The State of Educational Data Mining in 2009: A Review and Future Visions

RYAN S.J.D. BAKER
Department of Social Science and Policy Studies
Worcester Polytechnic Institute
Worcester, MA USA

AND

KALINA YACEF
School of Information Technologies
University of Sydney
Sydney, NSW Australia

We review the history and current trends in the field of Educational Data Mining (EDM). We consider the methodological profile of research in the early years of EDM, compared to in 2008 and 2009, and discuss trends and shifts in the research conducted by this community. In particular, we discuss the increased emphasis on prediction, the emergence of work using existing models to make scientific discoveries ("discovery with models"), and the reduction in the frequency of relationship mining within the EDM community. We discuss two ways that researchers have attempted to categorize the diversity of research in educational data mining research, and review the types of research problems that these methods have been used to address. The most-cited papers in EDM between 1995 and 2005 are listed, and their influence on the EDM community (and beyond the EDM community) is discussed.

1. INTRODUCTION

The year 2009 finds the nascent research community of Educational Data Mining (EDM) growing and continuing to develop. This summer, the second annual international conference on Educational Data Mining, EDM2009, was held in Cordoba, Spain, and plans are already underway for the third international conference to occur in June 2010 in Pittsburgh, USA. With the publication of this issue, the Educational Data Mining community now has its own journal, the Journal of Educational Data Mining. In addition, it is anticipated that in the next year, Chapman & Hall/CRC Press, Taylor and Francis Group will publish the first Handbook of Educational Data Mining.

This moment in the educational data mining community's history provides a unique opportunity to consider where we come from and where we are headed. In this article, we will review some of the major areas and trends in EDM, some of the most prominent articles in the field (both those published in specific EDM venues, and in other venues where top-quality EDM research can be found), and consider what the future may hold for our community.

2. WHAT IS EDM?

The Educational Data Mining community website, www.educationaldatamining.org, defines educational data mining as follows: "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in."

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [Witten and Frank 1999]. It has been proposed that educational data mining methods are often different from standard data mining methods, due to the need to explicitly account for (and the opportunities to exploit) the multi-level hierarchy and non-independence in educational data [Baker in press]. For this reason, it is increasingly common to see the use of models drawn from the psychometrics literature in educational data mining publications [Barnes 2005; Desmarais and Pu 2005; Pavlik et al. 2008].

3. EDM METHODS

Educational data mining methods are drawn from a variety of literatures, including data mining and machine learning, psychometrics and other areas of statistics, information visualization, and computational modeling. Romero and Ventura [2007] categorize work in educational data mining into the following categories:

- Statistics and visualization
- Web mining
    - Clustering, classification, and outlier detection
    - Association rule mining and sequential pattern mining
    - Text mining

This viewpoint is focused on applications of educational data mining to web data, a perspective that accords with the history of the research area. To a large degree, educational data mining emerged from the analysis of logs of student-computer interaction. This is perhaps most clearly shown by the name of an early EDM workshop (according to the EDM community website, the third workshop in the history of the community – the workshop at AIED2005 on Usage Analysis in Learning Systems [Choquet et al. 2005]) . The methods listed by Romero and Ventura as web mining methods are quite prominent in EDM today, both in mining of web data and in mining other forms of educational data.

A second viewpoint on educational data mining is given by Baker [in press], which classifies work in educational data mining as follows:

- Prediction
  - Classification
  - Regression
  - Density estimation
- Clustering
- Relationship mining
  - Association rule mining
  - Correlation mining
  - Sequential pattern mining
  - Causal data mining
- Distillation of data for human judgment
- Discovery with models

The first three categories of Baker's taxonomy of educational data mining methods would look familiar to most researchers in data mining (the first set of sub-categories are directly drawn from Moore's categorization of data mining methods [Moore 2006]). The fourth category, though not necessarily universally seen as data mining, accords with Romero and Ventura's category of statistics and visualization, and has had a prominent place both in published EDM research [Kay et al. 2006], and in theoretical discussions of educational data mining [Tanimoto 2007].

