Learning Analytics and Privacy-Respecting Privacy in Digital Learning Scenarios

Marvin Priedigkeit, Andreas Weich, Ina Schiering

Learning Analytics and Privacy - Respecting Privacy in Digital Learning Scenarios

Marvin Priedigkeit[1 [0000−0001−9739−8155]], Andreas Weich[1 [0000−0003−4368−5129]], and Ina Schiering[2 [0000−0002−7864−5437]]

[1] Georg Eckert Institute for International Textbook Research Member of the Leibniz Association, Braunschweig, Germany
{marvin.priedigkeit, andreas.weich}@gei.de
[2] Ostfalia University of Applied Sciences, Wolfenbttel, Germany
i.schiering@ostfalia.de

Abstract. With the rise of digital systems in learning scenarios in recent years as learning management systems, massive open online courses, serious games, and the use of sensors and IoT devices huge amounts of personal data are generated. In the context of learning analytics, this data is used to individualize contents and exercises, predict success or dropout. Based on a meta analysis it is investigated to which extent the privacy of learners is respected. Our research found that, although surveys have shown that privacy is a concern for learners and critical to adopt to establish trust in learning analytic solutions, privacy issues are very rarely addressed in actual learning analytic setups.

Keywords: learning analytics · privacy · educational datamining · MOOC, · LMS · serious games

1 Introduction

Based on the rise of digital systems in learning scenarios, the analysis of learning results and meta-data, also called learning analytics, is an area of research gaining importance. The society for Learning Analytics defines learning analytics as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs"[1].

An overview about the current state of the learning analytic research field was given by Papamitsiou et al. [22]. According to them, Massive Open Online Courses (MOOCs) are leading the field of learning analysis, in addition traditional learning management systems (LMS) [21], serious games [19] and sensor-based IoT devices as e.g. stylus [28] are investigated. In this context it is stated that "At the same time, the discussion about ethics and privacy in learning analytics is not new, and is also well-grounded in theoretical frameworks, however, those frameworks are still external to the core of the field itself." [22].

In this paper, a meta-analysis of research in the area of learning analytics is presented. There the consideration of privacy is investigated in the context of

usage groups, technologies, and intended users. To this aim user studies, technology proposals, discussion papers, and case studies are investigated. Based on this analysis a privacy risk classification is proposed.

2 Background

A huge motivation for using learning analytics is to predict the success rate of students and to identify students at risk. To this aim besides data of MOOCs and LMS such as log-in time, log-in frequency, or video pause behavior also historical performances and demographics [3, 24] are used. Moreover, even biometric sensor data were used to analyze students learning performance [28]. In addition to MOOCs, digital serious games are commonly used to analyze students' behavior when confronted with new tasks. For example, Kaeser and Schwartz let students predict the outcome of a digital tug of war game and trained a neural network to group them into clusters according to their learning behavior [19]. The prediction of drop out rates is not only limited to education in schools or universities. Companies, such as Udacity, had also developed methods to analyze enrolled participants at risk. For example, Udacity's researchers Kim et al. developed a deep neural network to identify enrolled participants at risk of failing to complete a course [17].

Learning analytics is already used since several years especially based on existing LMS data. Arnold and Pistilli described a learning analytic solution to predict student's success based on interactions within the Purdue university's learning management system and other sources such as demographic data and historical performance[3]. They used a proprietary algorithm to compute a signal light feedback for the students and focused more on the behavior change of the students at risk. In contrast, Pelaez et al. [24] used machine learning to identify subgroups of students at risk enrolled in the Introductory Psychology course at the San Diego state university. Their goal was to enable the administration to further support this specific sub-groups.

3 Related Work

An overview about the current state of the learning analytic research field was given by Papamitsiou et al. [22]. They analyzed 627 publications, which were published between 2011 and 2019 at the Learning Analytic and Knowledge Conference and the Journal of Learning Analytics, extracted key-terms from these publications manually as well as by statistic analysis, grouped them into 14 clusters, each representing a research theme or sub-field, and sorted these into 4 categories. According to this analysis, MOOCs are leading the field of learning analysis, while "At the same time, the discussion about ethics and privacy in learning analytics is not new, and is also well-grounded in theoretical frameworks, however, those frameworks are still external to the core of the field itself." [22].

In addition to the Learning Analytics and Knowledge Conference, the Educational Data Mining conference is an popular venue for data driven education. Chen et al. analyzed and compared both of them [7]. Beside different other properties, they extracted the eleven most common topics on both conferences. The "Predictive and Descriptive Analytics" topic is at the top of their list, while privacy is none of the top eleven topics.

Privacy issues in learning analytics research are for example incorporated in interviews and questionnaires for students and other stakeholders. For example, Whitelock-Wainwright et al. asked in a survey more than 2000 students in three countries about their expectations of learning analytics, including their privacy concerns [29]. According to them: "From a student perspective, we can see that, on average, they have strong ideal expectations towards the university ensuring all data remain secure or controlling the access from third party companies. However, responses to the predicted expectation scale show students' beliefs to not be as". Tsai et al. identified, that an ethics and privacy framework, including anonymity, is key to ensure that students feel comfortable while sharing their data for learning analytics purposes [27]. Approaches to give institutions support to introduce learning analytic solutions had been made as well [8, 14].

