Association for Information Systems

AIS Electronic Library (AISeL)

BLED 2019 Proceedings

BLED Proceedings

2019

From dirty data to multiple versions of truth: How different choices in data cleaning lead to different learning analytics outcomes

Justian Knobbout

Huub Everaert,

Esther van der Stappen

32ND BLED ECONFERENCE
HUMANIZING TECHNOLOGY FOR A SUSTAINABLE SOCIETY
JUNE 16 – 19, 2019, BLED, SLOVENIA, CONFERENCE PROCEEDINGS
A. Pucihar (ed. et al.)

University of Maribor Press

From dirty data to multiple versions of truth: How different choices in data cleaning lead to different learning analytics outcomes

JUSTIAN KNOBBOUT, HUUB EVERAERT & ESTHER VAN DER STAPPEN

Abstract Learning analytics is the analysis of student data with the purpose of improving learning. However, the process of data cleaning remains underexposed within learning analytics literature. In this paper, we elaborate on choices made in the cleaning process of student data and their consequences. We illustrate this with a case where data was gathered during six courses taught via Moodle. In this data set, only 21% of the logged activities were linked to a specific course. We illustrate possible choices in dealing with missing data by applying the cleaning process twelve times with different choices on copies of the raw data. Consequently, the analysis of the data shows varying outcomes. As the purpose of learning analytics is to intervene based on analysis and visualizations, it is of utmost importance to be aware of choices made during data cleaning. This paper's main goal is to make stakeholders of (learning) analytics activities aware of the fact that choices are made during data cleaning have consequences on the outcomes. We believe that there should be transparency to the users of these outcomes and give them a detailed report of the decisions made.

Keywords: • Data Cleaning • Learning Analytics • Student Data • Moodle Data • Outcomes •

CORRESPONDENCE ADDRESS: Justian Knobbout, Researcher/Lecturer, HU University of Applied Sciences Utrecht, The Netherlands, e-mail: justian.knobbout@hu.nl. Huub Everaert, Ph.D., Researcher/Lecturer, HU University of Applied Sciences Utrecht, The Netherlands, e-mail: huub.everaert@hu.nl. Esther van der Stappen, Ph.D., Researcher, HU University of Applied Sciences Utrecht, The Netherlands, e-mail: esther.vanderstappen@hu.nl.

DOI https://doi.org/10.18690/978-961-286-280-0.4 ISBN 978-961-286-280-0
Dostopno na: http://press.um.si

1 Introduction

Virtual Learning Environments (VLEs) are digital learning platforms where students can interact with course materials (presentations, digital readers, instructional video's et cetera), can test their knowledge via quizzes, and can interact with each other and instructors via e.g., the discussion board. They support learning and simultaneously enable the collection of data on learner behavior in the system. Data from virtual learning environments are used for learning analytics activities, cf. Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García (2014); Conijn, Snijders, Kleingeld, & Matzat (2016); Rienties, Toetenel, & Bryan (2015); Romero, Ventura, & García (2008). Objectives of learning analytics vary but often involve student behavior modelling, prediction of performance and increase in (self) reflection and (self) awareness (Papamitsiou & Economides, 2014).

Importantly, raw data exported from virtual learning environments need to be cleaned and transformed before it is of any use to educators and students. In general, data cleaning takes up to 80% of analytical time (Brink, Richards, & Fetherolf, 2016). However, in the current learning analytics field, details about cleaning and transforming are often overlooked or, at best, not described and discussed in literature. For example, searching the terms *data cleaning* or *data preprocessing* in the Learning Analytics & Knowledge conference proceedings 2011 till 2018 (n = 438) only yield 17 papers describing either cleaning or preprocessing of learner data before analyzing the data. To make matters even more complex, full-scale and multimodal learning analytics require aggregated data from multiple sources, amplifying the effects of data cleaning on the analysis' outcomes. As we will show in this paper, data cleaning is problematic as (unspoken) choices can lead to a wide variety of outcomes and, subsequently, pedagogical interventions. Using a raw data set with VLE data, we will construct twelve different, cleaned sets and use these to calculate the time-spent-on the online part of six courses. With these data sets, we can provide an answer to our research question: *"What are the effects of (unspoken) choices made during the cleaning process of student data on the outcomes when these data are in turn used for learning analytics?"*.