The fifth category of Baker's EDM taxonomy is perhaps the most unusual category, from a classical data mining perspective. In discovery with models, a model of a phenomenon is developed through any process that can be validated in some fashion (most commonly, prediction or knowledge engineering), and this model is then used as a component in another analysis, such as prediction or relationship mining. Discovery with models has become an increasingly popular method in EDM research, supporting sophisticated analyses such as which learning material sub-categories of students will most benefit from [Beck and Mostow 2008], how different types of student behavior impact students' learning in different ways [Cocea et al. 2009], and how variations in intelligent tutor design impact students' behavior over time [Jeong and Biswas 2008].

Historically, relationship mining methods of various types have been the most prominent category in EDM research. In Romero & Ventura's survey of EDM research from 1995 to 2005, 60 papers were reported that utilized EDM methods to answer research questions of applied interest (according to a post-hoc analysis conducted for the current article). 26 of those papers (43%) involved relationship mining methods. 17 more papers (28%) involved prediction methods of various types. Other methods were less common. The full distribution of methods across papers is shown in Figure 1.

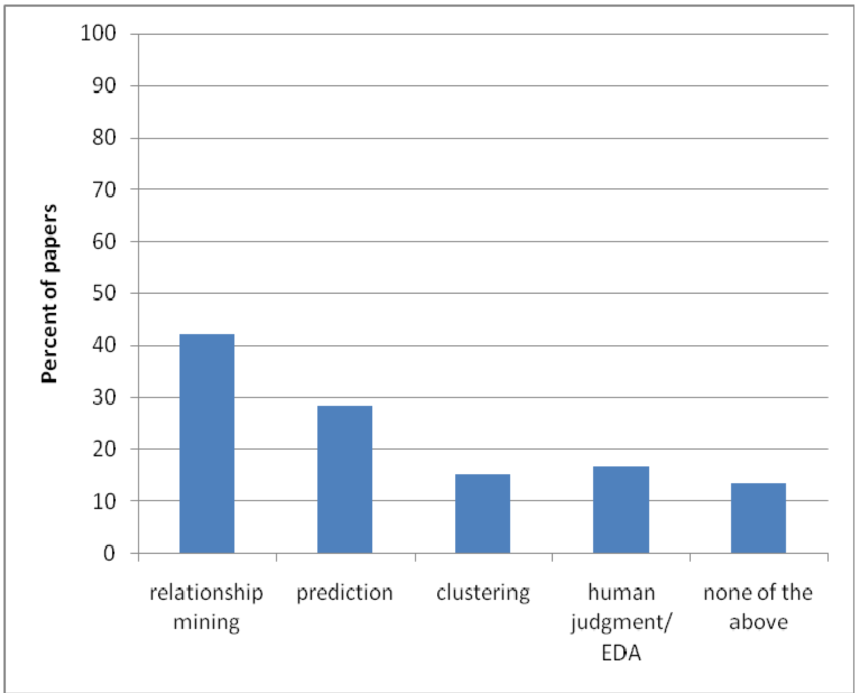

Figure 1. The proportion of papers involving each type of EDM method, in Romero & Ventura's [2007] 1995-2005 survey. Note that papers can use multiple methods, and thus some papers can be found in multiple categories.

4. KEY APPLICATIONS OF EDM METHODS

Educational Data Mining researchers study a variety of areas, including individual learning from educational software, computer supported collaborative learning, computer-adaptive testing (and testing more broadly), and the factors that are associated with student failure or non-retention in courses.

Across these domains, one key area of application has been in the improvement of student models. Student models represent information about a student's characteristics or state, such as the student's current knowledge, motivation, meta-cognition, and attitudes.