Moreover, Drachsler et al. proposed a checklist for trusted learning analytics in order to help institutions to implement learning analytics with privacy concerns in mind [10]. They conclude that ethics and privacy concerns should be on the same level as functional requirements. Others such as Siemens and Pardo analysed similar privacy issues [23]. However, reports of common practices in higher education show that privacy and ethics are rarely considered sufficiently in concrete learning scenarios [12].

4 Methodology

To investigate the consideration of privacy in the field of learning analytics a broad variety of publications as user studies, technology proposals, discussion papers, case studies are taken into account. We identified relevant literature mainly by two different approaches. First by searching based on corresponding search terms such as "learning analytic privacy study", "empirical learning analytic" on scientific search engines and databases, such as Google scholar and IEEE Xplore. The second approach was to consider proceedings from relevant conferences, i.e. the LAK conference and the EDM conference. In addition, papers identified by these two approaches were used as a starting point for further research via considering references.
For publications, selected in this first phase, the following criteria were evaluated. To this aim, we define the terms participants and audience. *Participants* are persons, which participate in a study and whose data are collected, while *audience* are people, which are addressed as the intended audience by the aim of the study. In the following, the criteria of the intended classification are explained.

Participant type

This criterion describes which type of persons participates mainly in the

study. For example typical types of participants are k12 pupil (learners from primary to secondary education) or students enrolled in courses at a university.

Number of participants
This criterion describes how many participants were investigated in the study. In the context of studies participants could be people, participating in a lab study, students enrolled in a online course, which will be analysed by the study or someone answering questions in a survey.

Target audiance
As described above we distinguish the terms participant and target audience. This could be the same type as the participants or a different group. A study which investigates how MOOC users interact with the system in order to develop a traffic light warning system for future MOOC users, the target audience and the participant types match, both would be MOOC user. Though, the individual participants, whose data are used to develop such a system, might not benefit from this system as they probably already finished their MOOC. If in a similar study a system collects data from a class of k12 pupil and warns the corresponding teacher, the target audience would be teachers, while the type of participant would be pupil.

Feedback type
This describes if a study contains a feedback system and how this feedback is provided, for example a traffic light system to warn MOOC users, that they are at risk to fail the course. The feedback system is always based on the participants data.

Context
This criterion describes in which context the participants contribute to a study. In the case of students using their university LMS to solve home work excises, which data is collected and used in context of a study, the context would be the LMS. It is not necessarily the case, that participants are aware, that they or their data contribute to a study, i.e. a learning analytics scenario. A typical example, when participants are not aware that their data is used in the context of a study are demographic data of enrolled students. We chose to integrate this criterion as transparency is an important privacy requirement and privacy concerns of participants might be influenced if they know that they contribute to a study or not.

Data source type
This criterion describes which type of data is used in the study. Typical data sources are LMS interactions such as login frequency, video playback behaviour or the output data of various biosensors attached to participants in a lab study. Another example of data source types are answers to a survey

or demographic data.

Processing
This describes which methods are used to evaluate and analyse the collected data. A typical processing method is frequency analysis for surveys or machine learning methods such as deep neural networks.

Aim
The aim criteria describes, what the study should achieve for the target audience. For example giving the administration at a university guidance how to implement learning analytics into their institution.

Privacy
This criterion describes if and to what extent privacy is considered in the investigated research. A survey with questionnaires regarding privacy concerns or security perceptions of participants is an example, which should be identified by this criterion. Another typical example is if the study design incorporated privacy concerns upfront or if they evaluate privacy implications only afterwards based on the developed system.

In the second stage, these criteria were used to analyze the selected publication from our first stage and identify relevant publications. As we are interested in studies which actually interact with participants or are at least are addressed to a specific audience like the educational administration, publications which only presented technical solution, but had not tested them with participants were excluded. Additionally, publications that do not provide information about their target audience or aim, were excluded. As learning analytic is an emerging and adapting topic, publications published before 2012 were not considered. The extant publications form our final pool of publications. Based on these, we used the criteria to identify clusters within our pool of publications. Our findings will be described in the following section.

5 Analysis

Based on the proposed criteria in the methodology section we identified 23 research publications, which will be analyzed in this section. First, a brief overview of the publications is given grouped by different contexts. Afterward, other criteria will be analyzed. Finally, relations between publications will be described and clusters of publications will be identified and analysed.

5.1 Overview of Search Results

To gain a better understanding of the search results, for every context example publications are in the following summarized.