The remainder of this paper is structured as follows. First, an in-depth description of learning analytics and data cleaning is given based on existing literature. Then, the research question and method are described, followed by the presentation of our results. Finally, we provide five recommendations based on the outcomes of our study, as well as directions for future work.

2 Related work

In this section, we will present existing literature related to our study. First, we will provide a definition of learning analytics and an overview of the learning analytics process. Next, a thorough description of data cleaning and its implications is given.

2.1 Learning analytics

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 environment in which it occurs" (Siemens et al., 2011). Learning analytics aim to improve learning processes at the level of students and teachers (Siemens & Long, 2011) and is, for example, used to analyze student behavior within digital learning environments, monitor the usage of course material, and predict whether students will fail a certain course or drop out entirely. The process of learning analytics consists of four steps: 1) learners generate learning data, 2) these data are captured, collected and stored, 3) analysis and visualization are performed, and 4) the design and use of data-driven pedagogical interventions (Clow, 2012) – see also Figure 1. Consequently, when the data is incorrect or incomplete, the analysis and subsequent interventions may be sub-optimal or even completely erroneous.

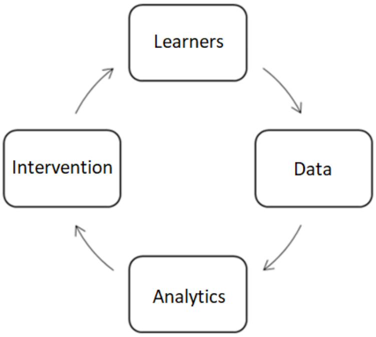


Figure 1: Learning Analytics Cycle (Clow, 2012).

2.2 Time-on-task

Study-time is the time students spend on studying learning materials, using (metacognitive) tools, solving questions etcerera and can be used as measure of affected learning (Knobbout & van der Stappen, 2018). In several studies, a positive correlation between study-time and achievements of students has been found, cf. Marzano (2003); Scheerens & Bosker (1997)). Estimating 'time-on-task' in the 'traditional' classroom is based on estimates by students and/or observations in classrooms. In a virtual learning environment (VLE), on the other hand, it is common to use the number of clicks (Wolff, Zdrahal, Nikolov, & Pantucek, 2013) or the time between certain clicks as measure for time-on-task (Kovanović et al., 2015).

Wolff et al. (Wolff et al., 2013) showed that "even fairly coarse grain data about students' activities" is useful in predicting retention (p. 148). Unfortunately, it is not perfectly clear what part of the clicks were used "[w]hile the issue of data cleaning for all data within the [Open University] was not resolved, it was possible to gain enough knowledge about the data […] to start building models" (p. 146). From their point of view, it is import to note that in predicting failing students, changes in the student's own VLE activity, compared to their previous activity, are indicative. A relative reduction of clicks hints an failing student. Kovanović et al. (2015) deal explicitly and extensively with the thorny methodological issues of estimating time-on-task in VLE's. Their primary goal is "to raise awareness of the issue of accuracy and appropriateness surrounding time-estimation within the broader learning analytics community, and to initiate a debate about the challenges of this process" (p. 184). It is regarded good practice in different academic fields to discuss methodological issues and learning analytics should not become an exception to this rule. In this study, we extent the work of Kovanović et al. by estimating time-on-task for multiple parallel courses and by showing different options to handle missing data, i.e., records of events unlinked to any of the courses in the dataset.