Modeling student individual differences in these areas enables software to respond to those individual differences, significantly improving student learning [Corbett 2001]. Educational data mining methods have enable researchers to model a broader range of potentially relevant student attributes in real-time, including higher-level constructs than were previously possible. For instance, in recent years, researchers have used EDM methods to infer whether a student is gaming the system [Baker et al. 2004], experiencing poor self-efficacy [McQuiggan et al. 2008], off-task [Baker 2007], or even if a student is bored or frustrated [D'Mello et al. 2008]. Researchers have also been able to extend student modeling even beyond educational software, towards figuring out what factors are predictive of student failure or non-retention in college courses or in college altogether [Dekker et al. 2009; Romero et al. 2008; Superby et al. 2006].

A second key area of application of EDM methods has been in discovering or improving models of a domain's knowledge structure. Through the combination of psychometric modeling frameworks with space-searching algorithms from the machine learning literature, a number of researchers have been able to develop automated approaches that can discover accurate domain structure models, directly from data. For instance, Barnes [2005] has developed algorithms which can automatically discover a Q-Matrix from data, and Desmarais & Pu [2005] and Pavlik et al [Pavlik et al. 2009; Pavlik, Cen, Wu and Koedinger 2008] have developed algorithms for finding partial order knowledge structure (POKS) models that explain the interrelationships of knowledge in a domain.

A third key area of application of EDM methods has been in studying pedagogical support (both in learning software, and in other domains, such as collaborative learning behaviors), towards discovering which types of pedagogical support are most effective, either overall or for different groups of students or in different situations [Beck and Mostow 2008; Pechenizkiy et al. 2008]. One popular method for studying pedagogical support is learning decomposition [Beck and Mostow 2008]. Learning decomposition fits exponential learning curves to performance data, relating a student's later success to the amount of each type of pedagogical support the student received up to that point. The relative weights for each type of pedagogical support, in the best-fit model, can be used to infer the relative effectiveness of each type of support for promoting learning.

A fourth key area of application of EDM methods has been in looking for empirical evidence to refine and extend educational theories and well-known educational phenomena, towards gaining deeper understanding of the key factors impacting learning, often with a view to design better learning systems. For instance Gong, Rai and

Heffernan [2009] investigated the impact of self-discipline on learning and found that, whilst it correlated to higher incoming knowledge and fewer mistakes, the actual impact on learning was marginal. Perera et al. [2009] used the Big 5 theory for teamwork as a driving theory to search for successful patterns of interaction within student teams. Madhyastha and Tanimoto [2009] investigated the relationship between consistency and student performance with the aim to provide guidelines for scaffolding instruction, basing their work on prior theory on the implications of consistency in student behavior [Abelson 1968].

5. IMPORTANT TRENDS IN EDUCATIONAL DATA MINING RESEARCH

In this section, we consider how educational data mining has developed in recent years, and investigate what some of the major trends are in EDM research. In order to investigate what the trends are, we analyze what researchers were studying previously, and what they are studying now, towards understanding what is new and what attributes EDM research has had for some time.

5.1. Prominent Papers From Early Years

One way to see where EDM has been is to look at which articles were the most influential in its early years. We have an excellent resource, in Romero and Ventura's (2007) survey. This survey gives us a comprehensive list of papers, published between 1995 and 2005, which are seen as educational data mining by a prominent pair of authorities in EDM (beyond authoring several key papers in EDM, Romero and Ventura were conference chairs of EDM2009). To determine which articles were most influential, we use how many citations each paper received, a bibliometric or scientometric measure often used to indicate influence of papers, researchers, or institutions. As Bartneck and Hu [2009] have noted, Google Scholar, despite imperfections in its counting scheme, is the most comprehensive source for citations – particularly for the conferences which are essential for understanding Computer Science research.

The top 8 most cited applied papers in Romero and Ventura's survey (as of September 9, 2009) are listed in Table 1. These articles have been highly influential, both on educational data mining researchers, and on related fields; as such, they exemplify many of the key trends in our research community.

The most cited article, [Zaïane 2001], suggests an application for data mining, using it to study on-line courses. This article proposes and evangelizes EDM's usefulness, and in this fashion was highly influential to the formation of our community.

