Int J Artif Intell Educ (2016) 26:387–392
DOI 10.1007/s40593-015-0043-2

ARTICLE

Evaluation Methods for Intelligent Tutoring Systems Revisited

Jim Greer[1] · Mary Mark[2]

Published online: 23 April 2015

Abstract The 1993 paper in IJAIED on evaluation methods for Intelligent Tutoring Systems (ITS) still holds up well today. Basic evaluation techniques described in that paper remain in use. Approaches such as kappa scores, simulated learners and learning curves are refinements on past evaluation techniques. New approaches have also arisen, in part because of increases in the speed, storage capacity and connectivity of computers over the past 20 years. This has made possible techniques in crowd sourcing, propensity-score matching, educational data mining and learning analytics. This short paper outlines some of these approaches.

Keywords Evaluation methods · Intelligent tutoring systems · Adaptive learning environments

Introduction

We were pleased to learn that the article, "Evaluation methodologies for intelligent tutoring systems" by Mark and Greer, which came out two decades ago in 1993, is among the highly-cited articles in IJAIED. We were also surprised to learn that in the last 10 years the article was cited about as often as it was in the first 10 years. Being essentially a survey paper, it took a high-level look at evaluation, bringing together systems evaluation methods from computer science with better known educational and social science evaluation methods.

The special issue of IJAIED on Evaluation, in which the paper appeared, was a reaction to the rather weak evaluation methods commonly used to make claims about

✉ Jim Greer
jim.greer@usask.ca

[1] ARIES Laboraty, Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

[2] Chemical Heritage Foundation, Philadelphia, PA, USA

the effectiveness of ITSs and other learning environments at that time. Learner satisfaction surveys were too often the primary metric for evaluating systems in the 80's and early 90's. Laboratory-style control group studies with tiny sample sizes often were delivered with a chorus of excuses for insignificance. The evaluations of ITSs and other adaptive learning environments improved in quality and rigor over the years. We believe this has occurred in part due to that IJAIED special issue.

The broad concept of an ITS popular in 1993 has since narrowed. Now the moniker tends to suggest a stand-alone one-on-one computer tutor, and many AIEd researchers identify their work using different terms. Nonetheless, the 1993 paper, while labeled as evaluation methodologies for ITSs, applies readily to the wider range of smart, intelligent, or adaptive educational environments that people build today. What needs to be evaluated in educational assessments of today's learning environments differs little from what needed to be evaluated two decades ago. Similar questions about formative and summative evaluation are still addressed. Component and whole system validation and system comparison with human expert performance continues to be done routinely. It would be inaccurate to claim that a research project of today using a method of evaluation described in our 1993 paper was, in fact, influenced by our paper. We certainly did not invent these research methodologies.

While many of the approaches described in our original paper remain in use, new approaches to evaluating intelligent tutors and other adaptive learning environments continue to evolve, due in part to increases in computer speed, storage, and connectivity. Approaches that hold great promise for evaluating educational environments include:

- using the open web and crowd sourcing to evaluate systems
- comparing decisions made by adaptive learning environments and human experts
- using simulated learners in the evaluation of learning environments
- washing out selection bias while evaluating educational interventions
- examining learning curves for evaluating systems
- performing evaluations derived from mega quantities of micro measurements

We briefly comment below on each of these six approaches, indicating ways in which the survey presented in 1993 might be expanded.

The Wisdom (or not) of Crowds

Fuelled in part by the development of platforms such as Amazon's Mechanical Turk, crowdsourcing is used and studied by the business, gaming and Human Computer Interaction (HCI) communities. Types of crowdsourcing can be classed as contests, collaborative communities, labour markets and complementors, each of which has strengths and weaknesses (Boudreau and Lakhani 2013). With clever incentives involving fun, altruism and money, such approaches can draw on a potentially huge sample of users who can help to investigate a system and give extensive feedback.