In the *biosensor context*, different approaches are investigated to attach existing commercial level sensors to learners measuring their body's signals and

Index	Publication
1	Ebner, Martin, and Matthias Pronegg. "Use of Learning Analytics Applications in Mathematics with Elementary Learners." *International Journal of Academic Research in Education* 1.2 (2015): 26-39.
2	Wampfler, Rafael, et al. "Affective State Prediction in a Mobile Setting using Wearable Biometric Sensors and Stylus." *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*. 2019.
3	Kim, Byung-Hak, Ethan Vizitei, and Varun Ganapathi. "GritNet: Student performance prediction with deep learning." *arXiv preprint arXiv:1804.07405* (2018).
4	Sclater, Niall. "Developing a code of practice for learning analytics." *Journal of Learning Analytics* 3.1 (2016): 16-42.
5	Whitelock-Wainwright, Alexander, et al. "Assessing the validity of a learning analytics expectation instrument: A multinational study." *Journal of Computer Assisted Learning* 36.2 (2020): 209-240.
6	Pelaez, Kevin. "Using a Latent Class Forest to Identify At-Risk Students in Higher Education." *Journal of Educational Data Mining* 11.1 (2019): 18-46.
7	Tsai, Yi-Shan, Alexander Whitelock-Wainwright, and Dragan Gašević. "The privacy paradox and its implications for learning analytics." *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 2020.
8	Käser, Tanja, and Daniel L. Schwartz. "Exploring Neural Network Models for the Classification of Students in Highly Interactive Environments." *International Educational Data Mining Society* (2019).
9	Karumbaiah, Shamya, Ryan S. Baker, and Valerie Shute. "Predicting Quitting in Students Playing a Learning Game." *International Educational Data Mining Society* (2018).
10	Drachsler, Hendrik, and Wolfgang Greller. "Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics." *Proceedings of the sixth international conference on learning analytics & knowledge*. 2016.
11	Hunt-Isaak, Noah, et al. "Using online textbook and in-class poll data to predict in-class performance."
12	Pirkl, Gerald, et al. "Any problems? a wearable sensor-based platform for representational learning-analytics." Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. 2016.
13	Hernández-García, Ángel, et al. "Applying social learning analytics to message boards in online distance learning: A case study." *Computers in Human Behavior* 47 (2015): 68-80.
14	Hernández-Lara, Ana Beatriz, Alexandre Perera-Lluna, and Enric Serradell-López. "Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn." *Computers in Human Behavior* 92 (2019): 600-612.
15	Blikstein, Paulo, et al. "Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming." *Journal of the Learning Sciences* 23.4 (2014): 561-599.
16	Yu, Taeho, and Il-Hyun Jo. "Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education." *Proceedings of the fourth international conference on learning analytics and knowledge*. 2014..
17	Khalil, Hanan, and Martin Ebner. ""How satisfied are you with your MOOC?"-A Research Study on Interaction in Huge Online Courses." *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 2013.
18	Fianu, Eli, et al. "Factors affecting MOOC usage by students in selected Ghanaian universities." *Education Sciences* 8.2 (2018): 70.
19	Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." *arXiv preprint arXiv:1201.3417* (2012).
20	Merceron, Agathe, and Kalina Yacef. "Educational Data Mining: a Case Study." *AIED*. 2005.
21	Pigeau, Antoine, Olivier Aubert, and Yannick Prié. "Success Prediction in MOOCs: A Case Study." *International Educational Data Mining Society* (2019).
22	Whitehill, Jacob, et al. "Beyond prediction: First steps toward automatic intervention in MOOC student stopout." *Available at SSRN 2611750* (2015).
23	Mcbroom, Jessica, Irena Koprinska, and Kalina Yacef. "How does student behaviour change approaching dropout? A study of gender and school year differences." *the 13th International Conference on Educational Data Mining (Upcoming)*.

Table 1. Pool of publications

determining how they feel, while participating in paper equivalent exercises [25]. The identified publications in this context exclusively use students as participant type but could be applied to any kind of learner as well. As every participant needs to be attached to these sensors, the amount of participants is relatively low, compared to other contexts.

Publications from a *serious game* context let participants play specially designed games and analyze their behavior doing so. In our pool of publications, different kinds of serious games are considered. Some are designed like video games where participants need to solve multiple levels to succeed, others are designed to simulate business interactions, where a group of participants needs to decide how a fictive company should adjust its strategy [16, 15]. Typically, serious games designed like video games had k12 pupils as the type of participants while serious games designed as simulations focused on students as participants.

In addition to that, we identified publications with the context *guides and surveys* which aim to give educational administration support to implement and understand learning analytic solutions. While surveys have students as participants, guides do not have participants [26, 27]. Hence for this type of publication, not all criteria could be applied.

LMSs is the most common context in our pool of publications. Publications in this context use different existing data sources such as grade books or demographic data to further understand the learning process of the learners, predict their performance, or determine groups at risk [24]. They almost exclusively have students as participants and tend to target teachers as an audience.

