The Learner Data Institute: Emerging Science Convergence and Research Opportunities

Vasile Rus[1], Stephen E. Fancsali[2], Philip Pavlik, Jr.[1], Deepak Venugopal[1], Steve Ritter[2], Dale Bowman[1], and The LDI Team
[1]The University of Memphis
[2]Carnegie Learning, Inc.
vrus@memphis.edu, sfancsali@caregielearning.com

ABSTRACT
This paper provides an update of the Learner Data Institute (LDI; www.learnerdatainstitute.org) which is now in its third year since *conceptualization*. Funded as a *conceptualization* project, the LDI's first two years had two major goals: (1) develop, implement, evaluate, and refine a framework for data-intensive science and engineering and (2) use the framework to start developing prototype solutions, based on data, data science, and science convergence, to a number of core challenges in learning science and engineering. One major focus in the third, current year is synthesizing efforts from the first two years to identify new opportunities for future research by various mutual interest groups within LDI, which have focused on developing a particular prototype solution to one or more related core challenges in learning science and engineering. In addition to highlighting emerging data-intensive solutions and innovations from the LDI's first two years, including places where LDI researchers have received additional funding for future research, we highlight here various core challenges our team has identified as being at a "tipping point." Tipping point challenges are those for which timely investment in data-intensive approaches has the maximum potential for a transformative effect.

Keywords
big data in education, data science, science convergence, learning engineering, adaptive instructional systems, intelligent tutoring systems.

1. INTRODUCTION
This paper provides an update of the whereabouts of the Learner Data Institute (LDI; www.learnerdatainstitute.org). The present paper is part of a series published each year since the conception of LDI (Rus et al., 2020; Rus et al., 2021), which provided an introduction to LDI and early activities and outcomes (Rus et al., 2020) and the resulting LDI convergence framework (Rus et al., 2021). We emphasize here the developments of the past 12 months (since the 2021 paper), focusing on the key elements of the science convergence framework, its development, implementation, evaluation, and refinement, and key outcomes such as solutions to key challenges in the learning ecosystem and additional funding obtained by LDI members to continue their efforts started as part of the LDI. A more comprehensive lists of our outcomes including additional funding obtained by our team can be found on our website: LDI; www.learnerdatainstitute.org.

The LDI is a "frameworks" project funded by the United States' National Science Foundation (NSF) under the Data-intensive Research in Science and Engineering (DIRSE) program to make the learning ecosystem more effective, efficient, engaging, equitable, relevant, and affordable. It is part of the NSF's Harnessing the Data Revolution[1] (HDR) Institutes effort. "HDR Institutes… enable breakthroughs in science and engineering through collaborative, co-designed programs to formulate innovative data-intensive approaches to address critical national challenges" (NSF-HDR, 2021). LDI focuses on data-intensive approaches to developing and improving learning environments that include adaptive instructional systems as a means to address the challenge of offering access to high-quality education to everyone. Our vision is for LDI to: (i) serve as a hub to identify investment opportunities for data-intensive approaches to core learning science and engineering challenges to accelerate progress toward equitable learning and achievement in education; (ii) foster, support, and build a portfolio of inter-related, inter-disciplinary "Scale-up Projects" to research, develop, and disseminate data-intensive solutions across multiple academic and non-academic communities that currently cannot easily communicate with each other, embodying a process of science convergence; (iii) bridge the HDR ecosystem with the educational data science and learning engineering community and the broader education world, and, in particular, serve as the education & training hub for the HDR ecosystem, assisting other teams with developing data science training platforms for their communities.