When applying crowdsourcing to the evaluation of educational systems, there are issues of suitability of the volunteers to the teaching and learning goals of the system, the veracity of reactions and responses, and the seriousness with which volunteers take their interactions with the system. Kittur et al. (2008) examine the feasibility of crowdsourced evaluation with Mechanical Turk, noting its potential for gathering user measurements for formative studies with techniques such as surveys, rapid prototyping, and quantitative performance measures. They also report the danger of users "gaming" the systems, with disastrous results, if tasks are not carefully constructed. Crowdsourcing is likely to be most useful in gathering formative data from a wide base of users, and getting a quick reality check on a system before investing too much effort into a more controlled study. Crowdsourcing may not be a reliable or valid means of evaluating a system.

Kappa Scores and Human Experts

Measuring agreement between a pair (or among members of a panel) of experts using inter-rater reliability measures is a common occurrence. If one needs to compare the decisions or choices of an automated intelligent system to those of human experts, this type of criterion-based evaluation can help determine if the intelligent system is in line with what an expert might do. Kappa measures are a consensus estimate of inter-rater reliability. They are rarely used for tests of statistical significance. Rather Kappa measures give an indication of better versus poorer agreement between or among decision makers (Stemler 2004).

Several researchers in the AIED community have made use of Kappa in evaluation of behaviours of intelligent systems (Rourke et al. 2001). Use of Kappa in evaluating adaptive educational environments is usually aimed at demonstrating that the system's decisions align with what a human expert would do. There are two fundamental problems with such comparisons. When it comes to decisions or choices about educational actions, especially where subtlety of context plays a role, agreement among human experts is often quite low. This gives wide latitude for the decisions or choices of a computer system, while still remaining in line with human experts. The second problem is the "gold standard". Imagine that a computer system is far superior to any human expert in its judgments or decisions: Kappa measures of agreement would be quite low, even though the system was performing exceptionally well. When comparing a system with a human expert, a good Kappa measure tells us that the system is behaving neither much worse nor much better than a human expert, but not whether the human experts are a desirable basis of comparison.

Simulated Learners

Bias and small sample size continue to be problems for experimental evaluations of learning environments. Sometimes a handful of graduate students colleagues of the researcher are recruited to "evaluate" a system. Even if independent volunteers are recruited as experimental subjects, they may be few in number and may not reflect the target population for whom the system is intended.

Frustrated with small or unrepresentative sample sizes, some researchers have turned to synthetic subjects, i.e., simulated learners. Simulated learners are software systems or agents that can be designed to reflect the population of intended learners and scaled up to provide as large a sample size as desired. However, bias can be introduced by overfitting the simulated learner to the learning environment, which may lead to unrealistic learning predictions.

A good survey of the prospects for use of simulated learners in evaluating systems can be found in Vanlehn et al. (1994). More recent work by Matsuda et al. (2014) provides a good example of system evaluation with simulated learners. Simulated learners can help with formative assessments of learning environments or in validating specific components of a system, but testing with simulated learners cannot replace full summative evaluation and studies of educational impact. There is always the possibility that in a real situation, the students and the system may behave in unanticipated ways and surprise the experimenter!

Washing out Selection Bias

When formal experimental studies are carried out, a frequent concern is selection bias. Selection bias can occur when relying on volunteer subjects. Some people are more likely to volunteer than others and one cannot be sure that the volunteers are representative of the target population. This sort of problem occurs in classroom research settings where a subset of students decide to volunteer for an experimental treatment, such as use of a computer-based learning environment. Unless enough volunteers came forward to set up a randomized control group and a treatment group, the non-volunteers become a proxy control. Similar issues can occur when one section of a course uses a learning environment while another section of the same course acts as a proxy control group. Differences between the two groups may reflect variables besides the learning environment.

Propensity-score matching is a type of nonrandomized study that can be used to minimize selection bias and estimate the effects of treatments on outcomes. It works by matching students inside the treatment group with doppelgangers in a comparator group. Only corresponding students with a high degree of similarity on relevant variables should be paired (Rosenbaum and Rubin 1985; Austin 2011). It is assumed that to work well, the comparator group must be much larger than the treatment group so that a representative set of comparators can be identified as the control. This promising method has been only recently applied in the evaluation of learning environments, in particular in studying MOOCs, where the availability of large subject groups and advances in big data and machine learning make it feasible to detect well-matched comparator subjects (Brooks et al. 2015).