The second and fourth most cited articles, [Zaïane 2002] and [Tang and McCalla 2005] center around how educational data mining methods (specifically association rules, and clustering to support collaborative filtering) can support the development of more sensitive and effective e-learning systems. As in his other paper in this list, Zaiane makes a detailed and influential proposal as to how educational data mining methods can make an impact on e-learning systems. Tang and McCalla report an instantiation of such a system, which integrates clustering and collaborative filtering to recommend content to students. The authors present a study conducted with simulated students; successful evaluation of the system with real students is presented in [Tang and McCalla 2004].

The third most-cited article, [Baker, Corbett and Koedinger 2004] gives a case study on how educational data mining methods (specifically prediction methods) can be used to open new research areas, in this case the scientific study of gaming the system (attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning the material). Though this topic had seen some prior interest (including [Aleven and Koedinger 2001; Schofield 1995; Tait et al. 1973]), publication and research into this topic exploded after it became clear that educational data mining now opened this topic to concrete, quantitative, and fine-grained analysis.

The fifth and sixth most cited articles, [Merceron and Yacef 2003] and [Romero et al. 2003], present tools that can be used to support educational data mining. This theme is carried forward in these groups' later work [Merceron and Yacef 2005; Romero, Ventura, Espejo and Hervas 2008], and in EDM tools developed by other researchers [Donmez et al. 2005].

The seventh most cited article [Beck and Woolf 2000] shows how educational data mining prediction methods can be used to develop student models. They use a variety of variables to predict whether a student will make a correct answer. This work has inspired a great deal of later educational data mining work – student modeling is a key theme in modern educational data mining, and the paradigm of testing EDM models' ability to predict future correctness – advocated strongly by Beck & Woolf – has become very common (eg [Beck 2007; Mavrikis 2008]) .

Table 1. The top 8 most cited papers, in Romero & Ventura's 1995-2005 survey.
Citations are from Google Scholar, retrieved 9 September, 2009.

Article	Citations
Zaïane, O. (2001). Web usage mining for a better web-based learning	110

environment. *Proceedings of Conference on Advanced Technology for Education*, 60–64.	
Zaïane, O. (2002). Building a recommender agent for e-learning systems. *Proceedings of the International Conference on Computers in Education,* 55–59.	89
Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.	83
Tang, T., McCalla, G. (2005) Smart recommendation for an evolving e-learning system: architecture and experiment, *International Journal on E-Learning*, 4 (1), 105–129.	63
Merceron, A., Yacef, K. (2003). A web-based tutoring tool with mining facilities to improve learning and teaching. *Proceedings of the 11th International Conference on Artificial Intelligence in Education*, 201–208.	54
Romero, C., Ventura, S., de Bra, P., & Castro, C. (2003). Discovering prediction rules in aha! courses. *Proceedings of the International Conference on User Modeling*, 25–34.	46
Beck, J., & Woolf, B. (2000). High-level student modeling with machine learning. *Proceedings of the 5th International Conference on Intelligent Tutoring Systems,* 584–593.	43
Dringus, L.P., Ellis, T. (2005) Using data mining as a strategy for assessing asynchronous discussion forums, *Computer and Education Journal* , 45, 141–160.	37

5.2. Shift In Paper Topics Over The Years

As discussed earlier in this paper (see Figure 1), relationship mining methods of various types were the most prominent type of EDM research between 1995 and 2005. 43% of papers in those years involved relationship mining methods. Prediction was the second most prominent research area, with 28% of papers in those years involving prediction methods of various types. Human judgment/exploratory data analysis and clustering followed with (respectively) 17% and 15% of papers.

