I/ITSEC IEEE Learning Standards Special Event

Subject: Recommended Practices for Evaluating Artificial Intelligent Systems (AISs)

Author: Louise Yarnall

Discussion Topic: Evaluation and critique descriptors needed to inform AIS buyers about different AIS capabilities

Background: Many education institutions have adopted artificial intelligent systems (AISs)—also known as adaptive learning products—that use computer algorithms to parse learning analytic data collected as students study within courseware and online. In theory, such products mimic many of the best aspects of a personal tutor: generating personalized feedback, study reminders, and content recommendations, and providing real-time dashboards that report on progress. However, these products have shown inconclusive results in studies of learner outcomes and cost effectiveness (Yarnall, Means, & Wetzel, 2016). To support better outcomes, standards should encourage vendors to share more details and evaluative evidence to help education consumers make decisions about when, where, and how to integrate these products into instruction. The current state of the art remains far from achieving this goal. Learners, particularly those in 4-year universities, have questioned the quality of these AIS learning experiences (Yarnall et al., 2016). Also, vendors still use the detailed data that they collect on learners primarily for internal user studies that serve their own product marketing and development needs; standards should explore ways to encourage vendors to report evaluative findings from such data to inform education customers (Yarnall, Boyce, Wetzel, Snow, & Murphy, 2018).

The increasing use of AIS products in institutions of education raises the stakes around the need to provide more transparency and better evidence to users about these products. This is particularly the case for high-cost institutions of higher education, which face questions from students about whether classroom experiences that increasingly depend on AIS courseware provide learning value commensurate with the costs of tuition. This paper discusses some evaluation standards that the Institute of Electrical and Electronics Engineers (IEEE) might consider to foster more transparency about AIS products.

Based on our past research, educators particularly need, at minimum, two additional types of information about adaptive learning products: The "inside" features associated with these products' learning design, such as how and when AISs track learner progress, and the "outside" conditions associated with effective integration of AIS activities with classroom activities. There are two core reasons in support of such evaluative standards. First, providing greater transparency around the "inside" learning design features in these products can help educators to align these products with their own instructional assumptions and approaches. Second, providing greater transparency around conditions of usage and adoption—based on empirical data that the vendor has collected and analyzed—helps educators more accurately predict how well the product might function in their learning contexts and how long it will take to achieve optimal results. These suggested standards for sharing "inside" and "outside" product features and findings will be discussed in more detail below.

Standards that foster transparency about what is "inside" an adaptive learning product

1. *Transparency around the models used to interpret learner progress*. One possible evaluation standard might require vendors to label their products to describe the assumptions of an adaptive learning product's models that guide learners. Broadly speaking, the adaptive learning literature defines two distinct types of models: Macroadaptive and microadaptive (Desmarais & Baker, 2012; Galyardt, 2015; Koedinger, Brunskill, Baker, Mclaughlin, & Stamper, 2013). A standard

might require vendors to describe how much of their product is macroadaptive and how much is microadaptive because such information can help instructors know when and how to assign students to engage with the product and what learning results to expect. For example, if a product is primarily macroadaptive, the product tracks progress on units, chapters, quizzes, and tests through an entire course. An instructor might use a predominantly macroadaptive product as a tool to track student engagement and performance around assigned readings and units. If a product is primarily microadaptive, the product estimates the accuracy of each step that a learner takes on multi-step tasks and procedures, providing real-time feedback, encouragement, and hints. An instructor might use a macroadaptive product to engage learners in extended practice and remediation on complex procedures.