Currently, the LDI team consists of 60+ researchers, developers, and practitioners from three continents spanning many disciplines and backgrounds. LDI is led by the Institute of Intelligent Systems

[1] https://www.nsf.gov/cise/harnessingdata/

at The University of Memphis and main corporate partner Carnegie Learning (co-lead and developer of commercial-grade adaptive systems serving over 400,000 students in 2,000+ school districts); the assembled team now spans 14 main organizations on 3 continents including NSF-funded partners such as the Institute for Data, Econometrics, Algorithms, and Learning (IDEAL; HDR TRIPODS project at Northwestern University lead) and the LearnSphere project: Building a Scalable Infrastructure for Data-Driven Discovery and Innovation in Education (NSF DIBBs project; Carnegie Mellon University lead) as well as US Army's Generalized Intelligent Framework for Tutoring project (Sottilare et al, 2012). LDI includes 6 additional corporate partners, 3 laboratory schools (The Early Learning & Research Center, Campus Elementary School, and University Middle School in Memphis, TN), 3 K-12 school districts - Shelby County Schools (Memphis, TN area; 200 schools, 100,000 students), Brockton Public Schools (Boston, MA area; 24 schools, 15,000 students), Val Verde Unified School District (Los Angeles, California area; 21 schools, 20,000 students), and one teacher training program at Christian Brothers University.

Together, we intend a rigorous test of the hypothesis that *emerging learning ecologies that incorporate adaptive instructional systems (AISs) are capable of providing affordable, effective, efficient, equitable, and engaging individualized assistance for both learners and instructors, and that the characteristics, parameters, and impacts of these systems, for example, effectiveness (in terms of learning gains), can be improved over time given sufficient attention to evidence, captured as data, and expertise, provided by teams of interdisciplinary researchers like ours*.

2. DATA SCIENCE AND AISs — A TRANSFORMATIVE MIX FOR THE LEARNING ECOSYSTEM

The LDI is founded on the key observation that data science and AISs are a powerful mix with potentially transformative impact on the learning ecosystem.

Big educational data (edu-data) and recent advances in data science create tremendous opportunities to reveal facets along which learner experiences can be tailored or adapted in ways heretofore impossible. A particular learning environment may result in different learning outcomes for different (groups of) students because of students' idiosyncratic prior knowledge, experience(s), interest(s) and motivation(s). A small minority of students, for example, that approach a problem in a unique way could be overlooked in a small dataset, but larger datasets give us the possibility to detect and account for individual differences in learning. To this end, our mission is to harness the data revolution to further our understanding of how people learn.

AISs can monitor and scaffold learners at a fine level of granularity (e.g., capturing every single step during instructional activities) and with respect to many aspects of learning (e.g., cognitive, behavioral, affective, social, motivational facets of learning) at scale (i.e., for millions of learners and teachers and across many topics and domains) and across time periods (e.g., across grade-levels). Such rich data, when collected, can be characterized as *deep* (many data instances from millions of learners), *wide* (capturing many aspects of the learning process at a fine granularity level), and *long* (longitudinal, i.e., across time and grade levels). Such big edu-data, together with advanced data science methods, are likely to offer insights about learning and instruction and lead to the development of effective and affordable instructional tools that were not possible before. This is promising enough to believe that the learning ecosystem is at a tipping-point to be transformed.

Indeed, LDI is built on the belief that AISs constitute a necessary catalyst to enable the transformation of the learning ecosystem through harnessing the data revolution because, as noted earlier, AISs can monitor and scaffold learners at a very fine granularity level, at scale, and across time. It should be noted that much of education data, (e.g., currently collected by schools), relies on a set of predefined competencies or standards to monitor student progress. Such data only reveal what students know or mastered and what they don't know (didn't master yet), but such data often do not reveal much about the learning and instructional process. That is, much of the school data focus on "where the student is" but not what they do during instructional activities. Fundamentally, teachers and schools in general lack the capacity to monitor and store data about all students at every single step of the learning and instruction process. LDI will thus offer schools a new powerful framework to understand, monitor, and intervene at a fine-grain level with potentially transformative effects on the learning ecosystem.

3. FRAMEWORK FOR SCIENCE CONVERGENCE

A major goal of LDI conceptualization phase has been to develop, implement, test, and refine a *framework for data-intensive research in science and engineering* enabling science convergence, aligning with the Growing Convergence Research (GCR) "big idea" identified by the National Science Foundation.