A very different pattern is seen in the papers from the first two years of the Educational Data Mining conference [Baker et al. 2008; Barnes et al. 2009], as shown in Figure 2. Whereas relationship mining was dominant between 1995 and 2005, in 2008-

2009 it slipped to fifth place, with only 9% of papers involving relationship mining. Prediction, which was in second place between 1995 and 2005, moved to the dominant position in 2008-2009, representing 42% of EDM2008 papers. Human judgment/exploratory data analysis and clustering remain in approximately the same position in 2008-2009 as 1995-2005, with (respectively) 12% and 15% of papers.

A new method, significantly more prominent in 2008-2009 than in earlier years, is discovery with models. Whereas no papers in Romero & Ventura's survey involved discovery with models, by 2008-2009 it has become the second most common category of EDM research, representing 19% of papers.

Another key trend is the increase in prominence of modeling frameworks from Item Response Theory, Bayes Nets, and Markov Decision Processes. These methods were rare at the very beginning of educational data mining, began to become more prominent around 2005 (appearing, for instance, in [Barnes 2005] and [Desmarais and Pu 2005]), and were found in 28% of the papers in EDM2008 and EDM2009. The increase in the commonality of these methods is likely a reflection of the integration of researchers from the psychometrics and student modeling communities into the EDM community.

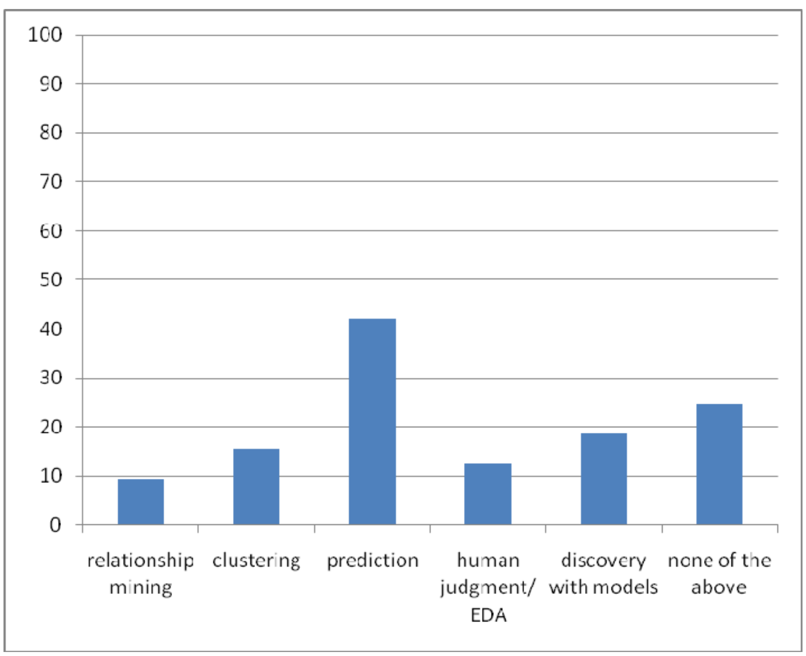

Figure 2. The proportion of papers involving each type of EDM method, in the proceedings of Educational Data Mining 2008 and 2009 [Baker, Barnes and Beck 2008; Barnes, Desmarais, Romero and Ventura 2009]. Note that papers can use multiple methods, and thus some papers can be found in multiple categories.

It is worth noting that educational data mining publications in 2008 and 2009 are not limited solely to those appearing in the proceedings of the conference (though our analysis in this paper was restricted to those publications). One of the notable metrics of our community's growth is that the proceedings of EDM2008 and EDM2009 alone accounted for approximately as many papers as were published in the first 10 years of the community's existence (according to Romero & Ventura's review). Hence, EDM appears to be growing in size rapidly, and the next major review of the field is likely to be a time-consuming process. However, we encourage future researchers to conduct such a survey. In general, it will be very interesting to see how the methodological trends exposed in Figures 1 and 2 develop in the next few years.

5.3. Emergence of public data and public data collection tools

One interesting difference between the work in EDM2008 and EDM2009, and earlier educational data mining work, is where the educational data comes from. Between 1995 and 2005, data almost universally came from the research group conducting the analysis – that is to say, in order to do educational data mining research, a researcher first needed to collect their own educational data.

