OEA Learning Analytics

Use Case Defined

January 2023

*­­­­­­­­­­­­­­­­­­*

*This is a preliminary document and may be changed substantially prior to final release of the software described herein. The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions,* ***it should not be interpreted as a commitment on the part of Microsoft****, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.*

***This template is for informational purposes only. Microsoft makes no warranties, express or implied, in this document.***

***Complying with all applicable copyright laws is the responsibility of the user.***

*© 2023 Microsoft Corporation. All rights reserved.*

# 

# 1) The Use Case Problem

**Defining the Problem: What problem does this use case seek to solve?**

**Learning Analytics**: This use case package is designed to support educators understand and support learners in their classes and courses, and academic support staff like tutors. In particular, the use case is designed to provide answers to questions like:

* Are students engaged and learning in their course?
* What's the impact of student well-being on engagement and academic performance?
* What are the best ways to keep students engaged, especially in hybrid or remote learning?
* How are students making progress towards learning outcomes?
* Which learning resources are students engaging with most, and how does engagement with specific learning resources relate to outcomes?
* Which learners most need support to succeed in the course, and what type of support will help them?

--------------------------------------------------------------------------------------------------------------------------

Learning Analytics is a field of research developed over the last 2 decades that examines how data from learning resources, applications, and platforms can inform learning design and other decisions to improve learning outcomes. Much of the prior research has used data from individual applications or platforms such as Learning Management Systems or discrete subject-specific applications (e.g., math or reading). With the greatly expanded use of digital learning tools over the past years (accelerated during the pandemic), teachers and faculty have begun using many different applications in the same class or course.

Building from prior learning analytics research in the context of today’s multi-app learning environments, this Open Education Analytics use case package refers to learning analytics as 1) enabling analytics that combine data from several different learning applications and platforms; 2) provides a data foundation for analyzing combined data to answer questions across the continuum from student well-being, engagement, and outcomes; 3) provides insights from which faculty, tutor, or student nudges or recommendations can be created.

Learning happens through many modalities: space (face to face or remote); time (synchronous vs asynchronous); and scale (individual learning, small group collaboration, large scale groups). While it is unlikely that there is rich data on all these modalities, learning analytics should acknowledge and appreciate that digital signals provide only part of the overall picture of learning. Data from face-to-face interactions or learning work that takes place offline may be limited or not available.

The Open Education Analytics Working Group on Learning Analytics collaborated to create this open source GitHub package on [Open Education Analytics (OEA)](https://github.com/microsoft/OpenEduAnalytics) to enable other education systems to use the technical assets included in the package.

**Important Note:** It is strongly recommended to education systems or institutions planning to use this package establish a process for obtaining consent for using the data included in this package. This consent should be part of the system or institution’s broader data governance policy that clearly specifies who can have access to what data, under what conditions, for what purposes, and for what length of time.

# 2) The Use Case Stakeholders

Considering the benefits of a use case – and its potential harm – requires the consideration of different stakeholders and their points of view. Stakeholders typically include the people who are responsible for, will use, or will be affected by the use case. Stakeholders are defined by their role: their duties, contextual identity, or circumstances in relationship to the use case.

* **Direct stakeholders** interact with the data from a use case directly and make decisions or take actions based on results. They include data users, system developers, and even data system maintenance staff.
* **Indirect stakeholders** are affected by the use case data system but, unlike direct stakeholders, they do not have a role that requires them to use or maintain the system. Indirect stakeholders can include groups who may be affected by the downstream effects of the system, such as parents, students, or future employers.
* **Malicious actors** include hackers and others who may intentionally misuse the system. Considering malicious actors is important to supporting safe and reliable data systems.

**Who are the stakeholder groups for this use case, and how are they involved in its development?**

|  |  |  |
| --- | --- | --- |
| **Possible Stakeholder Groups** | **Relationship to Use Case** | **Involvement in Use Case** |
| Educators (Faculty, Instructors, Curriculum Designers, Learning Engineers, Tutors, or Teachers) | Direct Stakeholders | May use these analytics to understand how they can support and engage their students and improve the learning outcomes for all students in a class or course or for an individual student. |
| Technology, IT, learning analytics, and research departments and staff | Direct Stakeholders | Will provision data and develop the analytic products used by Educators. |
| Learners | Indirect Stakeholders | Potential beneficiaries of academic or well-being support or personalized learning experiences. |
| Parents or Guardians | Indirect Stakeholders | Potential beneficiaries of academic or well-being supports or personalized learning experiences for their learners. |
| School System or Institutional Leaders | Indirect Stakeholders | May be needed to provide funding or be able to use the data outputs to improve learning across the system. |
| Potential Malicious Actors | Indirect Stakeholders | Potential for hacking this data system to prevent its effective use for learning or to harm individual or groups of teachers or students. |

**Engaging stakeholders in the use case development process.**

Including stakeholders in the early thinking and conceptualization of a data use case is a good way to ensure that the use case output will be accepted, trusted, and used by key stakeholders. For example, conducting interviews or focus groups with representatives of each stakeholder group can provide early insights into the conceptual model framing the use case and the appropriateness of specific data sources to be used. At a later stage in the project, involving key stakeholder groups in designing the use case outputs (such as a dashboard or notification) can be essential to that product’s eventual effective use.