The LDI's leadership team and participants have designed, prototyped, and tested a process and a corresponding set of tools designed to transform what is currently a loosely coupled group of research centers, AIS commercial providers, and governments research labs engaged in similar but disparate research and development efforts into a set of interacting teams (Berry, 2011; Lilian, 2014), in aggregate constituting a physical and virtual community of practice (Lave & Wenger, 1991). We have not and will not attempt to "tighten" the coupling between participating research centers. As Weik (1991) has argued in respect to educational systems, loosely coupled systems have several advantages over tightly coupled ones—not least flexibility, survivability (with dysfunction in individual nodes tolerable), and increased likelihood of beneficial "mutations." Rather, LDI's leadership has intended to design and test a set of processes and tools that will support the independent work of the participating research centers, facilitate the flow of information and ideas within and across these centers, and help to keep participants focused on common problems without the need for direct intervention (e.g., in the form of a top-down, tightly controlled research agenda). This paper illustrates some of the outcomes of the designed, implemented, and tested convergence framework. More details about the elements of the convergence framework are available in our two previous papers (Rus et al., 2020; Rus et al., 2021).

3.1 Investment Areas and Outcomes

Our strategy to accomplish the LDI mission of transforming the learning ecosystems has been to focus on a number of carefully selected research priorities, targeting key aspects of the learning ecosystem which we believe are at a "tipping point" (i.e., a point at which timely investment in data-intensive approaches focusing on those critical aspects has the maximum potential for a transformative effect).

The identified research priorities were the result of an intense science convergence process involving a number of activities (e.g., brainstorming sessions or "ideas labs" followed by iterative discussions for ranking and selection at "all-hands" virtual meetings, engagement with Expert Panels, etc.). That is, the identified research priorities reflect our collective interdisciplinary wisdom that timely investment in data-intensive approaches will have the maximum potential for a transformative effect.

Identified key investment opportunity areas or thrusts include:

- **Investment Opportunity Area 1**: *Scaling Up Access To Learning Data* – From Impoverished Datasets To Learning Data Convergence To Comprehensive Learner Models
- **Investment Opportunity Area 2**: *Novel, Richer, More Powerful, Scalable, and Accurate Data-intensive Solutions to Core Education Tasks*
- **Investment Opportunity Area 3**: *Human Technology Frontier – Pushing For Wider Adoption and Integration Of AISs*

Investment Opportunity Area 1: *Scaling Up Access To Learning Data.* Access to "big" education data (big edu-data) is vital to enabling progress in educational data science. To this end, a key long term goal of LDI is learning data convergence: collecting and aligning (more) comprehensive data about the same learner(s) across skills, disciplines, and modalities (cognitive, meta-cognitive, emotional, motivational, behavioral, social) and across time (e.g., K-12 grade-levels), as well as data about the learning process and environment.

One area of progress in addressing this "impoverished data" challenge in the education ecosystem (Rus et al., 2020, 2021) and moving towards learning data convergence was the 2021 release of a new collection of MATHia (formerly Cognitive Tutor (Ritter et al. 2007) datasets by Carnegie Learning via LearnLab DataShop (Koedinger et al., 2010). This collection of datasets comprises the "Carnegie Learning MATHia 2019-2020" DataShop project[2] and includes anonymized learner data from 5,000+ learners, working in math content primarily targeted to Grade 7 middle school learners, across an entire academic year. Datasets within the project also provide data for tens of thousands of learners working on particular pieces of MATHia content or "workspaces," enabling approaches that Fancsali et al. have called "non-monolithic" analytical strategies that consider the characteristics of particular pieces of instructional content (Fancsali et al., 2021). Altogether, the collection of datasets represent tens of millions of student actions.

LDI researchers at Carnegie Learning are also involved in an effort to take a more comprehensive view of math learners, by considering reading skills in learning with AISs like MATHia, in what is called the *Math and Reading Acquisition Co-Adaptive System (LDI PIs: Steve Ritter and Stephen Fancsali; Source: Department of Education Institute for Education Sciences; Amount: $1,999,985; Period: 2021-2024).* This work considers new and improved supports for reading within MATHia and mixes theory- and data-driven approaches to identifying granular knowledge components and instructional design features that may be (more) related to reading as opposed to math. Crucial to these efforts are datasets that provide external (to an AIS like MATHia) measures of reading ability (e.g., end-of-year scores on reading or English Language Arts tests or indicators that a student is an emerging English learner) to better understand whether targeted supports are likely to have an impact to improve the learning experience of students with reading difficulties, emerging readers, and/or English learners.