Finally, we identified the context of *MOOCs* which is the second commonest context in our pool. Publications in this context revolve around different types of participants, which participate in full virtual courses. Their behavior such as the time they read a text or answer a quiz is analyzed [17]. These publications focus on adult users, either students or general participants. Moreover, they are the only ones that do target general participants at all. Besides students, these general participants take part in courses e.g. for general training purposes.

5.2 Analysis of Criteria

The general findings for each criterion from the methodology section summarized as a starting point of the analyses followed by a thorough investigation of relations and correlations between different publications.

Participant type

Publications usually had exactly one type of participant. We choose to term k12 pupil as pupil, students enrolled in universities as students, and participants in a MOOCs, which are not exclusively for students enrolled at a university or pupil, as general participants. The most common type of participant is students. This is in line with our expectations, as students are usually of legal age (not minors), therefore it is easier to integrate them into studies. Moreover, compared to regular MOOC users, universities already collect different demographic data by enrollment. This forms a solid

research basis of data and legal factors.

Number of paricipants
The number of participants varies widely, from 6 participants in a biosensor based lab study to over 10000 participants in a MOOC based study. However, even within the same type of context, the number of participants varies widely. For example, in the context of LMS, the number of participants varies from about 100 to over 3000 participants.

Target audiance
We observed, that the type of participants and the target audience differ in most publications. Moreover, our research has identified, that typically different types of teachers are targeted, for example, teachers for k12 pupils and teachers at universities. As all the different types of teachers mainly have the task to help learners achieving their individual learning targets, we choose to summarize K12, university, and MOOC teachers as teachers in general. Other target audiences that occur, are pupils, students, and general users, which are matching the participant types. Additionally, two other target audiences occurred, education administration and research.

Feedback type
Our research has found, that the usual type of feedback is either to inform the according type of participant or the according teacher or to deliver no direct feedback at all. It should be noted, that most publications only reported their findings, but did not take the extra step to develop a feedback system or inform at least the participants.

Context
Different contexts could be identified by our research. We summarized interaction within any kind of LMS such as forum posts, day, and time of exercise download or submission behaviour as LMS context. Different kinds of biosensors, such as eye trackers, heartbeat detectors, or stylus pressure, are also summarized as biosensors. In addition to them MOOCs, surveys, and serious games also occurred as the context of publications.

Data source type
Different data sources had been found. We observed an obvious correlation between the data source type and the context of the study. Studies within the context MOOC, mostly had different kinds of MOOC related data sources such as clickstreams, while publications within the context of serious games had very game-specific data sources. Although plenty of data sources, including sensitive data such as past grades and biosignals, are used in the pool of publications, data minimization approaches could not be identified.

Prosessing
Our research has identified different approaches concerning machine learn-

ing, such as neural networks and latent class trees [19, 17, 24]. However, we termed every machine learning approach and even linear regression as machine learning. Besides machine learning, the most common processing method were different statistic methods such as frequency analysis for the evaluation of surveys, which we termed as statistical analysis.

Aim

The aim of the publications varies as well. Some papers aim to understand participant expectations for learning analytic systems, others want to understand how participants learn or predict various aspects of their learning outcome, such as grades or their risk to fail a certain course. Though some publications aim to understand learning, this aim is limited to understanding certain easy to measure factors of this complex system.

Privacy

We observed, that privacy is rarely addressed in our pool of publications. Some surveys tackle privacy concerns in order to give guidance on how to implement learning analytics or by questionnaires in a survey but the majority do not concern them at all.

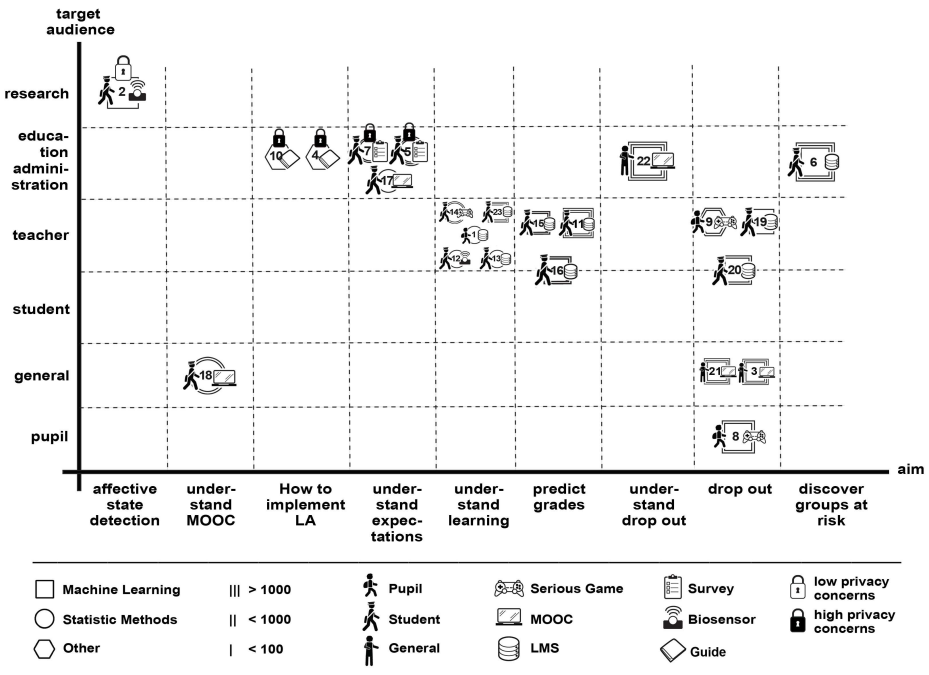

Fig. 1. Visualization of our pool of publications. The numbers refer to the entry in table 1

