EXCTRA - EXploiting the Click-TRAil

Assessing the benefits of Learning Analytics

ir. M.A. (Rianne) Conijn

prof.dr. C.C.P. (Chris) Snijders

dr. U. (Uwe) Matzat

dr.ir. P.A.M. (Ad) Kleingeld

W.J.M. (Wouter) Nij Bijvank

Management summary

Currently, Learning Management Systems (LMS) are being used in the majority of educational institutions to provide learning materials online. As a by-product of these systems—every click is recorded—one gets a rich amount of data about students' online behaviour. Recently many researchers have started to investigate these data. Interpreting and contextualizing data about students, to improve learning and teaching, is also known as learning analytics. This document describes the project "EXCTRA - EXploiting the Click-TRAil. Assessing the benefits of Learning Analytics". The main objective of the project, materialized through three reports, is to figure out whether and how learning management system (LMS) data can be used to predict student performance. In particular, early prediction of student performance constitutes an important input for a diversity of educational interventions aiming at reducing student failure.

The first report consists of a literature review, which explicitly identifies gaps in research on the prediction of student performance, which we address in the other reports. The second report consists of a manual that can be used to convert raw LMS log data into analysable data. The manual facilitates the analysis of LMS data by teachers who are not familiar with the data-handling techniques needed for preparation of the LMS data. The third report describes an empirical study using LMS data from seventeen blended courses with 4,989 students taught at Eindhoven university of Technology, combined with data from a test for prospective students (the "TU/e Study Choice Check"). Among other matters, it examines to what extent LMS data can be used for the construction of student performance across the different courses.

REPORT ONE: Literature review

The literature review describes three categories found in the relatively new field of learning analytics. By far the most common topic in learning analytics is the prediction of student performance. These studies show how a wide variety in variables extracted from the data, using a wide variety of analytical methods used, can reveal relations between online behaviour and course performance. Little theory is used to motivate the inclusion of predictor variables, which makes it hard to draw general conclusions about which variables are best in predicting student performance. In addition, most current studies predict student performance only at the *end* of the course, basically only considering whether predicting student performance is possible in principle, but at a time when interventions are not possible anymore. Additionally, often only LMS data are used, while additional student characteristics (for instance high-school grade point average) and performance data (for instance in-between test scores) have been shown to be robust predictors over decades.

The second category in learning analytics consists of analytics and visualization tools, which are made to assist researchers, teachers, and students to analyse and interpret the (complex) LMS data. Several of these tools exist, but they really are in their infancy at this point in time, and in any case quite diverse and mostly applied in just a couple of places, for instance restricted to a handful of courses. The third emerging theme in learning analytics focusses on the actual implementation of the analyses to improve learning and teaching. Research on this theme shows much more promise

and should be extended to get insight in the impact of learning analytics and insight in which interventions are useful in which situations.

REPORT TWO: A manual for pre-processing LMS data

LMS data are stored in large "raw" log tables which are hard to transform into analysable data tables. Moreover, for the prediction of student performance, the data needs to be merged with performance data (grades), which are often stored in a different database with a different data structure. This so-called pre-processing of the data takes a lot of time and effort, especially for teachers and researchers who lack background in data transformation. In fact, we feel this is one of the main reasons why LMS log data are relatively rarely used by educational researchers: they typically do not have the data-handling skills necessary to convert the raw data to an analysable data set. Therefore, our second report offers a manual for pre-processing the raw LMS data and performance data into data which can be used for further analyses, including scripts and explanations of the decisions during the pre-processing process so that any researcher willing to invest a couple days should be able to create an data set that is analysable through standard statistical techniques.