BE-PAL: A Behavior Personal Assistant to Transform Classroom Learning and Teaching and Improve Academic Achievement for All Learners. This project is about developing an effective, efficient, and scalable process and tools to assist schools meet the high behavior management standards set by the federal mandate (IDEA, 1997) while mitigating the costs associated with addressing system-wide needs for behavior assessment and intervention. It will transform the current practice of manual behavior data collection, analysis, and intervention utilized within a positive behavior multi-tiered system of support (MTSS) resulting in improved classroom behavior, more instruction time, and improved academic achievement for all learners. Specifically, the goal is to develop and validate a novel monitoring and intervention system in the form of an educational technology called BE-PAL (BEhavior Personal Assistant for Classroom Learning) whose goal is to enable continuous, accurate, and scalable monitoring, detection, and support, through individualized intervention, of behavioral excesses in classrooms. It should be noted that the BE-PAL system can also detect and reinforce desired behavior in classroom environments, functioning as a feedback tool for all students, including those exhibiting good behavior, thus having a positive impact on all learners.

The BE-PAL project will have a transformative impact on classroom learning and teaching. Indeed, BE-PAL will lead to improved classroom behavior, e.g., more on-task behavior, and free teachers from many behavior management duties, enabling them to spend more time on actual instruction rather than behavior management tasks. More instruction time, which will benefit all learners, combined with more on-task behavior, i.e., for students who benefit from the MTSS and who otherwise would engage in off-task behavior, will together have an amplified positive impact on all students' academic achievement. The proposed BE-PAL system will provide a foundation for extending existing personalized learning technologies to address behavior, leading to more comprehensive personalized educational systems that address behavior, cognitive, social, emotional, and motivational aspects of learning. It is important to note that existing personalized education technologies, such as intelligent tutoring systems, often ignore behavioral aspects. Funding for this project is currently sought by LDI members Rus, Elswick, and Casey.

Investment Opportunity Area 2: *Novel, Richer, More Powerful, Scalable, and Accurate Data-intensive Solutions to Core Education Tasks*.

This investment opportunity area focuses on improving existing methods and models with respect to their scaling and extension using big edu-data and developing novel, richer, more powerful, scalable, and accurate computational models for a number of core educational tasks. The goal is to improve our understanding of how learners learn, improve the effectiveness and efficiency of AISs, make AISs more affordable and scalable horizontally (across topics and domains), and scale AISs vertically (offering training on higher-level skills such as deep conceptual understanding and collaborative problem solving). Examples of such data-intensive solutions and prototypes are illustrated next.

Neuro-symbolic approaches For Student Learning Strategy Prediction: Solution and Software Prototype (Shakya, Rus, &

[2] https://pslcdatashop.web.cmu.edu/Project?id=720

Venugopal, 2021; https://github.com/anupshakya07/SSPM). Predicting student problem-solving strategies is a complex problem but one that can significantly impact automated instruction systems since they can adapt or personalize the system to suit the learner. While for small datasets, learning experts may be able to manually analyze data to infer student strategies, for large datasets, this approach is infeasible. We develop a Machine Learning model to predict strategies from student data. While Deep Neural Network (DNN) based methods such as LSTMs can be applied for this task, they often have long convergence times for large datasets and like several other DNN-based methods have the inherent problem of overfitting the data. To address these issues, we develop a Neuro-symbolic approach for strategy prediction, namely a model that combines strengths of symbolic AI (that can encode domain knowledge) with DNNs. Specifically, we encode relationships in the data using Markov Logic and use symmetries among these relationships to train an LSTM more efficiently. In particular, we use an importance sampling approach where we sample the training data such that for clusters/groups of symmetrical instances (instances where the strategies are likely to be symmetric), we only pick representative samples for training the model instead of using the whole group. Further, since some groups may contain more diverse strategies than the others, we adapt the importance weights based on previously observed samples. Through empirical evaluation on the KDD EDM challenge datasets, we show the scalability of our approach. Funding to extend this effort has been secured as illustrated below.

