# Analyzing the relationships between learning analytics, educational data mining and AI for education

Hugues Labarthe[1,2], Vanda Luengo[2] and François Bouchet[2]

[1] Incubateur Académique, Rectorat de Créteil, 94700 Créteil, France
[2] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, 75005 Paris, France
hugues.labarthe@ac-creteil.fr, vanda.luengo@lip6.fr, francois.bouchet@lip6.fr

**Abstract.** Baker and Siemens have well explained the theoretical differences and similarities between the educational data mining (EDM) and learning analytics (LA) communities in their 2012 seminal paper, in which they also wished for bridging the gap between both communities. Moreover, since its creation as an independent conference in 2009, EDM has been evolving in parallel with the intelligent tutoring systems (ITS) / artificial intelligence for education (AIED) community. But what are the actual links that exist between these three communities in terms of members and research topics: to what extent do they overlap and work together? Are they getting closer from each other or drifting apart? Is each community specific to researchers with different backgrounds, modeling and analysis techniques? Those are some of the questions we investigate using a quantitative analysis led between 2007 and 2017 through: a social network analysis of the 3 communities, involving the 1822 scientists who participated in program committees and/or appeared as authors of the associated journals (IJAIED, JEDM and JLA); and a text analysis of abstracts of articles published in these journals. Results reveal the clear differences between these communities, their topics, practices and research methods.

**Keywords:** learning analytics, educational data mining, artificial intelligence in education, social network analysis, text analysis, communities

## 1. Introduction

At the beginning of the 2010s, two communities progressively structured themselves to study learning data: the *Society for Learning Analytics Research* (SoLAR) and the *International Educational Data Mining Society* (IEDMS). In the meantime, the International AIED Society, gathered around the encompassing "Artificial Intelligence for EDucation" (AIED) theme, also started to analyze more and more data coming for their systems (in particular, intelligent tutors). Thus, three research communities have been tackling similar issues, and there has now been enough history for a data-based approach (valued by all three communities) to examine what distinguishes them and what brings them together.

The theme of *Educational Data Mining* first appeared during the ITS (*Intelligent Tutor Systems*) conference in Montreal in 2000 [1]. But it is really in 2005, with the first workshop on EDM held in Pittsburgh in conjunction with the AAAI (*Association*

*for the Advancement of Artificial Intelligence*) conference that the theme started to take off. Most of the research work presented at that time were led on data coming from ITS [2]. The first state of the art work was published in 2007 by Romero et Ventura [3], and was followed by the creation of the yearly EDM conference in 2008, and of its associated journal, the *Journal of Educational Data Mining* (JEDM), in 2009. In parallel, and independently, the *Society for Learning Analytics Research* (SoLAR) was founded in 2011 with its associated yearly conference, LAK (*Learning Analytics and Knowledge*), followed in 2014 by its own journal, the *Journal of Learning Analytics* (JLA). Finally, the AIED community has been structured for three decades around two alternating bi-yearly conferences, AIED (*Artificial Intelligence for Education*), which became yearly in 2017, and ITS (*Intelligent Tutoring Systems*), as well as a journal, IJAIED (*International Journal of Artificial Intelligence for Education*).

Very early on, the two new communities have acknowledged each other and the differences that exist between them, mainly in the background of its lead members (semantic web for LA, educational software for EDM), the analysis techniques they mostly use (social network analysis for LA, more machine learning for EDM), and their overall goal (empowering learners and teachers while leaving them in charge for LA, automated adaptation by the computer for EDM). Those key differences are well summarized in [4], in which the authors also call for joining the forces of the two communities to build upon each other's strengths. Although the interactions have been happening [5], both communities have also kept their respective identities [6], which have been established through publications to federate their respective domains [7–9].

Overall, a decade after the first EDM conference, and three after the first ITS, the three communities are thriving, and we can wonder about the relationships between each other and their respective impact on education. We decided to study three types of data: (1) the reviewers for the conferences associated to each community (AIED/ITS, EDM, LAK); (2) the authors of the papers published in the journals associated to each community (IJAIED, JEDM, JLA); (3) the abstracts of the papers published in the journals associated to each community. Using these datasets, we performed exploratory analyses of the overlap of the communities as well as of their individual specificities.

