instead of a coefficient of a variable term. An empirical study showed that when the shallow prerequisite knowledge was given, SimStudent yielded a more accurate model of induction errors that showed a better fit to actual students' learning data—i.e., making the same errors that students commonly make. *The results from the empirical study provide a theoretical account for the relationship between conceptual prerequisites and skill acquisition when learning Algebra*. *The SimStudent technology therefore has a non-trivial potential for contributing to theory development on human learning through simulation studies*.

This project was funded by the Pittsburgh Science of Learning Center: "Towards a Theory of Learning Errors: Application of a Synthetic Student to Model How Students Learn Errors," for which I was a Principal Investigator from September 1, 2008 to August 31, 2009.

Studying the Theory of Learning by Teaching with the Synthetic Peer

Despite ample evidence showing that students learn when they teach (Roscoe & Chi, 2008), little is known about how and why students learn by teaching others. Therefore, I launched a project to investigate a cognitive theory of learning by teaching using SimStudent as a synthetic peer (aka, a teachable agent). I developed an online game-like learning environment in which students learn algebraic equations by interactively teaching SimStudent.

So far, I have conducted 8 evaluation studies with more than 2000 middle school students who participated in their actual algebra classrooms. The most important findings include the following: (a) By teaching the synthetic peer (i.e., SimStudent), students improve their skills of solving equations (Matsuda, Yarzebinski, Keiser, Raizada, Stylianides, et al., 2012). (b) Answering the tutee's questions facilitates students' learning when they enter elaborated answers (Matsuda, Cohen, et al., 2012). (c) The maturity of the student's prior knowledge has a significant influence on the effect of learning by teaching (Matsuda et al., 2013). (d) The students' and the synthetic tutee's learning are highly correlated. When SimStudent commits to shallow learning, the students' learning is also subpar (Matsuda, Yarzebinski, Keiser,

Raizada, Cohen, et al., 2012). (e) Learning by teaching requires adaptive scaffolding, and it is scaffolding on how to teach (as opposed to how to solve) the target task that facilitate tutor learning (Matsuda, Weng, & Wall, 2020). *A methodological strength of this line of research is the capability of the SimStudent technology that allows us to conduct tightly controlled studies in authentic classroom settings and collect detailed learning process data in combination with learning outcome data*.

I have been the Principal Investigator on four grants on this line of research: (1) NSF, Research on Education and Learning (REAL). "Learning by Teaching a Synthetic Peer: Investigating the effect of tutor scaffolding for tutor learning". 10/1/2013 to 9/30/2016. (2) NSF, Research and Evaluation on Education in Science and Engineering (REESE). "Learning by Teaching a Synthetic Student: Using SimStudent to Study the Effect of Tutor Learning". 8/1/ 2009 to 7/31/2013. (3) US Department of Education, IES. "Learning by Teaching a Synthetic Student: Using SimStudent to Study the Effect of Tutor Learning". 6/1/2009 to 5/31/2013. (4) US Department of Education, IES. "Developing an online learning environment for learning algebra by teaching a synthetic peer". 9/1, 2018 to 8/31/2023.

1.2 Studying Efficient Learning-Engineering Methods for Building Online Courseware at Scale

The primary aim of this line of research is to investigate evidence-based, transformative engineering *methods* to efficiently create adaptive online courseware (similar to Coursera, edX, and Moodle, as opposed to a study management system like Blackboard, but with *adaptive pedagogy*). I hypothesize that students' learning log data (e.g., performance and assessment records) and instructional materials data (e.g., written instructions and videos) will provide us with insights into designing effective online courseware.

Data-driven Online Courseware Development Environment

We have developed an integrated development environment (IDE) for evidence-based creation of online courseware, called IDEA (**I**ntegrated **D**evelopment **E**nvironment with learning **A**nalytics). The ideas behind the IDEA project were inspired by the contextual interview with actual online courseware developers that I have conducted with my prior research team at Carnegie Mellon University.