An overview of our pool of publications and their according criteria is visualized in Figure 1. The target audience and the aim are important criteria with multiple values, therefore we choose them as axis values for our visualization. They give a good overview of our pool of publications. We encoded our additional criteria by additional shapes or symbols. Therefore every study has a shape, which determines the processing criteria, whereas the size of the shape is used to visualize the number of participants. Additionally, every study can have three additional symbols. The left symbol determines the participant type, while the right symbol determines the context criterion. Finally, every study can have either a hollow or a solid lock symbol, representing low or respective high privacy concerns. The lock is not present if the study does not concern privacy at all.

Relations between different criteria could be identified and will be described in the following. We observed first of all typical relations between the type of participants and the target audience. In general, the type of participant and the target audience of a study are different. Although the main participant type of publications is students, they are almost never the target audience. This is true across all aims. To a lesser degree, the same is true for pupil, if they are the type of participants, mostly the teacher is the according target audience for them. However, in one of three publications, they are also their own target audience and therefore this is much more common compared to the student participant group. In contrast to that, if general users are participants in a study, they usually are the target audience as well.
Overall, teachers and the education administration are the most common target audience, while the combined group of actual learners, in the different contexts, are a less common target group. Therefore we identified an imbalance between the group of learners as participants, which contribute with their data, and the according target group. In most publications, the group of learners does not receive any feedback nor are they the target group.

Also, a strong relation between the aim criterion and the target audience criterion was observed. Many aims are exclusive for a certain target audience, such as *How to implement LA* and *understand expectations*. In contrast, we observed, that aims that are involved in the prediction of learners' performance such as *predict grades* and *drop out* are spread through multiple target audiences. Another observation is, that the education administrative target group is the group with the most differentiated aims.

The processing method is spread through different aims. However, we observed, a relation between machine learning-based processing and aims, that resolve around drop out of students. Almost any study that aims to understand drop out, discover groups at risk or determine drop out uses machine learning approaches to process the collected data. In contrast, aims such as *How to implement LA* and *understand expectations* obviously never uses machine learning processing methods.

Finally, relations between the context and privacy concerns have been observed. Whenever the context is either *guide* or *survey* high privacy concerns

have been considered. In contrast, no privacy concerns have been taken into account in any other aims.

5.3 Proposed Clusters based on Analysis

Beyond relations between different criteria, we were able to identify major cluster of publications, which share multiple relations with each other. These clusters will be described in the following.

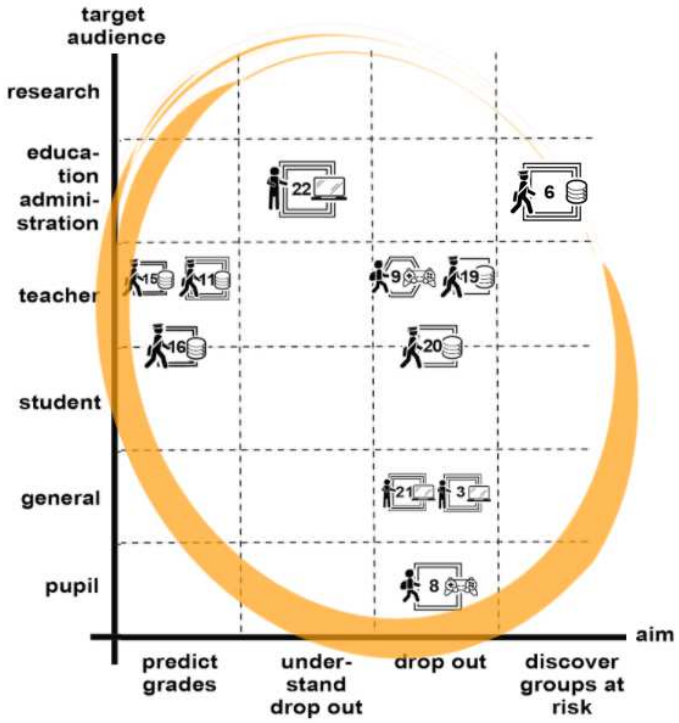

Fig. 2. Cluster of publications using machine learning to determine learners performance

The first major cluster of publications is shown in figure 2. Publications within this group target every type of audience involved in the actual learning process. They aim to predict the learners' performance and using machine learning as processing method. Even though the prediction of a learner's performance might influence the perception of the learners' real performance, especially by his teacher, the actual process of how the machine learning processing method determines this prediction is not transparent to the learners' nor their teachers'.