Use of Learning Curves

Educational researchers have long struggled to find effective ways to measure changes in learning due to some educational intervention. Standard measures of learning gains in pre-post measures when comparing some treatment and control are rarely significant. A more subtle means of measuring learning gains can be achieved by analyzing learning curves. Originating in machine learning, learning curves plot error rates in learner performance against an estimate of student knowledge level. This can be used to determine whether new modifications enhance or degrade performance. Martin et al.

(2005) offer a fine illustration of the methodology and provide an excellent guide to using learning curves in evaluation of adaptive educational systems. Martin et al. (2011) give further examples and discuss the use of performance curves in both formative and summative studies.

Educational Data Mining and Learning Analytics

Many educational systems gather fine-grained click-stream data documenting every learner action. Similarly video logs, eye-tracking data, LMS log files and MOOC platforms accumulate huge amounts of data. Attempting to make sense of very fine-grained log data poses a significant technical challenge for data mining. Meaningful patterns or features can sometimes be identified in advance through a priori understanding of the domain of learner activity. Meaningful patterns or features may also be found with machine learning and pattern matching techniques. Aggregating fine-grained data into meaningful patterns requires a data fusion and data engineering process as well as good understanding of statistical and data mining methods. Data visualization tools can also be helpful in recognizing patterns that may not have been anticipated in advance. These techniques may have applications in user modelling and profiling, domain modelling, and the study of learning components and instructional principles (U.S. Department of Education, 2012). Applying these techniques of educational data mining and learning analytics to large quantities of fine grained learner data is likely the most actively growing area in evaluation of learning environments (Baker & Siemens 2014).

Conclusion

Issues in the evaluation of ITSs and other adaptive intelligent learning support systems have not changed substantively over the past 20 years. However, methods for addressing them have developed both through refinements to existing evaluation approaches and newer techniques. Rigorous analysis of these refinements and new techniques is proceeding. With new challenges facing researchers and developers of AIEd systems, new approaches to evaluation undoubtedly will continue to evolve.

References

Austin, P. C. (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46 (3), [Special issue on Propensity Score Analysis], pp. 399–424.

Baker, R., & Siemens, G. (2014). *Educational data mining and learning analytics. Cambridge Handbook of the Learning Sciences*.

Boudreau, K. J., & Lakhani, K. R. (2013). Using the crowd as an innovation partner. *Harvard Business Review, April 2013, 91*(4), 60–69.

Brooks, C., Chavez, O., Tritz, J. & Teasley, S. (2015) Reducing Selection Bias in Quasi-Experimental Educational Studies. 5th International Conference on Learning Analytics and Knowledge (LAK'15), Poughkeepsie NY, March 2015.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical Turk. *In Proceedings of the SIGCHI conference on human factors in computing systems, 6*, 453–456.

Martin, B., Koedinger, K. R., Mitrovic, A., & Mathan, S. (2005). *On using learning curves to evaluate ITS. In 2005 AIED conference* (pp. 419–426). Amsterdam: Ios Press.

Martin, B., Mitrovic, A., Koedinger, K. R., & Mathan, S. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction, 21*(3), 249–283. doi:10.1007/s11257-010-9084-2.

Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2014). Teaching the Teacher: Tutoring SimStudent Leads to More Effective Cognitive Tutor Authoring. *International Journal of Artificial Intelligence in Education*, 1–34.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education (IJAIED), 12*, 8–22.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9 (4).

U.S. Department of Education, Office of Educational Technology (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. Washington, D.C.

Vanlehn, K., Ohlsson, S., & Nason, R. (1994). Applications of simulated students: an exploration. *Journal of Artificial Intelligence in Education, 5*, 135–135.