IDEA provides courseware developers with data-driven feedback on the structure of the courseware (e.g., a lack of formative assessment for a particular learning objective) that will inform the potential improvement of course activities. IDEA also provides courseware developers with a visually interactive (aka WYSIWIG) authoring environment. The system provides data-driven feedback while authors are editing the courseware content. We have applied IDEA to an existing online courseware platform, Open Learning Initiative (OLI), and iteratively improved a few online courses. The results from an evaluation study showed that the improved OLI course on Discrete Math indeed yielded shorter learning time and better learning outcome when used as a gate-way course for undergraduate freshman at the School of Computer Science at CMU. Through the IDEA project, *we have demonstrated the fidelity of implementation of the data-driven, iterative online courseware engineering methods, which we argue, is a critical component of successful dissemination for effective online education*.

The IDEA project, "Data-Driven Methods to Improve Student Learning from Online Courses," was funded by NSF, Research on Education and Learning program (DIR), and I was the Principal Investigator from 8/1/2014 to 7/31/2017.

Scalable Learning Engineering Methods

More recently, I launched another project to study learning-engineering methods to build adaptive online courseware, called PASTEL (**P**ragmatic methods to develop **A**daptive and **S**calable **T**echnologies for next generation **E**-**L**earning).

The primary aim of the PASTEL project is to investigate evidence-based, transformative engineering methods to allow instructional engineers to efficiently create effective online courseware, called CyberBook. CyberBook is an integration of cognitive tutors into online courseware. I hypothesize that by integrating cognitive tutors (that provide adaptive instruction on mastery learning) and traditional online courseware (that provide media-rich, multi-modal instruction), we can create adaptive online courseware

that fosters students' *synergetic competency* that, by definition, is robust learning with a tight connection between conceptual and procedural understanding.

Currently, the PASTEL project focuses on six technologies: (1) a text-mining method to discover latent concepts and skills in the didactic instructional text used in the courseware, (2) a web-based authoring tool to create intelligent tutoring systems by demonstrating solutions and seamlessly integrating them to the courseware, (3) an application of deep learning for question generation from didactic instructional text, (4) an application of reinforcement learning for optimally sequencing the courseware contents based on individual students' competency, (5) a stochastic prediction method to detect students at risk, and (6) a quality-assurance technique for courseware contents based on students' performance on the courseware. Practically, these technologies will allow courseware developers to efficiently create adaptive online courseware, which in turn will globally impact hundreds and thousands of diverse students' learning.

So far, we have prototyped each of the six PASTEL technologies and integrated them into the Open edX platform as a basis of CyberBook. Preliminary results from empirical evaluations include the followings (but not limited to): (a) A neural-network based detector has high recall of 0.79—i.e., 79% of actually at-risk students are correctly detected, but relatively low precision, 0.25—i.e., only 25% of at-risk decisions are correct, which means that the current detector is too pessimistic to categorize non-at-risk students as at risk (Matsuda, Chandrasekaran, & Stamper, 2016). (b) The machine discovered skill model yielded a better prediction model than the human-generated skill model on an existing Biology online courseware on Open Learning Initiative (Matsuda, Furukawa, Bier, & Faloutsos, 2015). (c) A combination of a bi-directional transformer (BERT) and an existing question conversion technology can generate questions from given instructional text that are aligned with specific learning objectives (Shimmei & Matsuda, 2021). *The broader impact of this project includes a global dissemination of the next generation cyberlearning infrastructure on which researchers and educators can conduct a variety of research and practice toward a pragmatic application of adaptive online courseware and evidence-based learning engineering*.

The PASTEL project has been awarded two NSF grants for which I served as the PI: (1) Cyberlearning and Future Learning Technologies. Exploratory study on the Adaptive Online Course and its implication on synergetic competency. 8/1/2016 to 7/31/2018. (2) Cyberlearning and Future Learning Technologies. Collaborative Research: Cyberinfrastructure for Robust Learning of Interconnected Knowledge. Principal Investigator. 7/1/2020 to 6/30/2023.

2. Future Directions

For the next five to ten years, I will further extend my work to advance the theory of learning and teaching by utilizing advanced technology innovation as my secret weapon. The modern education system has many complicated problems that require deep understanding of the students', teachers', and practitioners' needs to meet a vigorous solution. Throughout my previous projects, I have empirically demonstrated my scholarly strength in technology innovation and theory development with the spiral iteration starting from the user-centered investigation. I believe that by continuing these efforts, we can solve many important and urgent problems in modern education.