Even though they use sensitive data such as past grades, none of these publications even address privacy issues at all. The next group is shown in figure 3.

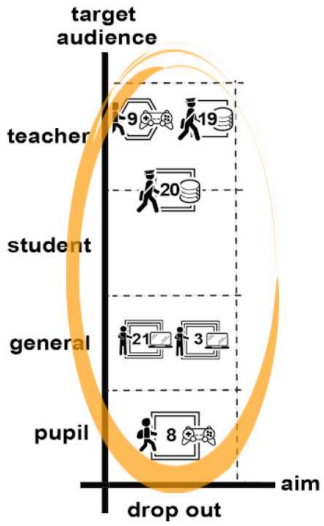

Fig. 3. Cluster of publications with a highly diverse target audience

We identified that drop out is a major aim across all types of target audience involved in the actual learning process. In contrast, most other aims are limited to a single target audience. Moreover, a mix of different contexts appears in this cluster. Though every type of participant appears in this cluster privacy is never concerned.

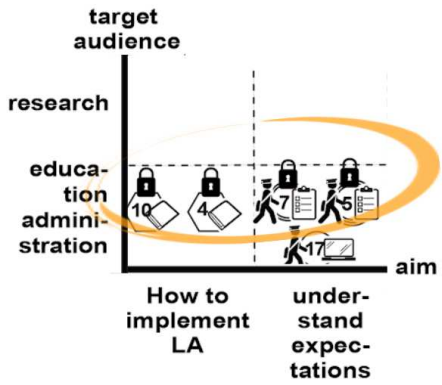

Fig. 4. Cluster of publications with high privacy concerns

Our last clusters contain publications with high privacy considered, as shown in figure 4. We identified that privacy is mainly addressed with surveys and guides, the target audience is education administration. Therefore the group of learners is not targeted whenever privacy is a high concern. Moreover, the contest of privacy concerning publications is almost exclusively limited to guides and surveys.

5.4 Privacy Risks

We observed that privacy is rarely concerned in the pool of publications. Therefore, in this section, we will further describe privacy threats in learning analytic systems and briefly outline measurements for them. To this aim, we will follow the six privacy protection goals to analyse learning analytic systems [13].

Confidentiality

The observed publications describe a broad variety of contexts and therefore plenty of different data is collected in the learning analytic scenarios. In addition to regular personal data such as name, address, and so on, proposed learning analytic systems also collect very sensitive data such as learning performance and grades. Moreover investigated systems also track detailed data from log-in times and patterns to learning behaviours such as the video playback behaviour [17]. This makes the learners transparent, as very individual behaviours such as how the learner engages with the material are tracked. Moreover, contexts such as the biosensor context even track physical feats, for example, heartbeat rate [28]. Therefore the confidentiality of this data is especially important and should be limited to authorized stakeholders such as the according teacher. To achieve this goal different established solutions could be used. From a privacy model and its strict implementation of rules, roles and responsibility to data protection by encryption with well

established cryptographic encryption schemes such as AES and ChaCha20 in combination with digital signature algorithm on elliptic curves like Ed25519 to prevent unauthorized interactions [9, 4, 5]. Although solutions exist the observed publications do not mention if or how confidentiality is achieved.

Integrity

Data from learning analytic systems or in general learning contexts are often intertwined with grades and in the long run with graduation. Therefore it is necessary that such sensitive data is integrous and not altered by unauthorized stakeholders and at least alternations should be documented and logged. This is not only true for grades but also for the fundamental data which leads to that grade, such as engagement with learning material and submissions. Unintended or malicious alternations of learners' data such as grades, or enrolled courses could lead to serious difficulties for example falsely missing requirements for graduation. Similar to confidentiality, integrity could be provided by established cryptographic solutions, but, also similar, our pool of publications does not mention if or how they tackle integrity as a goal for their proposed systems [5].

Availability

As data intertwined with grades and graduation stays relevant for a relatively long time, often the learners' whole lifetime, it should be assured that key data from learning analytic systems are available for the same time. However, in most scenarios, the data availability is not very time crucial and other goals such as confidentiality are more important and therefore should have priority. However, availability could fairly easily be introduced, even into existing solutions, by the redundancy of according data and offline backup solutions.

Transparency

In the observed learning analytic publications, it is not transparent for the learner who has actual access to their data and who might get access in the future. This is especially true in complex relationships, for example, who is actually reading and correcting submissions into a MOOC? The actual teacher which was recorded for the video material, the support staff, or even an algorithm? Therefore it is necessary to inform the learners up front, who has or will become accessible to what kind of data in an understandable way. Moreover, in contrast to our observations, learning analytic systems shall also target the actual learners as the target audience. Privacy notifications can be one way to raise the learners' awareness about their data usage and provide them transparent information, like who accessed which data [20].