Investigating techniques that couple Markov Logic and Deep Learning with applications to discovering strategies to improve STEM learning (LDI PIs; Deepak Venugopal and Vasile Rus; Source: National Science Foundation; Amount: $413,482.00; Period: 2020-2023). The goal of this project is to develop novel techniques to integrate different but complementary approaches in artificial intelligence (AI). This research combines the strengths of Deep Neural Networks (DNNs) and Markov Logic Networks (MLNs) to address key shortcomings of those techniques when used by themselves. In particular, the proposed work will address the limitation of DNNs with respect to utilizing background knowledge in learning a model. The fact that DNNs typically do not utilize background knowledge explicitly often results in models that over-fit the training data and generalize poorly on new datasets. On the other hand, statistical relational models such as Markov Logic Networks (MLNs) encode complex background knowledge explicitly but lack inference and learning capabilities that are as scalable and accurate as DNN-based methods. The project will develop novel techniques in which MLNs provide the DNN with task-specific background knowledge which helps the DNN to learn more generalizable models. Further, this project will apply these novel techniques to significantly improve personalized learning in adaptive instructional systems (AISs) for STEM topics. The project will yield i) general-purpose open-source software for learning and inference that can be used by a broad range of application domains and ii) specific models for core tasks in AIS-based learning (e.g. inferring student problem-solving strategies) that can significantly improve the adaptive capabilities of AISs which results in better student engagement and learning. The project will impact a number of communities including machine learning, artificial intelligence, artificial intelligence in education, and educational data mining. The outcomes of our work will be widely disseminated through publications in top conferences and journals, presentations, a website, social media, and training materials for researchers and practitioners.

Existing approaches that incorporate background knowledge into DNNs do so using a Bayesian framework where the types of priors are typically simple to ensure tractability of Bayesian inference. The main technical contribution of this project is to address this limitation by developing DNN models that incorporate rich relational knowledge specified in the form of an MLN. To do this, the project will i) develop new representations that encode symmetries (or exchangeability) in the MLN distribution as sub-symbolic embeddings, ii) develop efficient DNN-based learning algorithms for relational data by exploiting exchangeability of variables specified implicitly by the MLN and iii) develop interpretable generative models using Generative Adversarial Networks utilizing symmetries specified by the MLN to traverse across diverse modes in the distribution. The AIS tasks that will be developed as part of this project will use large-scale datasets and thus convincingly demonstrate the scalability of the proposed models in real-world problems. Further, the models developed for the AIS tasks will help us better understand student needs and learning processes which in turn can inform improvements of advanced educational technologies for STEM topics and help validate and refine human learning theories.

Scaling up AISs horizontally. Developing AISs for new domains is expensive due to the authoring bottleneck (Murray, 1999) which is the result of relying heavily on human experts. Automating as much as possible elements, e.g., domain modelling, of the AIS development and refinement process represents a major need, enabling affordable scalability across topics and domains for future learning ecologies that include AISs. This effort will leverage neuro-symbolic approaches to automatically extracting domain models from both structured (e.g., student performance data) and unstructured data (e.g., text in textbooks).

To this end, one LDI effort we would like to highlight is about the auto-authoring of conversational intelligent tutoring systems for code comprehension (Alshaikh, Tamang, & Rus, 2021). Such automated processes will reduce the time and cost to develop such ITSs, e.g., for porting to new programming languages such as from Java to Python. The tool allows non-experts, e.g., instructors, to create fully functional conversational ITSs based on Socratic dialogue for learning programming and develop code comprehension skills. We developed a proof of concept and validated experimentally for the Java programming language. Indeed, we conducted a controlled experiment with 45 students attending intro-to-programming courses in order to assess the effectiveness of auto-authored Socratic dialogues for code comprehension and learning. The results of the experiments showed that the auto-authored dialogues improved students' programming knowledge by 43% in terms of learning gains (pre-post test). Secondly, we conducted a survey of auto-authored tutoring dialogues by introductory to programming course instructors to evaluate the dialogues' quality to help students better comprehend and learn programming using a Likert scale (1 – completely disagree; 5 – completely agree). The results showed that the instructors rated the questions as agree or strongly agree. However, the instructors suggested that more improvement is required to help students develop a robust understanding of programming concepts.