I would continue studying the theory of the intelligent pedagogical agent and adaptive online courseware engineering for various learning and training applications, including formal/informal K-16 education, life-long learning, and industrial training. With these specific research projects, I will continue studying the artificial intelligence applications for education to achieve broad dissemination of effective life-long education for the diverse student population.

References:

Matsuda, N. (in press). Teachable Agent as an Interactive Tool for Cognitive Task Analysis: A Case Study for Authoring an Expert Model. *International Journal of Artificial Intelligence in Education*, 1-28.

Matsuda, N., Chandrasekaran, S., & Stamper, J. (2016). How quickly can wheel spinning be detected? In T. Barnes, M. Chi & M. Feng (Eds.), *Proceedings of the International Conference on Educational Data Mining* (pp. 607-608).

Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2005). Applying Programming by Demonstration in an Intelligent Authoring Tool for Cognitive Tutors *AAAI Workshop on Human Comprehensible Machine Learning (Technical Report WS-05-04)* (pp. 1-8). Menlo Park, CA: AAAI association.

Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2015). Teaching the Teacher: Tutoring SimStudent leads to more Effective Cognitive Tutor Authoring. *International Journal of Artificial Intelligence in Education, 25*, 1-34. doi: 10.1007/s40593-014-0020-1

Matsuda, N., Cohen, W. W., Koedinger, K. R., Keiser, V., Raizada, R., Yarzebinski, E., . . . Stylianides, G. J. (2012). Studying the Effect of Tutor Learning using a Teachable Agent that asks the Student Tutor for Explanations. In M. Sugimoto, V. Aleven, Y. S. Chee & B. F. Manjon (Eds.), *Proceedings of the International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL 2012)* (pp. 25-32). Los Alamitos, CA: IEEE Computer Society.

Matsuda, N., Cohen, W. W., Sewall, J., Lacerda, G., & Koedinger, K. R. (2008). Why Tutored Problem Solving may be better than Example Study: Theoretical Implications from a Simulated-Student Study. In B. P. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 111-121). Heidelberg, Berlin: Springer.

Matsuda, N., Epstein, S., Cohen, W. W., & Koedinger, K. R. (2008). Towards a theory of learning errors: application of a synthetic student to analyze students errors *Proceedings of Japan National Conference on Information and Systems in Education* (pp. 362-363). Kumamoto, Japan.

Matsuda, N., Furukawa, T., Bier, N., & Faloutsos, C. (2015). Machine beats experts: Automatic discovery of skill models for data-driven online course refinement. In J. G. Boticario, O. C. Santos, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Michaescu, P. Moreno, A. Hershkovitz, S. Ventura & M. C. Desmarais (Eds.), *Proceedings of the International Conference on Educational Data Mining* (pp. 101-108). Madrid, Spain.

Matsuda, N., Lee, A., Cohen, W. W., & Koedinger, K. R. (2009). A Computational Model of How Learner Errors Arise from Weak Prior Knowledge. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 1288-1293). Austin, TX: Cognitive Science Society.

Matsuda, N., Weng, W., & Wall, N. (2020). The effect of metacognitive scaffolding for learning by teaching a teachable agent. *International Journal of Artificial Intelligence in Education, 30*(1), 1-37.

Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Cohen, W. W., Stylianides, G. J., & Koedinge, K. R. (2012). Shallow learning as a pathway for successful learning both for tutors and tutees. In N. Miyake, D. Peebles & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 731-736). Austin, TX: Cognitive Science Society.

Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Stylianides, G. J., & Koedinger, K. R. (2012). Motivational factors for learning by teaching: The effect of a competitive game show in a virtual peer-learning environment. In S. Cerri & W. Clancey (Eds.), *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 101-111). Heidelberg, Berlin: Springer-Verlag.

Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., William, W. C., Stylianides, G. J., & Koedinge, K. R. (2013). Cognitive anatomy of tutor learning: Lessons learned with SimStudent. *Journal of Educational Psychology, 105*(4), 1152-1163. doi: 10.1037/a0031955

Shimmei, M., & Matsuda, N. (2021). Learning Association between Learning Objectives and Key Concepts to Generate Pedagogically Valuable Questions. In I. Roll & D. McNamara (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 320-324, short paper).