## 2. Data collection and cleaning

For each of the aforementioned datasets, we decided to consider a period of 11 years (2007-2017), which encompasses the whole existence of the EDM community. Although it may appear to give an emphasis to the data from that community, the LA community has published overall more intensively since its birth in 2011 (*cf.* table 1 further), and we therefore believe the 4 extra years are not affecting the validity of our results. Regarding the AIED community, although we had access to older data, we believed the changes in terms of popular scientific topics and approaches over time did not justify including it, and that it made more sense to use a similar period of 11 years.

The first dataset (reviewers) was collected mostly manually by extracting the list of reviewers' names included in the proceedings of each conference. We extracted the names from PDF version of the proceedings, selecting any name listed under the "Program Committee" and "Reviewers" sections, excluding others such as "Conference chairs" or "Organization committee". The choice of reviewers instead of authors was

justified by the fact that many conferences authors may appear only once, and that authoring a single paper in a conference does not necessarily imply a tight relationship with the associated community. Conversely, being invited to review papers for a conference usually indicates a sustained link (including but not limited to authorship), more relevant for a community analysis like the one we wanted to perform.

The second (authors) and third (abstracts) datasets were extracted automatically using a webcrawler tool (Scrapy) specially configured to extract from each website the information relative to published papers (title, authors, abstract, keywords, volume, issue, year). For IJAIED, information was extracted from both the Springer and ijaied.org websites, but only the ijaied.org data was kept because the Springer data started in 2013 only. We excluded from these datasets articles explicitly identified as an editorial, including guest editorials for special sections in the case of JLA, to focus only on research papers. A tedious review of names, surnames and even positions resulted in creating a single table, reducing a list of 4026 names to 1505 individuals. The abstracts were analyzed using Python packages for text analysis and visualization.

Overall, when not counting twice authors and reviewers who published/reviewed more than once for a given journal/conference, we see in Table 1 that AIED remains logically the dominant community of the three, with 687 reviewers and 386 authors. In terms of reviewers, EDM and LA are very close from each other and are far less than half of the reviewers for AIED. However, in terms of journal authors, despite a later start, the LA community has published almost 2.5 times more articles than the EDM one, with almost twice more individual authors.

**Table 1.** Conferences, Journals, Authors and Reviewers between 2007 and 2017

| Communities | Conferences | Conf. reviewers * | J. Issues | J. Articles** | J. Authors* |
|---|---|---|---|---|---|
| AIED | 11 | 687 | 11 | 161 | 386 |
| EDM | 10 | 238 | 9 | 54 | 151 |
| LA | 7 | 233 | 4 | 132 | 267 |
| Total * | 28 | 990 | 33 | 349 | 748 |

* Double count free ** Editorials free

## 3. Conference reviewers community analysis

First, we focus on the conference reviewers' dataset to analyze the evolution of the reviewers' network among the three communities from 2007 to 2017. In a decade, the number of scientists reviewing for each year conferences' papers has increased by 103%, reaching 415 reviewers in 2017, showing the significant vitality of these research fields (*cf.* Table 2). Moreover, the total number of scientists involved in these 28 conferences has increased considerably from 204 to 990 (+385%), showing that the growth in yearly reviewers came from a community more than twice larger overall. Despite a small drop in the number of yearly reviewers from 2008 to 2010, the number of scientists involved in these reviews has never stopped increasing, with two peaks: +29% in 2008 for the first EDM conference and +36% in 2011 for the first LAK conference.

**Table 2.** The Continuous Enlargement of the Program Committees, from 2007 to 2017

| **Reviewers** | **2007** | **2008** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total number | 204 | 131 | 136 | 124 | 244 | 265 | 280 | 293 | 266 | 314 | 415 |
| Cumulated | 204 | 263 | 306 | 330 | 449 | 535 | 623 | 681 | 761 | 848 | 990 |
| Annual growth % | | +29 | +16 | +8 | +36 | +19 | +16 | +9 | +12 | +11 | +17 |

Due to its anteriority in the field, we could make the hypothesis that the AIED/ITS conferences provided most of the reviewers for the two other communities. To test this hypothesis, we examined the overlap of reviewers between LAK/EDM and the AIED/ITS conferences (*cf.* Table 3). Until 2014, the AIED community has recruited two thirds of the reviewers, with 88 % of them exclusively dedicated to its Program Committee. Then, it decreases to only half of the total, and 70-75% of exclusive reviewers. It is a sign not only of the growth of the LA/EDM communities, but also of the increased porosity with the older AIED community. As we can see in Table 3, the two new communities have been relying upon this first one, at least at their beginning. These communities have progressively grown from one fifth of the network together, to one third each, with LAK having the fastest growth. The proportion of cross-conferences' reviewers for more than one conference has remained constant overall, at around 8-12%, with two peaks to 15% in 2010, and to 14-19% in 2015-2017.