2.3 Data cleaning

Data cleaning is an important part of the ETL (Extraction, Transformation and Load) process. According to VanderPlas (2016) the majority of the work in data science often "comprises cleaning and munging real-world data" (p. 188). Brink, Richards, and Fetherolf (2016) underline five common tasks, of which two - transforming original data to the target and create features that are more easily interpreted – are core business in working with large computer generated data files. Müller and Guido (2016) state that "in the real world, inconsistencies in the data and unexpected measurements are very common" (p. 19). Brink, Richards & Fetherolf (2016) estimate researchers are spending about 80% of their research time to munging, wrangling, combining or reshaping data. Special attention is given to utilizing expert knowledge. Although machine learning can reduce the need to create a set of expert-designed rules, that does not mean that prior knowledge of the application or domain should be discarded. Domain experts can help to identifying useful features that are more informative than the initial representation of the data (Müller & Guido, 2016).

2.4 Missing data

In (social sciences) papers and articles an often-subordinated subject is missing data. One of the most frequent and most ignored sources of bias is missing data (Baguley, 2012). Missing data is a stubborn problem in data analyses and, in general, we have to consider two issues: how much is missing and why it is missing. Thanks to eloquently written textbooks like 'Applied missing data analysis' (Enders, 2010), solutions to deal with missing data mechanisms are nowadays within reach for social researchers. In an overview of traditional techniques, Enders (2010) describes (listwise/pairwise) deletion, several imputation methods, averaging items in Likert scales, or last observation carried forward to address the problem and concludes that "most single case imputation methods produce biased estimates, even with Missing Completely at Random (MCAR) data. Stochastic regression imputation is the one exception and is the only traditional approach that yields unbiased estimates under a Missing At Random (MAR) mechanism" (p. 54). He demonstrates benefits of modern methods like maximum likelihood approaches and multiple imputation. Even in MCAR – which occurrence can hardly be safely assumed – the problems of missing data may become more serious if more cases are missing. "Unfortunately, there are as yet no firm guidelines for how much missing data can be tolerated for a sample of a given size" (Tabachnick & Fidell, 2007) (p. 63). Indirectly, Tabachnick & Fidell (2007) seem to consider about 5% missing or less of the sample size as 'manageable' in some way or the other. It also depends on the pattern of missing data. Choosing among different techniques for dealing with missing data may also depend on knowledge, confidence, and familiarity with the subject matter on part of the researcher. Van Belle (2011) among others advocates sensitivity analysis as a good idea based on "a thorough understanding of the subject matter" (p. 186).

It does not matter whether the above mentioned authors are working in the different fields varying form social or educational sciences, general data sciences to hard core machine learning and it seems fair to conclude that working with data is time consuming and in general comes with trouble, caveats or thorny issues. Fortunately, at the end of the process we will rely on some technical solutions, but working the data is in itself a muddy experience in which the data scientist/researcher has to rely on (several) subjective views and or decisions.

Educators are in the midst of a transition from learning analysis to learning analytics. The analysis of classical test scores is not enough. The availability of VLEs and the tracking of student behavior gives both students and educators much more opportunities to follow the learning of students in real–time and opportunities to intervene if necessary. At the same time, the upper limits of learning analytics are not well defined. Techniques borrowed from educational data mining, data science and machine learning combined with data from social-media become more and more intertwined (Daniel, 2017; Gibson & Ifenthaler, 2017). Technical solutions by themselves are not sufficient for successful use of educational data, as "[d]ata do not exist independently of the ideas, instruments, contexts and knowledge used to generate, process and analyze them" (Kitchin, 2014) (p. 2) thereby (implicitly) suggesting that data scientist are not aware of the pitfalls of data construction. As we will later show in this paper, most data scientists are aware of the true nature of data, that is, data are not neutral, objective and pre-analytic in nature. What often lacks is a thorough discussion of the possible solutions and consequences of a technical data issue, which is a major motive to conduct the study at hand.