2. *Transparency around the learning principle(s) informing personalized guidance to learners.* Another possible standard might focus on encouraging vendors to share more specific details about the kinds of automated supports that their products provide. Such information can help instructors better interpret dashboard data to understand the strengths and pitfalls that different students may experience with a product. For example, learners who like to learn by observation might show up in a courseware product's dashboard as high users of a class of automated supports known as "worked examples," but these same learners might also lag in their progression to independent problem solving tasks. As another example, learners who struggle with focus and persistence might benefit from products that support repetition or reinforcement, but they also might require monitoring to ensure they don't overuse another class of automated supports that allow them to skip lessons and opt for shorter content (Crooks, Klein, Jones, & Dwyer, 1996; Durlach & Ray, 2011; Metzler-Baddeley & Baddeley, 2009; Scheiter & Gerjets, 2007; Sung & Mayer, 2013; Sweller, 2003; Xiong & Beck, 2011). Also, product developers should be transparent when using more experimental types of automated supports that have less evidence of efficacy. Examples of experimental supports include those that infer learners' moods and use such data to trigger the appearance of supportive avatars or pop-up reminders to maintain learner engagement (Azevedo & Hadwin, 2005; Koedinger & Aleven, 2007; Mendicino, Razzaq, & Heffernan, 2009; Narciss, 2013; Roll, Baker, Aleven, & Koedinger, 2014; Shute, 2008; Sitzman & Ely, 2011).

Standards that foster transparency about what needs to be "outside" of an adaptive learning product

1. *Transparency around vendors' empirical evaluation findings about recommended usage and contextual conditions.* Vendors sometimes conduct evaluations of their products, but mostly for their internal purposes. A standard should encourage vendors to share more information from such evaluations to help instructors and students. Of particular promise are vendor evaluations that examine how different product usage rates and different local implementation conditions relate to academic achievement with students with various characteristics (such as, prior achievement, gender, race/ethnicity, institution type, subject domain, and socioeconomic status). The standard might also require vendors to be clear about the study designs that informed those findings. To be most useful, any evaluative claims about impacts on student achievement should control for all learners' baseline scores of subject matter knowledge and compare outcomes with a control group not using the product. If such evaluation standards have not been observed, then vendors accordingly should be required to disclose the limitations of any evaluative claims.
2. *Transparency around vendors' empirical findings about the costs associated with instructor professional development and technological integration.* Adaptive learning products vary widely in how much training instructors need to achieve optimal use. Therefore, another evaluation

standard might focus on transparently sharing vendors' findings about the amount of time it takes to familiarize instructors with all the critical features of their products, and their findings about which instructor product usage decisions are associated with optimal learning results. For example, some of our past research suggests that products that give instructors a high degree of control over which product features to activate or how to design adaptive learning experiences require more professional development to achieve optimal use.

References

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition - Implications for the design of computerbased scaffolds. *Instructional Science, 33,* 367–379. doi:10.1007/s11251-005-1272-9

Crooks, S. M., Klein, J. D., Jones, E. E., & Dwyer, H. (1996). Effects of cooperative learning and learner-control modes in computer-based instruction. *Journal of research on computing in education*, *29*(2), 109-123.

Desmarais, M. C., & Baker, R. S. J. D. (2012b). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modelling and User-Adapted Interaction, 22(*1), 9–38. http://doi.org/10.1007/s11257-011-9106-8

Durlach, P. J., & Ray, J. M. (2011). *Designing adaptive instructional environments: Insights from empirical evidence (Technical Report 1297).* Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. Retrieved from http://www.adlnet.gov/wp-content/uploads/2011/11/TR-1297.pdf

Galyardt, A. (2015). Move your lamp post: Recent data reflects learner knowledge better than older data. *Journal of Educational Data Mining,7*(2), 85–111. Retrieved from http://educationaldatamining.org/JEDM13/index.php/JEDM/article/view/JEDM100

Koedinger, K. R., Brunskill, E., Baker, R. S. J. d., Mclaughlin, E. A., & Stamper, J. (2013). New potentials for data-driven inteligent tutoring system development and optimization. *AI Magazine, 34*(3), 27–41. doi:http://dx.doi.org/10.1609/aimag.v34i3.2484

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology* Review,19(3), 239–264. http://doi.org/10.1007/s10648-007-9049-0

Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). A comparison of traditional homework to computer-supported homework. Journal of Research on Technology in Education, 41(3), 331-359.