REPORT THREE: Predicting student performance

In the third report we investigate how LMS data can be used. First of all, we characterize the TU/e courses with respect to the LMS features that they use. The courses utilizing Moodle LMS at Eindhoven University of Technology mainly use the LMS to provide content and quizzes. More interactive features such as a discussion forum, wikis, and peer-reviewed assignments were also used, but not consistently throughout many courses. Secondly, we show that LMS data can indeed be used to predict student performance at the course level. However, consistent with previous research, we find that the effects of the LMS predictors differ across courses. One could have naively hoped to find that, say, spending a lot of time online is predictive of a high grade, and we do find courses where this is the case, but we do not find many consistent results of this kind across all courses. Only the in-between assessment grades, the total number of sessions, and the time until the first activity were found to be robust predictors. Hence, it is hard to draw general conclusions, that is, conclusions that hold across all courses, about which LMS data are useful for predicting final exam grades. Still, the data can be used for prediction of student performance per course.

Thirdly, we find that learner data outperforms LMS data in the prediction of student performance. As soon as in-between assessment grades are added to LMS data, learner data has a much lower predictive value. The combination of LMS data and learner data is especially useful for the early prediction of student performance, before the in-between assessments are available. However, the predictions are quite far away from an accurate prediction (confidence intervals typically are the predicted grade plus or minus 1.35 points on scale of 0 to 10), indicating that one has to be careful in using these predictions for early interventions. Fourth, we considered the relationship between LMS data and learner data. We find that most learner data does not correlate strongly with LMS data. However, conscientiousness, time management, and in-between assessment grade did show significant correlations with most of the LMS variables, with low to moderate effect sizes. This offers

3

some promise that, at least for these concepts, LMS data might be of use to measure them continuously, as the university year progresses.

Implications for future use of learning analytics (at TU/e)

Taken together, these three reports show the potential of LMS data, its current limitations, and inform the future use of learning analytics. We end by summarizing the above and some additional findings.

1. It is easy to confuse what is meant by learning analytics, as it contains many different topics: using online LMS data [1] to help students learn better, [2] to help teachers understand (the progress of) their students better, [3] to better understand what leads to student course success, [4] to actually predict student course performance, [5] to get an overview at the university level of the kinds of blended learning that are being offered, or [6] to get an overview at the university level of which ways of blended learning work best. It is crucial to make sure what it is that one is talking about, when talking about learning analytics.

2. In addition, it makes sense to distinguish between **descriptive learning analytics** that is meant to represent overviews of what is happening in an online course (for instance, showing the average time online, or the percentage of material that was viewed) versus **predictive learning analytics** that aims to understand matters that are not obviously found in the LMS data (for instance, being able to predict which student is going to pass the course, based on the online behaviour in the first two weeks). The first part is relatively easy and "just" an ICT-problem, the second is complicated and not even clear that it can be solved.

3. The current learning analytics literature suggests that predicting student success at the course level is potentially possible. Predicting student success across courses seems much harder (or even impossible). Our empirical analyses on TU/e data support this.

4. The current learning analytics literature is not very well developed yet: there is not a lot of theory about what the online behaviour is actually measuring.

5. Current learning analytics tools as offered in the standard LMSs, are quite crude and usually only allow for the inspection of aggregate level course data.

6. All learning analytics require structural access to the data (LMS data, grade data, and performance data). As the pre-processing of the data takes a lot of time and effort, it makes a lot of sense to invest in a general pre-processing method or program at TU/e. The last thing one should want, is that everybody who is trying to do something with LMS data is investing a lot of time in pre-processing the data independently. Our own manual may provide a useful template for the construction of such a method to be used for the future Canvas data. Find a couple of people who do the pre-processing for the rest. This also allows more control over privacy-sensitive matters.

7. For teachers, future use of learning analytics could include the analyses of LMS data in a single (their own) course, including the prediction of student performance and providing interventions for students who are at risk of failing a course.

8. Moreover, LMS data can show which LMS features are rarely used, providing insights for the improvement of the course design that utilizes the full potential of blended courses.

9. To give teachers an overview of their course, visualizations or dashboards can be relatively easily created and evaluated. For students, these dashboards could also be used to inform them about their learning activities.

10. For researchers, future work should investigate generalizable predictors across courses or course offerings. Course characteristics and theoretical concepts referring to student learning behavior and processes need to be included to improve the accuracy and portability of LMS data and to get a better understanding of LMS data and how it can be used to improve learning and teaching**.**