**Outline how stakeholders will be involved in the development in different stages of the use case development:**

At an early stage in the use case process the primary stakeholders involved should be learners, educators, and Technology, IT, learning analytics, and research departments and staff. The project will likely be led by the technology team in partnership with educators to define the project scope to ensure educators can and will act on the data. Learners should be involved to ensure they see these uses of the data as acceptable and to define thresholds for privacy and security.

At a later stage in the project, as dashboards and information start to be available, these should be reviewed for validity with different stakeholder groups. For example, students who are identified as potential beneficiaries should be included to plan ways of using the data outputs that they will accept or respond positively to.

**What type of outputs are expected from this use case, such as AI models, dashboards, or notification systems?**

This use case should result in reports and real-time data dashboards that show patterns of engagement, well-being, and academic progress related to the instructional design of a class or course. Outputs may also include predictive models that identify students who would benefit from personalized learning experiences or support that may be delivered through automated or human-triggered recommendations or nudges.

# 3) Mapping Theory to Data

**For this use case, what prior research or conceptual model frames your theory of the problem?**

“LEARNING ANALYTICS is 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, as defined back in 2011 for the first LAK, this general definition still holds true even as the field has grown. Learning analytics is both an academic field and commercial marketplace which have taken rapid shape over the last decade. As a research and teaching field, Learning Analytics sits at the convergence of Learning (e.g., educational research, learning and assessment sciences, educational technology), Analytics (e.g., statistics, visualization, computer/data sciences, artificial intelligence), and Human-Centered Design (e.g. usability, participatory design, sociotechnical systems thinking).” [What is Learning Analytics? - Society for Learning Analytics Research (SoLAR) (solaresearch.org)](https://www.solaresearch.org/about/what-is-learning-analytics/)

* Dragan, G., Dawson, S., Siemens, S. Let’s not forget: Learning Analytics are about Learning. <https://www.sfu.ca/~dgasevic/papers_shared/techtrends2015.pdf>
  + The Gašević et al. (2014) analysis of the results of a regression model created by combining data from nine undergraduate courses in an Australian University showed that **only three variables – number of logins and number of operations performed on discussion forums and resources were significant predictors of academic performance.**
  + Early studies of Course Signals showed high levels of predictive accuracy and significant benefits in the retention of the students who took at least one course adopting the early alert software versus those who took a course without the Course Signals tool (Arnold & Pistilli, 2012).
* [SoLAR\_Position-Paper\_2020\_09.pdf (solaresearch.org)](https://www.solaresearch.org/wp-content/uploads/2020/09/SoLAR_Position-Paper_2020_09.pdf)
  + “EdTech systems are currently generating large amounts of low quality data that, while providing surface level insights, offers few avenues for improving student learning and the environments in which it occurs.”
  + “Information such as instructor details, learning objectives, teaching methodology, learning design etc. are increasingly becoming important in the most sophisticated approaches to LA. And yet this information is rarely stored in a form that can be accessed.”

**Mapping Theory to Data. From prior research or conceptual models what are they key data categories expected to inform this use case? What local data sources are available or needed for each category? Please note where no data is available for a Data Category**

A key part of the use case development process is deciding which data to use and how it should be mapped to the theory of the problem. Identifying which data should be viewed as a “feature” and which data is the “target outcome” is at the core of this mapping.

|  |  |
| --- | --- |
| **Key Data Category** | **Local Data Source** |
| Student enrollment in courses (Class rosters) and student demographics | Student Information Systems (SIS or MIS) provide associations between educators and students enrolled in their courses, as well as demographic information on those students. |
| Digital learning activities of students | For example, data on student engagement from digital platforms and applications used as part of a class or course. |
| Learning resources associated with class or course | For example, the readings, files, videos, experiments, projects, etc. used as part of instruction. |
| Well-being data | For example, social engagement, extra-curricular, health, safety, learning, or financial status or challenges. |
| Assessment data | Can include files, videos, or projects created by students, or data from students’ responses to structured traditional assessments. |
| Skills and career data | For example, information on students’ career goals, technical or employability skills, and skill gaps between goals and current skills. |
| Instructional Design | Information pertaining to the teaching methods or design of a class, such as the structure of learning project or experience. |

Note: [OEA Modules and Packages](https://github.com/microsoft/OpenEduAnalytics), support many of these education use cases through accelerating the ingestion of key data sources needed and providing resources to set up these use cases.