3 Research method

The choices made in the cleaning of student data extracted from VLEs has effect on the outcome of this process – the dataset which is used for analysis and visualization of learning. However, not much is written about this effect and, consequently, the differences between outcomes based on the assumptions and choices made by the people responsible for the cleaning of the raw data are also underexposed. This study's aim is to fill this gap in the current learning analytics knowledge based on answering the following research question: *"What are the effects of (unspoken) choices made during the cleaning process of student data on the outcomes when these data are in turn used for learning analytics?"*. As we will research how the made choices affect analytical outcomes of contemporary events whilst we do not have control over these events, a case study is a suitable research method for our study (Yin, 2014).

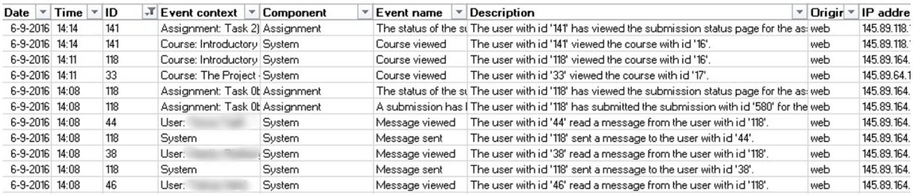
3.1 Case description

In this single case study, we analyze data from an international minor program. Students (n = 34) from the Netherlands, Finland, Spain, United Kingdom, Mexico, and Germany all participate in six blended courses (in this study named A to F), offered in 'traditional' classrooms, at an external workplace, as well as online via Moodle – a well-known VLE. In this study, we focus on data obtained from the latter.

Log files from Moodle are collected by exporting them via the administrator dashboard. This dashboard allows administrators to download all logs in comma separated value (.csv) format, which in turn can be processed in more specialized statistical software or learning analytics tools – in this study, we used IBM SPPS Statistics 24. The data are aggregated by us, i.e., events from all six courses are combined in one dataset. In compliance with the ethical procedures and guidelines that were applicable at the time the research was conducted, students were asked to give passive informed consent and all data were after collection immediately anonymized. Initially, the dataset comprises the variables as shown in Table 1 and Figure 2.

Table 1: Variables extracted from Moodle.

Variable	Description
Date	Date of the event taking place
Time	Time, in HH:MM-format, of the event taking place
User id	Moodle id of the user
Event context	Page of the VLE where event takes place
Component	Whether it involves an assignment or not
Event name	Name of the activity
Description	Description of the event, including course and user(s) id
Origin	Whether website or app is used
IP-address	IP-address from where Moodle is accessed
Id of affected user	In case of e.g., message sent or discussion board reaction



Date	Time	ID	Event context	Component	Event name	Description	Origir	IP addre
6-9-2016	14:14	141	Assignment: Task 2)	Assignment	The status of the su	The user with id '141' has viewed the submission status page for the as:	web	145.89.118.
6-9-2016	14:14	141	Course: Introductory	System	Course viewed	The user with id '141' viewed the course with id '16'.	web	145.89.118.
6-9-2016	14:11	118	Course: Introductory	System	Course viewed	The user with id '118' viewed the course with id '16'.	web	145.89.164.
6-9-2016	14:11	33	Course: The Project -	System	Course viewed	The user with id '33' viewed the course with id '17'.	web	145.89.64.1
6-9-2016	14:08	118	Assignment: Task 0t	Assignment	The status of the su	The user with id '118' has viewed the submission status page for the as:	web	145.89.164.
6-9-2016	14:08	118	Assignment: Task 0t	Assignment	A submission has I	The user with id '118' has submitted the submission with id '580' for the	web	145.89.164.
6-9-2016	14:08	44	User: [redacted]	System	Message viewed	The user with id '44' read a message from the user with id '118'.	web	145.89.164.
6-9-2016	14:08	118	System	System	Message sent	The user with id '118' sent a message to the user with id '44'.	web	145.89.164.
6-9-2016	14:08	38	User: [redacted]	System	Message viewed	The user with id '38' read a message from the user with id '118'.	web	145.89.164.
6-9-2016	14:08	118	System	System	Message sent	The user with id '118' sent a message to the user with id '38'.	web	145.89.164.
6-9-2016	14:08	46	User: [redacted]	System	Message viewed	The user with id '46' read a message from the user with id '118'.	web	145.89.164.