This necessity appears to be disappearing in 2008, due to two developments. First, the Pittsburgh Science of Learning Center has opened a public data repository, the PSLC DataShop [Koedinger et al. 2008], which makes substantial quantities of data from a variety of online learning environments available, for free, to any researcher worldwide. 14% of the papers published in EDM2008 and EDM2009 utilized data publicly available from the PSLC DataShop.

Second, researchers are increasingly frequently instrumenting existing online course environments used by large numbers of students worldwide, such as Moodle and WebCAT. 12% of the papers in EDM2008 and EDM2009 utilized data coming from the instrumentation of existing online courses.

Hence, around a quarter of the papers published at EDM2008 and EDM2009 involved data from these two readily available sources. If this trend continues, there will be significantly benefits for the educational data mining community. Among them, it will become significantly easier to externally validate an analysis. If a researcher does an analysis that produces results that seem artifactual or "too good to be true", another researcher can download the data and check for themselves. A second benefit is that researchers will be more able to build on others' past efforts. As reasonably predictive models of domain structure and student moment-to-moment knowledge become available for public data sets, other researchers will be able to test new models of these phenomena in comparison to a strong baseline, or to develop new models of higher grain-size constructs that leverage these existing models. The result is a science of education that is more concrete, validated, and progressive than was previously possible.

6. CONCLUSIONS

The publication of this first issue of the Journal of Educational Data Mining finds the field growing rapidly, but also in a period of transition. The advent of the EDM conference series has led to a significant increase in the volume of research published. In addition, public educational databases and tools for instrumenting online courses increase the accessibility of educational data to a wider pool of individuals, lowering the barriers to becoming an educational data mining researcher. Hence further growth can be expected.

It is possible that these trends will make educational data mining an increasingly international community as well. Between the papers in Romero & Ventura and the EDM2008 and EDM2009 proceedings, it can be seen that the EDM community remains focused in North America, Western Europe, and Australia/New Zealand, with relatively lower participation from other regions. However, the increasing accessibility of relevant and usable educational data has the potential to "lower the barriers" to entry for researchers in the rest of the world.

Recent years have also seen major changes in the types of EDM methods that are used, with prediction and discovery with models increasing while relationship mining becomes rarer. It will be interesting to see how these trends shift in the years to come, and what new types of research will emerge from the increase in discovery with models, a method prominent in cognitive modeling and bioinformatics, but thus far rare in education research.

At this point, educational data mining methods have had some level of impact on education and related interdisciplinary fields (such as artificial intelligence in education, intelligent tutoring systems, and user modeling). However, so far only a handful of articles have achieved more than 50 citations (as shown in Table 1), indicating that there is still considerable scope for an increase in educational data mining's scientific influence. It is hoped that this journal will play a role in raising the profile of the educational data mining field and bringing to educational research the mathematical and scientific rigor that similar methods have previously brought to cognitive psychology and biology.

ACKNOWLEDGEMENTS

We thank Cristobal Romero and Sebastian Ventura for their excellent review in 2005 of the state of Educational Data Mining, which influenced our article – and the field – considerably. We thank support from the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation, award number SBE-0354420.

REFERENCES

ABELSON, R. 1968. *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago.

ALEVEN, V. and KOEDINGER, K.R. 2001. Investigations into help seeking and learning with a Cognitive Tutor. In *Proceedings of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments*, 47-58. R. LUCKIN Ed.

BAKER, R.S., CORBETT, A.T. and KOEDINGER, K.R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, Maceio, Brazil, 531-540.

BAKER, R.S.J.D. 2007. Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. In *Proceedings of the ACM CHI 2007: Computer-Human Interaction conference*, 1059-1068.

BAKER, R.S.J.D. in press. Data Mining For Education. In *International Encyclopedia of Education (3rd edition)*, B. MCGAW, PETERSON, P., BAKER Ed. Elsevier, Oxford, UK.