**Table 3.** Total numbers of reviewers by and between conference

| **Conf.** | **Reviewers** | **2007** | **2008** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| All | Total # | 204 | 131 | 136 | 124 | 244 | 265 | 280 | 293 | 266 | 314 | 415 |
| AIED ITS | *% Total* | *100* | *89* | *93* | *81* | *72* | *72* | *66* | *68* | *46* | *45* | *47* |
| | *% Exclusive* | *100* | *91* | *90* | *81* | *88* | *88* | *88* | *88* | *69* | *70* | *75* |
| EDM | *% Total* | | *20* | *16* | *34* | *23* | *21* | *23* | *21* | *35* | *38* | *32* |
| | *% Exclusive* | | *58* | *41* | *55* | *62* | *61* | *58* | *61* | *55* | *68* | *64* |
| LAK | *% Total* | | | | | *16* | *18* | *24* | *23* | *40* | *38* | *37* |
| | *% Exclusive* | | | | | *78* | *79* | *69* | *76* | *75* | *78* | *79* |
| Cross conf. | *Total number* | | 11 | 13 | 19 | 24 | 26 | 33 | 29 | 50 | 43 | 60 |
| | *% of Line 1* | | *8* | *10* | *15* | *10* | *10* | *12* | *10* | *19* | *14* | *15* |

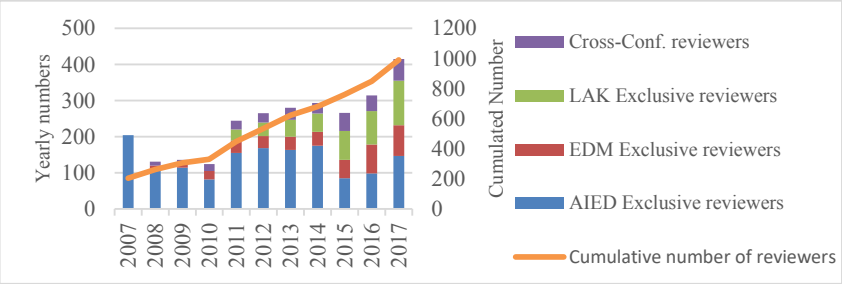


**Figure 1.** Evolution and distribution of the community of the reviewers for the conferences

Figure 1 illustrates the continuous growth of the overall reviewer community from 200 to almost 1000 in a decade, dominated by AIED during the first 8 years. From 2015 onwards, the number of cross-conferences reviewers has been growing too, which raises the question of knowing which communities overlap. But how many of them stayed in their original community and how many have been reviewing for more than a single conference? Overall, the 990 unique reviewers identified have been mentioned a little bit over 3000 times. Despite an average number of 3 conferences reviewed for each reviewer, 71% of them have appeared only in one community (711 nodes with outdegree=1). Figure 2 shows the number of reviewers who have been reviewers outside of their original community.