Investment Opportunity Area 3: *Human Technology Frontier – Pushing For Wider Adoption and Integration Of AISs*

This investment opportunity fosters a portfolio of efforts to push for wider adoption and integration of AISs with school-based and teacher-led learning activities at the Human-Technology Frontier, one other of NSF's ten Big Ideas for Future Investment and *detect* *and mitigate issues related to ethics, equity, inclusion, and diversity in education*. As a general principle, all LDI activities are informed and guided by our goal of using data science and AISs to promote ethics and equity in education (Riddle et al., 2015; Corbett-Davies & Goel, 2018; Gardner, Brooks, & Baker, 2019). We highlight next two successful projects which received further funding from the National Science Foundation and Department of Education's Institute for Education Sciences.

Collaborative Research: Exploring Algorithmic Fairness and Potential Bias in K-12 Mathematics Adaptive Learning (LDI PI: Steve Ritter; $987,015.00; Period: 2020-2023). This project engages innovative ways of considering student identity when investigating fairness and potential bias in algorithms for adaptive learning in AISs like Carnegie Learning's MATHia. Specifically, the project considers, in additional to socio-demographic categories, elements of students' self-expressed identity (Belitz et al., 2022) as potential categories along which bias challenges and opportunities for improved fairness may arise. In addition to investigating fairness and bias in algorithms used to estimate student knowledge, the project also considers data-driven "detector" models for inferring behavior like "gaming the system," which are widely studied in the educational data science literature (e.g., Baker et al., 2008). This project, like others, highlights the importance of working with more comprehensive datasets about learners to better ensure that AISs deliver high-quality experiences to all learners.

iCODE: Adaptive Training of Students' Code Comprehension Processes (LDI PI: Dr. Vasile Rus; Amount: $2 million, Source: Department of Education – Institute for Education Sciences, Period: 2022-2025).

The expected products of the project will be a novel and adaptive intervention to monitor, track, and scaffold code comprehension processes while CS majors, non-CS majors, and students from underrepresented groups (females, students of color, first generation status) engage in code comprehension activities. Furthermore, education materials such as examples for code comprehension activities and assessment instruments will be developed or adapted for the project. Descriptions of the overall intervention and its components, the development process, the results of the pilot summative experiment, and the feasibility and costs of implementing the intervention will be disseminated to stakeholder communities including the instructors involved in the project and their staff, STEM for All Video Showcase, the West TN STEM Hub, researchers, and the wider communities of practice. Dissemination efforts will include reports about our goals, findings and recommendations, short videos, presentations and publications, and professional development workshops focusing on the major findings of the project and on inclusive teaching practices. We will continuously share insights and products as they are generated during the project period so that teachers and other stakeholders can take advantage of new insights and products immediately and also to give them the opportunity to provide feedback to our team.

This Development and Innovation project proposes to develop and investigate a novel education technology called iCODE (improve source CODE comprehension) that targets code comprehension, a critical skill for both learners and professionals. Offering support to enhance learners' source code comprehension skills will have lasting positive effects for their academic success and future professional careers. iCODE will integrate reading strategies training, Animated Pedagogical Agents, inclusive and culturally responsive instructional design, and the Open ProSocial Learner Model, to improve code comprehension, learning, engagement, self-efficacy, CS identity, and retention in CS programs. By adapting to individual learner characteristics (prior knowledge, self-efficacy, engagement, socio-cultural factors) and code characteristics (language, cohesion, and readability), iCODE will benefit both CS and Non-CS majors, including underrepresented groups in CS and higher education such as women, students of color, and first generation college students. We will conduct a summative pilot study to evaluate our hypotheses that iCODE will improve code comprehension, learning, engagement, self-efficacy, CS identity, and retention in CS programs, which typically suffer from high attrition rates (30–40%, or even higher in introductory CS courses).