**OEA Modules that can be used for this use case:**

| **Data Category** | **OEA Module** |
| --- | --- |
| Student Information System data | For typical [Student Information System](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Student_and_School_Data_Systems) (SIS) data, including school rosters, grade level and demographic information. |
| Digital learning activities | [Microsoft Education Insights](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Microsoft_Education_Insights) and [Reading Progress](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Reading_Progress)  [Microsoft Graph](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Microsoft_Graph)  [Clever](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Clever) |
| Learning Management Systems (LMS) | [Canvas](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Canvas_Data)  Moodle – to be published later in 2023 |
| Assessment data | [iReady](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/iReady) |

**Note: Mapping theory to data with a ‘data dictionary.’**

A “data dictionary” allows the data team to examine specific data tables and data entities in the available datasets, and then map specific items to the Key Data Category. For example, in the OEA modules described above, data dictionaries are provided for each data source in the “test data” or “data” folders.

New data services like [Azure Purview](https://docs.microsoft.com/en-us/azure/purview/overview) can support this work through creating a holistic, up-to-date map of a data repository with automated data discovery, sensitive data classification, and end-to-end data lineage.

**Please see “Privacy and Security” section below for more information on ensuring that sensitive data is protected.**

4) Responsible AI Principles Applied

**In these next sections, please answer the questions under each of the headings describing how responsible AI principles will be applied to this use case.**

Fairness Principle

|  |
| --- |
| AI systems should treat everyone in a fair and balanced manner and not affect similarly situated groups of people in different ways. Human decision makers are susceptible to many forms of prejudice and bias, such as those rooted in gender and racial stereotypes. To ensure AI models are trained in a way that does not embed or re-enforce those biases, models must be tested for fairness. Microsoft has developed an open-source toolkit to support this called [Fairlearn](https://fairlearn.org/), which can be applied within the Azure analytical services used in the OEA reference architecture.  [Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6) on Fairness Principle.  **Who is most likely to be at risk of experiencing harm from this use case?**  This Learning Analytics Use Case does not currently involve the development of machine learning or a predictive model, so AI model bias is not applicable. However, data that describes different groups of students or contexts should be used to understand patterns that may reflect bias. |
| |  |  |  | | --- | --- | --- | | **Group or Subpopulation** | **Clearly Labelled in Dataset? Y/N** | **Planned Mitigations** | | 1. For example, immigrants or rural students |  | Provide dashboard data that shows if different groups of students have different levels of well-being, engagement, or outcomes. | | 1. For example, socio-economically disadvantaged, gender, language-based groups, racial, ethnic, or religious groups |  | Provide dashboard data that shows if different groups of students have different levels of well-being, engagement, or outcomes. | |

Reliability and Safety Principle

Systems should operate reliably and safely when they function in the world. AI systems must be designed with a view to the potential benefits and risks to different stakeholders and undergo rigorous testing to ensure they respond safely to unanticipated situations and do not evolve in ways that are inconsistent with the original shared purpose.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6) on Reliability and Safety Principle.

**What are possible risks faced by learners or educators from the analytics of this use case?**

1. Risk 1: students receive nudges, recommendations, or personalized experiences based on analytics that are not appropriate to their context.
2. Risk 2: students **do not** receive nudges, recommendations, or personalized experiences based on analytics that are needed or appropriate to their context.
3. Risk 3: teachers, faculty, or teachers use data and analytics to shape decisions and actions taken with whole class or with individual students, when some aspect of that data is incorrect.

**Planned Mitigations:**

The Learning Analytics solution developed for this use case needs to be made as robust and accurate as possible, and repeated, continuous testing and human checks of the system should be conducted both in initial design stages as well as during production of the system. For example, faculty or tutors should confirm directly with learners whether the data on which their thinking or actions is based. Additionally, technical and data teams must ensure all data sources are accurate, clean, and that no student is missed in the analysis.