Unlinkability

While learners are interested in a transparent understanding of how their data is processed, it is also necessary to hinder others to link different data about the learner. In scenarios that involve real-world interactions between

learners and teachers, a teacher could link individual learners with additional data, such as social media profiles. Therefore, unlinkability is very limited in these scenarios, as it is in non-digital classroom scenarios. In other learning scenarios, for example, MOOC's without real-world interactions, techniques such as data minimisation and pseudonymisation could be used and are proposed in some publications [6]. Nevertheless, the risk of reidentification is always present because of the amount and detail of data [18].
Especially in conjunction with machine learning approaches more sophisticated solutions are needed since usually a huge amount of data from LMS is used. To this aim, privacy-preserving machine learning is important, encompassing methods such as homomorphic encryption and differential privacy [2, 11]. Though we observed different publications that process data with machine learning methods such solutions have not to be mentioned

Intervenability
As the processed data consists of highly sensitive data, which is often stored for a long time, the Intervenability of learners should be taken into account as well. A learner should be able to intervene in every stage of the learning analytic system. From mistakenly clicked answers to the grading of submissions and the conclusions of the learning analytic system. However in order to intervene transparency is necessary. Though, our observed publications often target other stakeholders such as the teacher or education administration and only present them with the conclusions of the system. Therefore learners do not even know the conclusions of learning analytic systems, therefore they are unable to intervene in case of faulty data, such as wrongly calculated grades.

6 Discussion and Conclusion

In general learning, analytics has a huge potential in digital learning scenarios and the use of such approaches is gaining importance. On the other hand students' interaction with such systems provides a huge amount of personal data and at the moment privacy is rarely adequately considered.

Our study revealed that privacy issues are mostly addressed in surveys or guides, but very rarely in actual learning analytic implementations. Moreover, data usage is not transparent.

Trustfulness and privacy are key factors to successfully implemented accepted learning analytic systems [10, 23]. Therefore holistic systems, which already tackle privacy concerns in the preferable participatory, design process, are needed. Even though surveys and guides for the education administration are a first step to design privacy-preserving learning analytic systems, all in the learning process involved parties and their concerns need to be address by such systems to unfold the potential slumbering inside learning analytics. Moreover,

participants should always be considered as, at least secondary, audiences in order to make transparent, what happened to their data and to let them know the results.

Although pseudonymization techniques to ensure privacy is widely used in different scenarios [6], it has been shown, that pseudonymization or even minimal data collection techniques can be overcome, thus leading to reidentification [18]. Therefore more sophisticated approaches from the privacy and security community have to be included into learning analytic systems. Moreover cultural, social, political, and ethical considerations should be included in interdisciplinary projects. This is especially true, as different machine learning approaches are gaining attention within the learning analytic community in recent years. As they mostly operate on private data and pseudonymization is prone to fail, solutions for this dilemma must be found. Privacy-preserving machine learning is, among others, an approach to build learning analytic systems with privacy by design, which could solve this dilemma.

Acknowledgement This work was supported by the Leibniz Association and the Ministry for Science and Culture of Lower Saxony as part of Leibniz ScienceCampus Postdigital Participation Braunschweig.

References

1. What is learning analytics, https://www.solaresearch.org/about/what-is-learning-analytics/
2. Albrecht, M., Chase, M., Chen, H., Ding, J., Goldwasser, S., Gorbunov, S., Halevi, S., Hoffstein, J., Laine, K., Lauter, K., Lokam, S., Micciancio, D., Moody, D., Morrison, T., Sahai, A., Vaikuntanathan, V.: Homomorphic encryption security standard. Tech. rep., HomomorphicEncryption.org, Toronto, Canada (November 2018)
3. Arnold, K.E., Pistilli, M.D.: Course signals at Purdue: using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. pp. 267–270. LAK '12, Association for Computing Machinery, Vancouver, British Columbia, Canada (Apr 2012). https://doi.org/10.1145/2330601.2330666
4. Bernstein, D.J.: Chacha, a variant of salsa20. In: Workshop Record of SASC. vol. 8, pp. 3–5 (2008)
5. Bernstein, D.J., Duif, N., Lange, T., Schwabe, P., Yang, B.Y.: High-speed high-security signatures. Journal of cryptographic engineering **2**(2), 77–89 (2012)
6. Bosch, N., Crues, R.W., Paquette, L., Shaik, N.: hello,[redacted]: Protecting student privacy in analyses of online discussion forums. EDM (2020)
7. Chen, G., Rolim, V., Mello, R.F., Gaevi, D.: Let's shine together! a comparative study between learning analytics and educational data mining. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 544–553. LAK '20, Association for Computing Machinery, Frankfurt, Germany (Mar 2020). https://doi.org/10.1145/3375462.3375500
8. Corrin, L., Kennedy, G., French, S., Shum, S.B., Kitto, K., Pardo, A., West, D., Mirriahi, N., Colvin, C.: The ethics of learning analytics in australian higher