BAKER, R.S.J.D., BARNES, T. and BECK, J.E. 2008. 1st International Conference on Educational Data Mining, Montreal, Quebec, Canada.

BARNES, T. 2005. The q-matrix method: Mining student response data for knowledge. In *Proceedings of the AAAI-2005 Workshop on Educational Data Mining*.

BARNES, T., DESMARAIS, M., ROMERO, C. and VENTURA, S. 2009. Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings, Cordoba, Spain.

BARTNECK, C. and HU, J. 2009. Scientometric Analysis of the CHI Proceedings. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI2009)*, 699-708.

BECK, J. and WOOLF, B. 2000. High-level student modeling with machine learning. In *Proceedings of the International Conference on Intelligent tutoring systems*, 584-593.

BECK, J.E. 2007. Difficulties in inferring student knowledge from observations (and why you should care). *Proceedings of the AIED2007 Workshop on Educational Data Mining*, 21-30.

BECK, J.E. and MOSTOW, J. 2008. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 353-362.

CHOQUET, C., LUENGO, V. and YACEF, K. 2005. Proceedings of "Usage Analysis in Learning Systems" workshop, held in conjunction with AIED 2005, Amsterdam, The Netherlands, July 2005.

COCEA, M., HERSHKOVITZ, A. and BAKER, R.S.J.D. 2009. The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

CORBETT, A.T. 2001. Cognitive Computer Tutors: Solving the Two-Sigma Problem. In *Proceedings of the International Conference on User Modeling*, 137-147.

D'MELLO, S.K., CRAIG, S.D., WITHERSPOON, A.W., MCDANIEL, B.T. and GRAESSER, A.C. 2008. Automatic Detection of Learner's Affect from Conversational Cues. *User Modeling and User-Adapted Interaction 18*, 45-80.

DEKKER, G., PECHENIZKIY, M. and VLEESHOUWERS, J. 2009. Predicting Students Drop Out: A Case Study. In *Proceedings of the International Conference on Educational Data Mining*, Cordoba, Spain, T. BARNES, M. DESMARAIS, C. ROMERO and S. VENTURA Eds., 41-50.

DESMARAIS, M.C. and PU, X. 2005. A Bayesian Student Model without Hidden Nodes and Its Comparison with Item Response Theory. *International Journal of Artificial Intelligence in Education 15*, 291-323.

DONMEZ, P., ROSÉ, C., STEGMANN, K., WEINBERGER, A. and FISCHER, F. 2005. Supporting CSCL with automatic corpus analysis technology. In *Proceedings of the International Conference of Computer Support for Collaborative Learning (CSCL 2005)*, 125-134.

GONG, Y., RAI, D., BECK, J. and HEFFERNAN, N. 2009. Does Self-Discipline Impact Students' Knowledge and Learning? In *Proceedings of the 2nd International Conference on Educational Data Mining*, 61-70.

JEONG, H. and BISWAS, G. 2008. Mining Student Behavior Models in Learning-by-Teaching Environments. In *Proceedings of the 1st International Conference on Educational Data Mining*, 127-136.

KAY, J., MAISONNEUVE, N., YACEF, K. and REIMANN, P. 2006. The Big Five and Visualisations of Team Work Activity. In *Intelligent Tutoring Systems*, M. IKEDA, K.D. ASHLEY and T.-W. CHAN Eds. Springer-Verlag, Taiwan, 197-206.

KOEDINGER, K.R., CUNNINGHAM, K., A., S. and LEBER, B. 2008. An open repository and analysis tools for fine-grained, longitudinal learner data. In *Proceedings of the 1st International Conference on Educational Data Mining*, 157-166.

MADHYASTHA, T. and TANIMOTO, S. 2009. Student Consistency and Implications for Feedback in Online Assessment Systems. In *Proceedings of the 2nd International Conference on Educational Data Mining*, 81-90.