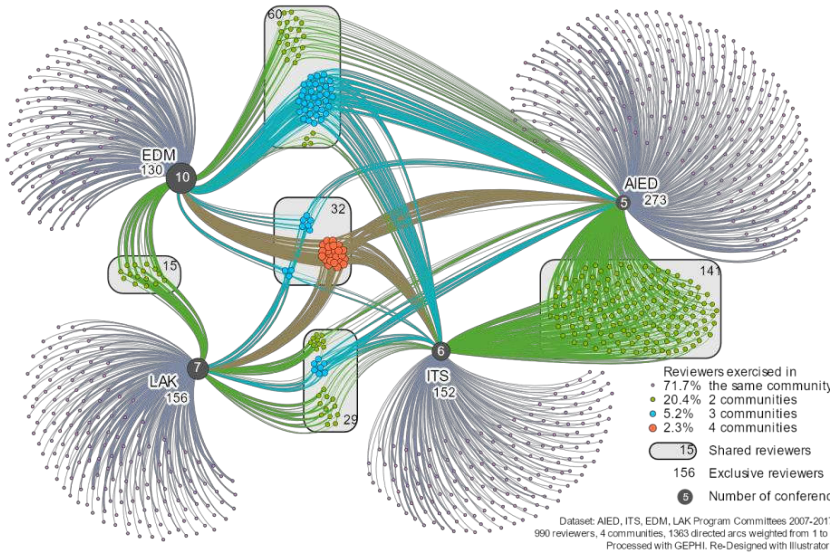


**Figure 2.** The community of reviewers for each conference

AIED/ITS conferences have been sharing a quarter of all their reviewers (141 out of 425): it could come from the fact that those conferences have been alternating over the period considered (odd years for AIED and even years for ITS) – although we see that both of them also have their own subset of reviewers. But beyond this particular case, the number of persons who really belong to two or more communities remains limited: only 13.7% of the reviewers (136 individuals) cross-reviewed between, at least, two of the following communities: AIED/ITS (considered as a single one), EDM and LAK. As illustrated by Figure 2, the common core of the three communities consists of 32 reviewers. The most surprising result was to see how the LAK community was the least related to the others, when compared with the bonds between EDM, ITS and AIED. The reviewers common to each pair of community, as well as to the three communities are in Table A in Appendix, and in Table 4 for a synthesis.

**Table 4.** Percentage of Shared Reviewers on All Reviewers for each pair of conferences

| AIED-ITS | EDM-ITS | EDM-AIED | LAK-ITS | EDM-LAK | LAK-AIED |
|---|---|---|---|---|---|
| 25 | 20 | 18 | 14 | 14 | 10 |

## 4. Journal authors community analysis

Using the second dataset, we considered the papers published in the communities' respective journals (IJAIED, JEDM and JLA). From 2007 to 2017, there are 996 signatures corresponding to 748 unique authors of 349 articles. 80% of these unique authors signed 1 paper; 14% signed 2, and 6% signed at least 3 of them. Overall, the low number of authors of more than one paper limits this analysis, but we performed the same cross-reference analysis as in the previous section for reviewers. It reveals that a dozen of authors published in each pair of journals (*cf.* Table B in appendix), and 8 central authors published in the three of them.

## 5. Textual analysis of journal abstracts

Scientific communities are centered around the scientists that are part of them, but also around some common themes. To identify the themes that are characteristics of each community, we have tried to identify the keywords characteristics of the papers published in the journal of each community, using the third dataset.

First, we performed a cleaning of the abstracts using Python Natural Language Toolkit (NLTK) to perform the usual first step (tokenization, lemmatization and stop words removal). Then we used the word_cloud package to identify visually if some keywords were appearing more in some abstracts than others (*cf.* Figure 3). All communities are obviously very centered on "student", "learning" and "usages". The LA and EDM communities also share the focus on data, which is missing from the AIED community.