education (2019), https://melbourne-cshe.unimelb.edu.au/research/research-projects/edutech/the-ethical-use-of-learning-analytics
9. Daemen, J., Rijmen, V.: Reijndael: The advanced encryption standard. Dr. Dobb's Journal: Software Tools for the Professional Programmer **26**(3), 137–139 (2001)
10. Drachsler, H., Greller, W.: Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 89–98. LAK '16, Association for Computing Machinery, Edinburgh, United Kingdom (Apr 2016). https://doi.org/10.1145/2883851.2883893
11. Dwork, C.: Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. pp. 1–19. Springer (2008)
12. Flanagan, B., Ogata, H.: Integration of learning analytics research and production systems while protecting privacy. In: The 25th International Conference on Computers in Education, Christchurch, New Zealand. pp. 333–338 (2017)
13. Hansen, M., Jensen, M., Rost, M.: Protection goals for privacy engineering. In: 2015 IEEE Security and Privacy Workshops. pp. 159–166. IEEE (2015)
14. Hermann, O., Hansen, J., Rensing, C., Drachsler, H.: Verhaltenskodex fr Trusted Learning Analytics (Mar 2020). https://doi.org/10.13140/RG.2.2.24859.41760
15. Hernández-Lara, A.B., Perera-Lluna, A., Serradell-López, E.: Applying learning analytics to students interaction in business simulation games. the usefulness of learning analytics to know what students really learn. Computers in Human Behavior **92**, 600–612 (2019)
16. Karumbaiah, S., Baker, R.S.J.d., Shute, V.J.: Predicting Quitting in Students Playing a Learning Game. In: EDM (2018)
17. Kim, B.H., Vizitei, E., Ganapathi, V.: GritNet: Student Performance Prediction with Deep Learning. EDM (2018)
18. Klose, M., Desai, V., Song, Y., Gehringer, E.: Edm and privacy: Ethics and legalities of data collection, usage, and storage. EDM (2020)
19. Kser, T., Schwartz, D.L.: Exploring Neural Network Models for the Classification of Students in Highly Interactive Environments. EDM '19, International Educational Data Mining Society (Jul 2019), https://eric.ed.gov/?id=ED599211
20. Murmann, P., Reinhardt, D., Fischer-Hübner, S.: To be, or not to be notified. In: IFIP International Conference on ICT Systems Security and Privacy Protection. pp. 209–222. Springer (2019)
21. Mwalumbwe, I., Mtebe, J.S.: Using Learning Analytics to Predict Students Performance in Moodle Learning Management System: A Case of Mbeya University of Science and Technology. The Electronic Journal of Information Systems in Developing Countries **79**(1), 1–13 (2017). https://doi.org/10.1002/j.1681-4835.2017.tb00577.x
22. Papamitsiou, Z., Giannakos, M.N., Ochoa, X.: From childhood to maturity: Are we there yet? Mapping the intellectual progress in learning analytics during the past decade. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 559–568. LAK '20, Association for Computing Machinery, Frankfurt, Germany (Mar 2020). https://doi.org/10.1145/3375462.3375519
23. Pardo, A., Siemens, G.: Ethical and privacy principles for learning analytics. British Journal of Educational Technology **45**(3), 438–450 (2014). https://doi.org/10.1111/bjet.12152
24. Pelaez, K., Levine, R., Fan, J., Guarcello, M., Laumakis, M.: Using a Latent Class Forest to Identify At-Risk Students in Higher Education. EDM '19 (2019). https://doi.org/10.5281/zenodo.3554747

25. Pirkl, G., Hevesi, P., Lukowicz, P., Klein, P., Heisel, C., Gröber, S., Kuhn, J., Sick, B.: Any problems? a wearable sensor-based platform for representational learning-analytics. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. pp. 353–356 (2016)
26. Sclater, N.: Developing a Code of Practice for Learning Analytics. Journal of Learning Analytics **3**(1), 16–42 (2016). https://doi.org/10.18608/jla.2016.31.3
27. Tsai, Y.S., Whitelock-Wainwright, A., Gaevi, D.: The privacy paradox and its implications for learning analytics. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 230–239. LAK '20, Association for Computing Machinery, Frankfurt, Germany (Mar 2020). https://doi.org/10.1145/3375462.3375536
28. Wampfler, R., Klingler, S., Solenthaler, B., Schinazi, V., Gross, M.: Affective State Prediction in a Mobile Setting using Wearable Biometric Sensors and Stylus (Jul 2019). https://doi.org/10.3929/ethz-b-000393912
29. Whitelock-Wainwright, A., Gasevic, D., Tsai, Y., Drachsler, H., Scheffel, M., Merino, P., Tammets, K., Delgado-Kloos, C.: Assessing the validity of a learning analytics expectation instrument: A multinational study. Journal of Computer Assisted Learning (Jan 2020). https://doi.org/10.1111/jcal.12401