Figure 2: Snippet of raw data set.

As a case for our study, we want to determine for each individual student how much time is spent on each of the six courses of the minor program and the underlying learning activities. This means we have to structure the data in such way that we can estimate the time-on-task for all events in the data set. We elaborate on this process and its results in the next section.

3.2 Cleaning of the data

Our focus in the ETL process of the Moodle data is on cleaning and transforming the data by deriving new calculated variables and values by splitting a column (existing variable) into multiple columns (new variables) and so disaggregating the data. Our VLE data records user id, event description and timing of an event. The variable Description (including the user id and course id) is split in different variables to identify the course the student is working on. We are willing to assume that a student's action in the VLE and thus creating an event in the data set is synonymous with studying. Therefore, we have to assume that opening of a second event implies the end of the first event and the time-spent-on the first event T1 amounts to t2 minus t1 – see Figure 3. Unfortunately, closing of the event is normally not registered in the VLE. Consequently, time-spent-on the last event in a session (T4 in Figure 3) cannot reliably be calculated.

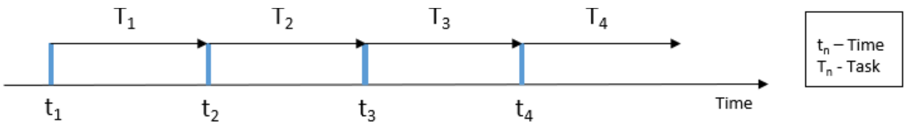


Figure 3: Calculation of time-spent-on task by using the start of new event.

Another issue is missing data: many events are not linked to a specific course. For example, when a student sends a message to another student, Moodle does not know to what course (if any at all) the message relates and therefore omits the inclusion of a course id in the event description. This proves problematic when calculating the total-time-spent-on a course. In Figure 4 we see that a student is working on course D at t2. Later, at t5, he is involved in course C. In order to link the other events (t1, t3, t4, t6, and t7) to a specific course to compute total-time-spent-on a course, we must make some assumptions.

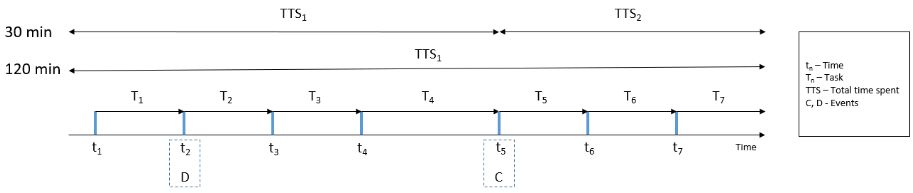


Figure 4: Total-time-spent on different courses, based on varying session times.

First, we must decide whether the event on t_4 is to be associated with a session in which the student is working on course C or course D. In the literature, a session or study-period often ends 30 minutes after the last click (see discussion and overview of time-on-task in (Kovanović et al., 2015)). Moodle's default setting, however, automatically ends sessions after 120 minutes. That are two main versions we worked with in this study, but there is no logical reason to limit ourselves to these options – why not 60 or 90 minutes? By deciding to end a session after 30 minutes of inactivity, we also assumed that the course worked on in the 30 minutes version is D at t_1, t_2, t_3 and t_4, while the student started with course C at t_5. We can now calculate the total-time-spent (TTS) during this session by adding all T_x within the session. In the default Moodle version, on the other hand, the timing between all events is smaller than the 120 minutes cut-off time. In such a study period (see Figure 4), we can calculate the total-time-spent during the session but do not know to what (portion of a) course to assign it. It can be DDDDCCC, but also DDCCCCC or whatever permutation possible. Obliviously, this is of influence when computing total-time-spent-on a course.