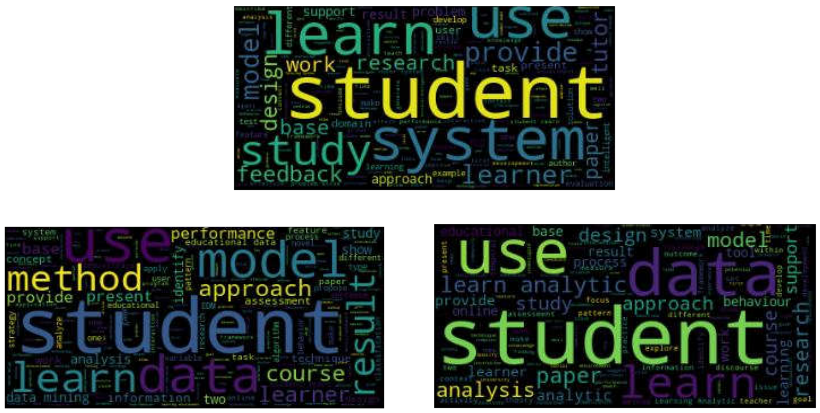


**Figure 3.** Word clouds for IJAIED (top), JEDM (left) and JLA (right) abstracts

However, more than the similarities between the communities, we are interested in what distinguish them from one another. To identify the keywords representative from each community, we extracted from the compilation of the abstracts of each journal the associated keywords using the Rapid Automatic Keywords Extraction (RAKE) algorithm. To avoid the fact that it may overrepresent keywords cited many times by the same article, we kept only the keywords that appeared in at least 20% of the abstracts from each journal. We obtained a set of 110 keywords appearing in at least 29 abstracts from IJAIED, 79 keywords appearing in at least 10 abstracts from JEDM, and 80 keywords appearing in at least 26 abstracts from JLA. Then we extracted (a) the keywords from JEDM not appearing in JLA, (b) the keywords from JLA not appearing in JEDM, (c) the keywords from IJAIED not appearing in JLA nor JEDM. They are summarized in Table 5. Overall, we see that the EDM community remains very anchored in a discovery approach (investigate, evidence, assess, understand, experiment…) when the LA community is more in the practice (support, inform, development, act, teach…). Although the particular techniques used in the papers do not appear with this analysis, the focus of EDM community on a more mathematical approach (features, log, class…) is visible, when compared to LA which focuses on "text", "chi square" and "ratings". As for the AIED community, its roots in tutor systems to provide feedback while modeling skills and knowledge from the student is also clearly visible.

**Table 5.** Keywords specific to each community based on abstracts

| Journals | Keywords |
|---|---|
| JEDM but not JLA | large, propose, technique, behavior, group, compare, ability, educational data mining, improve, demonstrate, ask, investigate, evidence, problem, make, assessment, new, cover, concept, information, analyze, log, discover, apply, assess, finding, feature, class, relate, understand, collect, experiment, task, search, state, type |

| | |
|---|---|
| JLA but not JEDM | support, focus, inform, call, analytics, development, learn analytics, high, time, n, explore, chi, rater, ever, learning, age, tool, LA, go, use, act, put, analytic, text, teach, different, pre, end, lea, two, pose, relation |
| IJAIED only | skill, tutor, instruct, evaluation, domain, interaction, interact, era, test, line, train, know, add, view, ten, well, AI, way, feed, effective, p, prove, low, computer, ratio, art, mode, solve, evaluate, tutor system, feedback, e tutor, effect, q, knowledge, par, help, stem, late, differ, port, adapt, instruction, come |

## 6. Conclusion

Through an analysis of the social networks of the conference reviewers and journal authors from the AIED, EDM and LA community, we have shown that Siemens and Baker's call has been heard, as more and more scientists are at the frontiers between the communities with 139 shared reviewers and 48 shared authors. The research themes however remain clearly distinct, as shown by the keywords analysis of the journal abstracts, with an emphasis on agents and tutors for AIED, automation and prediction for EDM, and visualization for LA. However, these are the different pieces of the same puzzle: enhancing learning experience through technology.

This work presents some limits: we focused on 3 important communities, but which do not represent the whole field of educational technology – extending this approach to other communities such as the "user modeling" one, or more local communities (EC-TEL in Europe) would provide a larger overview of the domain. We could also include conference authors and abstracts in our analysis, to see if more diversity of themes can be identified that way. The lack of information regarding authors' faculties for reviewers as well as for many authors did not allow us to confirm the fact that LAK is closer to education than the other communities. Finally, we have not considered the temporal aspects of the network evolution over the decade, but only the final outcome. Nonetheless, we hope that this work will contribute in structuring the communities, and encourage more scientists to follow the trend towards more interactions between them.

## References

1. Gauthier, G., Frasson, C., VanLehn, K. eds: Intelligent Tutoring Systems: 5th International Conference, ITS 2000, Montreal, Canada, June 19-23, 2000 Proceedings. Springer-Verlag, Berlin Heidelberg (2000).
2. Koedinger, K.R., Corbett, A.T.: Cognitive tutors : technology bringing learning science to the classroom. In: Sawyer, R.K. (ed.) The Cambridge Handbook of the Learning Sciences. pp. 61–77. Cambridge University Press (2006).
3. Romero, C., Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert Syst. Appl. 33, 135–146 (2007).
4. Siemens, G., Baker, R.S.J. d.: Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge. pp. 252–254. ACM, New York, NY, USA (2012).