Metzler-Baddeley, C., & Baddeley, R. J. (2009). Does adaptive training work? Applied Cognitive Psychology, 23(2), 254–266. doi:10.1002/acp.1454

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. Digital Education Review, 23(1), 7–26. Retrieved from http://revistes.ub.edu/index.php/der/article/view/11284

Roll, I., Baker, R. S. J. d., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. Journal of the Learning Sciences, 23(4), 537–560. doi:10.1080/10508406.2014.883977

Scheiter, K., & Gerjets, P. (2007). Learner control in hypermedia environments. Educational Psychology Review, 19(3), 285–307. http://doi.org/10.1007/s10648-007-9046-3

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. http://doi.org/10.3102/0034654307313795

Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. Psychological Bulletin, 137(3), 421–42. doi:10.1037/a0022777

Sung, E., & Mayer, R. E. (2013). Online multimedia learning with mobile devices and desktop computers: An experimental test of Clark's methods-not-media hypothesis. Computers in Human Behavior, 29(3),639–647. doi:10.1016/j.chb.2012.10.022

Sweller, J. (2003). Evolution of human cognitive architecture. In B. H. Ross (Ed.) The psychology of learning and motivation: Advances in research and theory, Vol. 43 (pp. 216-261). New York, NY: Academic Press.

Xiong, X., & Beck, J. E. (2014). A study of exploring different schedules of spacing and retrieval interval on mathematics skills in ITS environment. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI (pp. 504–509). Springer International. doi:10.1007/978-3-319-07221-0

Yarnall, L., Boyce, J., Wetzel, T., Snow, E., & Murphy, R. (2018). *An analysis of the relation between student usage and course outcomes for MyLab Math and MyLab Foundational Skills*. New York, NY: Pearson Education & SRI Education. Retrieved from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/reports/Pearson-MyLab-Math-ALMAP-Technical-Report.pdf

Yarnall, L., Means, B., & Wetzel, T. (2016). *Lessons learned from early implementations of adaptive courseware*. Menlo Park, CA: SRI Education. Retrieved from https://www.sri.com/sites/default/files/brochures/almap_final_report.pdf

- Discussion:
  - Talking Point #1: There are many features inside of adaptive learning products, but little transparency around developers' learning models and assumptions. Standards should drive more transparency around these features so that educators can better align these products with their own instructional approaches.
  - Talking Point #2: Vendors of adaptive learning products know that educators use their products differently in different contexts and they have access to empirical data about different usage patterns and contexts, but they fail to share such data with users to help them

understand how to achieve optimal results. Standards should drive transparency around what vendors have discovered about what works in various settings with their products.

- Talking Point #3: Vendors make many marketing claims about how their adaptive courseware will achieve learning results more effectively, but these claims often are not backed by data collected according to the standards of scientific evaluation. Standards should foster best practices of evaluation among vendors.
- Talking Point #4: Vendors of adaptive learning products know which features of their products require a steep learning curve for educators and substantial institutional investment in technology upgrades, but such requirements are not frequently shared in an open manner that permits product comparison. Standards should drive transparency around the costs involved to build readiness for educators and institutions to achieve best results from adaptive learning products.

- Recommendations:
  - Recommendation #1: Vendors should describe the core features and assumptions of their learning designs in a way that clarifies how learners' progress is tracked and how the software guides their learning.
  - Recommendation #2: Vendors should share more empirical data with customers so they understand what ways to adjust the learning context and instructional activities to help learners achieve optimal results.
  - Recommendation #3: Vendors should adhere to high standards of evaluation of their products and base marketing claims on those standards.
  - Recommendation #4: Vendors should work with customers to gather and report data about what ramp-up costs will be required to ensure teachers are adequately trained to use these products as intended.