Transparency Principle

Transparency requires visibility into all levels of decision-making and design of an AI system. Designers should clearly document their goals, definitions, and design choices, and any assumptions they have made. Those who build and use AI systems should be forthcoming about when, why, and how they choose to build and deploy them, as well as their data and systems' limitations. Information should be readily available on the quality of the predictions and recommendations the AI system makes. Transparency also encompasses intelligibility, which means that people (in this case, educators, tutors, students, etc.) should be able to understand, monitor, and respond to the technical behavior or recommendations of AI systems.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Transparency Principle

**What steps will the analytics or AI process include?**

This Learning Analytics Use case does not involve the development of machine learning or a predictive model, so AI transparency on model limitations is not applicable currently.

**Who will develop the analytics or models?**

[Must be completed by the team from the educational organization implementing this use case.]

**How will the limitations of the analytics or AI model be communicated to stakeholders and users?**

Limitations on the quality of the data sources should be noted as information within the dashboards or outputs themselves. For example, if a complete inventory of learning resources is not available, or if data on face to face interactions (in person classes) is not available.

**What means will be built into the system for correction and model feedback by those who provide data and who use its outputs?**

[Must be completed by the team from the educational organization implementing this use case.]

Privacy and Security

Private or personal data should not be collected or incorporated in analytics or AI products for education unless all groups have agreed this data is necessary to achieve the shared purpose of a specific analytics or AI project. Additionally, the people providing the data need to give permission for the data to be used for this purpose, such as through school policy at enrollment. Ideally, data providers should directly understand the value that they will receive because of sharing their data. Finally, the security of that data must be protected, guidelines or policies developed for which roles can access which data, and the level of anonymization needed for specific use case purposes defined.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Privacy and Security principle

Identifying sensitive data, such as personal information, should be part of the use case development process. For individual datasets, sensitive data should be pre-identified, and scripts written to pseudonymize or anonymize specific data fields before they “land” in Stage 2 data lakes (using the OEA Framework) and are accessed by researchers or data scientists. Only this protected data should be used in Stage 3 data lakes for presentation in dashboards.

**How will access to sensitive data be secured and protected in the data environment?**

The Learning Analytics outputs (e.g., dashboards) should only be available to those stakeholders who should have role-based access to such information, implemented through Azure Active Directory, and as governed by the education organization’s policy for data governance. Role-based Access Control (RBAC) is a standard technology solution for controlling access to data.

For more information on how RBAC works though Microsoft Azure Active Directory (which is part of the OEA Architecture), see this article on RBAC for data lake access: [Access control model for Azure Data Lake Storage Gen2 - Azure Storage | Microsoft Learn](https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model) and this broader article on Azure Active Directory: [What is Azure Active Directory? - Microsoft Entra | Microsoft Learn](https://learn.microsoft.com/en-us/azure/active-directory/fundamentals/active-directory-whatis).

**Does the dataset contain any personally identifiable information (PII) and how will that data be protected and governed?**

Learning Analytics use cases will generally include datasets that contain PII from the Student Information System and class rosters. That data can be protected by pseudonymizing this dataset in all dashboards and published insights.

Accountability

Accountability requires that people who develop and deploy AI systems be held responsible for how they operate. AI systems should never be left to operate unchecked, irrespective of the degree to which they may be capable of acting autonomously. This is what is meant by the phrase “humans in the loop.” A part of this is ensuring documentation of the decisions made during the AI system development. This document can be used for that purpose.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Accountability principle

**Who is responsible for reviewing the Use Case documentation and ensuring that the implementation meets responsible AI principles?**

[Must be completed by the team from the educational organization implementing this use case.]

The decision makers (technology, data, or learning engineers) in the educational organization who developing the system should be responsible for continued implementation of the principles of responsible analytics and AI described in this document. They should write, review, and update this document continuously, especially if decisions or data changes.

**How will stakeholders and end users be trained on the appropriate use of the system?**

[Must be completed by the team from the educational organization implementing this use case.]

**How will the analytics or AI system be monitored over time to ensure analytics and prediction perform reliably? Who will be responsible for this?**

[Must be completed by the team from the educational organization implementing this use case.]

Inclusion

The datasets used in learning analytics and AI determine the insights and predictions produced. If those datasets do not represent the whole population of learners, if the data quality is poor, or if certain types of data are not included in the models, it will decrease the accuracy, validity, and inclusiveness of the insights. Similarly, if the way the insights are acted upon by the system does not include all groups (e.g., students with disabilities), it can reinforce exclusion from learning opportunities.

[Video](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6) on Inclusion principle

**What are the constraints of these local data sources for this specific use case?**

In general, extensive data on all student sub-populations is available through an educational organization’s Student Information System dataset.

[Must be completed by the team from the educational organization implementing this use case.]

|  |  |
| --- | --- |
| **Dataset Name** | **Constraints or Limited Representativeness** |
| 1. *For example* | [Must be completed by the team from the educational organization implementing this use case.] |