5. Baker, R.S.J. d., Siemens, G.: Educational Data Mining and Learning Analytics. In: Sawyer, R.K. (ed.) Cambridge Handbook of the Learning Sciences. pp. 253–274. Cambridge University Press, New York, NY (2014).
6. Balacheff, N., Lund, K.: Multidisciplinarity vs. Multivocality, the Case of "Learning Analytics." In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 5–13. ACM, New York, NY, USA (2013).
7. Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J. d. eds: Handbook of educational data mining. Taylor & Francis Group, Boca Raton (2011).
8. Sawyer, R.K. ed: The Cambridge handbook of the learning sciences. Cambridge University Press, New York, NY (2014).
9. Gasevic, D., Dawson, S., Mirriahi, N., Long, P.D.: Learning Analytics – A Growing Field and Community Engagement. J. Learn. Anal. 2, 1–6 (2015).

# Appendix

**Table A.** Name of reviewers for more than one conference in 2007-2017

| Communities | Reviewers |
|---|---|
| AIED – EDM: 60 shared | Agnihotri L., Aïmeur E., Aleven V., Arroyo I., Barnes T., Beck J., Biswas G., Bosch N., Boticario J. G., Champaign J., Chi M., Conati C., Cox R., Crossley S., D'Mello S., Dragon T., Dufresne A., Feng M., Forbes-Riley K., Fossati D., Goldin I., González-Brenes J., Grafsgaard J. F., Heiner C., Hicks A., Hsiao S. I-H., Hutt S., Isotani S., Keshtkar F., Kim J., Koedinger K. R., Lallé S., Larranaga M., Litman D., Liu R., Lynch C., MacLellan C., Martin B., Matsuda N., Mavrikis M., Mojarad S., Mostafavi B., Mostow J., Muldner K., Olney A., Pavlik P., Porayska-Pomsta K., Rau M. A., Ritter S., Rodrigo Ma. M. T., Rus V., San Pedro M. O. Z., Santos O. C., Shaw E., Stewart A., Wang Y., Weibelzahl S., Williams J. J., Zapata-Rivera D. |
| AIED – LAK: 29 shared | Allen L. K., Brooks C., Brusilovsky P., Carmichael T., Daniel B., Dascalu M., Dessus P., Dillenbourg P., Dimitrova V., Fujita N., Greer J., Hatala M., Henze N., Herder E., Hoppe H. U., Kirschner P., Lindstaedt S., Maillet K., Martinez-Maldonado R., Ogata H., Reffay C., Roll I., Sampson D., Schmidt A., Sergis S., Suthers D., Teplovs C., Zervas P., Zouaq A. |
| EDM – LAK: 15 shared | Alexandron G., Conde M. A., Drachsler H., Gobert J., Klamma R., Lang C., Merceron A., Monroy C., Pardo A., Pechenizkiy M., Romero C., Siemens G., Verbert K., Wolpers M., Worsley M. |
| AIED – EDM – LAK: 32 shared | Azevedo R., Baker R.S.J.D, Blink M., Bouchet F., Boyer K. E., Desmarais M., Eagle M., Fancsali S., Gasevic D., Graesser A. C., Heffernan N. T., Jovanovic J., Kay J., Lester J., Luengo V., Mazza R., McCalla G., McLaren B. M., Mitrovic T., Nkambou R., Paquette L., Pardos Z., Pelánek R., Pinkwart N., Reimann P., Penstein-Rosé C., Sahebi S., Snow E. L., Stamper J., Trausan-Matu S., Yacef K., Yudelson M. |

**Table B.** Name of authors for more than one journal in 2007-2017

| Journals | Authors |
|---|---|
| IJAIED & JEDM: 15 shared | Azevedo R., Boyer K. E., Chung G. K.W.K., Conati C., D'Mello S., Goldin I., Harley J. M., Koedinger K. R., Lester J., Luckin R., Miller L. D., Nugent G., Person N., Samal A., Soh L.-K. |
| IJAIED & JLA: 14 shared | Blair K. P., Chin D. B., Cutumisu M., Gowda S. M, Heffernan N. T, Hoppe H. U., Kay J., Linn M. C., Paquette L., Pardos Z., Rau M. A., San Pedro M. O. Z., Schwartz D. L., Segedy J. R. |
| JEDM & JLA: 11 shared | Bannert M., Blikstein P., Cai Z., Crossley S., Kinnebrew J. S., Kitto K., Recker M., Schneider B., Sonnenberg C., Winne P. H., Yacef K. |
| All: 8 shared | Allen L. K, Baker R.S.J.D, Biswas G., Graesser A. C., McNamara D. S., Pelánek R., Penstein-Rosé C., Snow E. L |