4. CONCLUSIONS

To sum up, our strong team of interdisciplinary experts, developers, and practitioners is working together to move current practices beyond the small-scale studies to bring the learning sciences into the era of big data and interdisciplinary science convergence. The impact of LDI will be felt far and wide, propagating and evolving beyond the lifetime of the award and beyond our own team, acting as an agent of change for how research questions are conceived and addressed through interdisciplinary, collaboration, and co-designed research and development. The proposed processes, methods, and studies pave the way for taking these outcomes to other domains.

ACKNOWLEDGMENTS

The Learner Data Institute is sponsored by the National Science Foundation (NSF; award #1934745). The opinions, findings, and results are solely the authors' and do not reflect those of NSF.

5. REFERENCES

[1] Alshaikh, Z., Tamang, L., & Rus, V. (2021). *Experiments with Auto-generated Socratic Dialogue for Source Code Understanding*. In Proceedings of the 13th International Conference on Computer Supported Education - Volume 2: CSEDU, ISBN 978-989-758-502-9; ISSN 2184-5026, pages 35-44. DOI: 10.5220/0010398100350044

[2] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., and Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, 2008.

[3] Belitz, C., Ocumpaugh, J., Ritter, S., Baker, R. S., Fancsali, S. E., & Bosch, N. Constructing categories: Moving beyond protected classes in algorithmic fairness. *Journal of the Association for Information Science and Technology*, 1-6, 2022. https://doi.org/10.1002/asi.24643

[4] Berry, G. R. (2011). Enhancing effectiveness on virtual teams: Understanding why traditional team skills are insufficient. *The Journal of Business Communication, 48*(2), 186-206.

[5] Corbett-Davies, S. & Goel, S. (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, arXiv:1808.00023, 2018.

[6] Fancsali, S.E., Li, H., Sandbothe, M., Ritter, S. *Targeting design-loop adaptivity*. In Proceedings of the Fourteenth International Conference on Educational Data Mining. 2021.

[7] Gardner, J., Brooks, C., & Baker, R. S. J. d. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis, in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 225–234.

[8] *Growing Convergence Research*. (NSF-GCR, 2020). National Science Foundation's Growing Convergence Program, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505637 (accessed online on June 15, 2020)

[9] *Harnesssing the Data Revolution*. (NSF-HDR, 2021). National Science Foundation's Harnessing the Data Revolution, https://www.nsf.gov/cise/harnessingdata/ (accessed online on June 11, 2021)

[10] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. A data repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press. 2010.

[11] Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

[12] Lilian, S. C. (2014). Virtual teams: Opportunities and challenges for e-leaders. *Procedia-Social and Behavioral Sciences, 110*, 1251-1261.

[13] Riddle, T., Bhagavatula, S., Guo, W., Muresan, S., Cohen, G., Cook, J., and Purdie-Vaughns, V. (2015). Mining a Written Values Affirmation Intervention to Identify the Unique Linguistic Features of Stigmatized Groups. Proceedings of EDM 2015.

[14] Ritter, S., Anderson, J.R., Koedinger, K.R., & Corbett, A. The Cognitive Tutor: Applied research in mathematics education. Psychonomics Bulletin & Review, 14(2), 249-255, 2007.

[15] Rus, V., Fancsali, S.E., Bowman, D., Pavlik Jr., P., Ritter, S., Venugopal, D., Morrison, D., and The LDI Team (2020). The Learner Data Institute: Mission, Framework, & Activities. In V. Rus & S.E. Fancsali (Eds.) *Proceedings of The First Workshop of the Learner Data Institute*, The 13th International Conference on Educational Data Mining (EDM 2020), July 10-13, Ifrane, Morroco (held online).

[16] Rus, V., Fancsali, S.E., Venugopal, D., Pavlik Jr., P., Graesser, A.C., Bowman, D., Ritter, and The LDI Team. (2021). The Learner Data Institute - Conceptualization: A Progress Report, *Proceedings of The Second Workshop of the Learner Data Institute*, The 14th International Conference on Educational Data Mining (EDM 2021), June 29 - July 2 , Paris, France (held online).

[17] Sottilare, R.A., Brawner, K.W., Goldberg, B.S., & Holden, H.K. (2012). *The Generalized Intelligent Framework for Tutoring (GIFT)*. Downloaded from www.gifttutoring.org on November 30, 2012.