MAVRIKIS, M. 2008. Data-driven modeling of students' interactions in an ILE. In *Proceedings of the 1st International Conference on Educational Data Mining*, 87-96.

MCQUIGGAN, S., MOTT, B. and LESTER, J. 2008. Modeling Self-Efficacy in Intelligent Tutoring Systems: An Inductive Approach. *User Modeling and User-Adapted Interaction 18*, 81-123.

MERCERON, A. and YACEF, K. 2003. A Web-based Tutoring Tool with Mining Facilities to Improve Learning and Teaching. In *11th International Conference on Artificial Intelligence in Education.*, F. VERDEJO and U. HOPPE Eds. IOS Press, Sydney, 201-208.

MERCERON, A. and YACEF, K. 2005. Educational Data Mining: a Case Study. In *Artificial Intelligence in Education (AIED2005)*, C.-K. LOOI, G. MCCALLA, B. BREDEWEG and J. BREUKER Eds. IOS Press, Amsterdam, The Netherlands, 467-474.

MOORE, A.W. 2006. Statistical Data Mining Tutorials. Downloaded 1 August 2009 from http://www.autonlab.org/tutorials/

PAVLIK, P., CEN, H. and KOEDINGER, K.R. 2009. Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models. In *Proceedings of the 2nd International Conference on Educational Data Mining*, 121-130.

PAVLIK, P., CEN, H., WU, L. and KOEDINGER, K. 2008. Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor. In *Proceedings of the 1st International Conference on Educational Data Mining*, 77-86.

PECHENIZKIY, M., CALDERS, T., VASILYEVA, E. and DE BRA, P. 2008. Mining the Student Assessment Data: Lessons Drawn from a Small Scale Case Study. In *Proceedings of the 1st International Conference on Educational Data Mining*, 187-191.

PERERA, D., KAY, J., KOPRINSKA, I., YACEF, K. and ZAIANE, O. 2009. Clustering and sequential pattern mining to support team learning. *IEEE Transactions on Knowledge and Data Engineering 21*, 759-772

ROMERO, C. and VENTURA, S. 2007. Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications 33*, 125-146.

ROMERO, C., VENTURA, S., DE BRA, P. and CASTRO, C. 2003. Discovering prediction rules in aha! courses. In *Proceedings of the International Conference on User Modeling*, 25–34.

ROMERO, C., VENTURA, S., ESPEJO, P.G. and HERVAS, C. 2008. Data Mining Algorithms to Classify Students. In *Proceedings of the 1st International Conference on Educational Data Mining*, 8-17.

SCHOFIELD, J. 1995. *Computers and Classroom Culture*. Cambridge University Press Cambridge, UK.

SUPERBY, J.F., VANDAMME, J.-P. and MESKENS, N. 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, 37-44.

TAIT, K., HARTLEY, J.R. and ANDERSON, R.C. 1973. Feedback Procedures in Computer-Assisted Arithmetic Instruction. *British Journal of Educational Psychology 43*, 161-171.

TANG, T. and MCCALLA, G. 2004. Utilizing Artificial Learners to Help Overcome the Cold-Start Problem in a Pedagogically-Oriented Paper Recommendation System. In *Proceedings of the International Conference on Adaptive Hypermedia*, 245-254.

TANG, T. and MCCALLA, G. 2005. Smart recommendation for an evolving e-learning system: architecture and experiment. *International Journal on E-Learning 4*, 105-129.

TANIMOTO, S.L. 2007. Improving the Prospects for Educational Data Mining. In *Proceedings of the Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling, at the 11th International Conference on User Modeling (UM 2007)*, 106-110.

WITTEN, I.H. and FRANK, E. 1999. *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Fransisco, CA.

ZAÏANE, O. 2001. Web usage mining for a better web-based learning environment. In *Proceedings of conference on advanced technology for education*, 60-64.

ZAÏANE, O. 2002. Building a recommender agent for e-learning systems. In *Proceedings of the International Conference on Computers in Education*, 55–59.