To deal with the problem of events not linked to courses – which is essentially a missing data issue – we defined six scenarios:

- In the first scenario (strict) we disregarded sessions with events not referring to any course. This way, we do not have to make assumptions to what course a session relates. The downside, however, is that we lose sessions and, thus, information.
- In the second scenario (wide 1), we filled out the missing values by carrying the last observation forward till the next observed course or the end of the study session.
- In the third scenario (wide 2), we simply relied on the most frequent course in a study period as the one and only; overwriting missing values in that particular time frame.

In the other three scenarios, we imputed the missing values with randomly assigned courses weighted by the number of known courses worked on:

- In the fourth scenario (wide 3), the weight was based on the number of all courses observed on a weekly basis of all students together and all missing values of a single student in a particular time frame got the same random course assigned (for instance, AAA or BBB)
- In the fifth scenario (wide 4), the same is done as in wide 3 but several missing values in a particular computed study-period were independently randomized (for instance, DBA, or CAC or just FFF).

- In the final scenario (wide 5), the weight is computed by the number of courses directly chosen by an individual student on a weekly basis and missing values were imputed as in wide 4.

We just want to show that all scenarios are plausible in one way or the other, and indeed, we could have chosen other ways to deal with missing values. At this point we are not interested in the stability of the different approaches. In order to compute the total-time-spent-on a course (TTS$_{A}$, TTS$_{B}$ et cetera) in the different versions and session, we recomputed the study sessions by taking t-last minus t-first of a row of equal courses in order to estimate time-spent-on a course. See Figure 5 for a schematic representation of some of the scenarios.

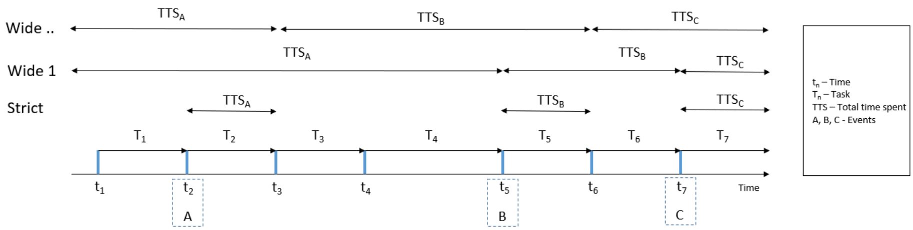


Figure 5: Schematic representation of scenarios Strict, Wide 1 and one of the other Wides.

3.3 Data processing

In line with our own recommendations (see section 5.1), we provide a summary of assumptions and decisions made in the processing of our data:

- Events related to accessing the VLE with phones or mobile apps creates records without any information other than that a mobile device is used and can be removed from the dataset;
- Activities as changing passwords or failed login attempts are not related to learning and thus can be removed from the dataset;
- Our research focusses on learners so event caused by other users (teachers, administrators etcetera) can be removed from the dataset;
- All remaining events in the dataset represent learning activities in the VLE;

- Learning sessions end either 30 or 120 minutes after the start of the last event in said session;

Data is cleaned by applying one of the six methods described in section 3.2.

4 Results

Now we have 12 different data sets – the six scenarios how to deal with missing data and two different sessions times (30 versus 120 minutes). With these data sets, we now calculate the time-spent-on the six courses of the minor program.

4.1 Identifying events and courses

In total, our raw dataset comprised 148,285 events. After removing events related to accessing the VLE with phones or mobile apps removing non-learning activities, and limiting ourselves to student users, we end up with 57,811 events. Of all these events, just 12,334 events (21% of relevant events) are directly linked to a course – see Figure 6. This leaves 45,477 events (79%) unaccounted for and the only way to link the registered student activity to a course is within a study session based on the Moodle default of 120 minutes or the 30 minutes often used in academic studies.

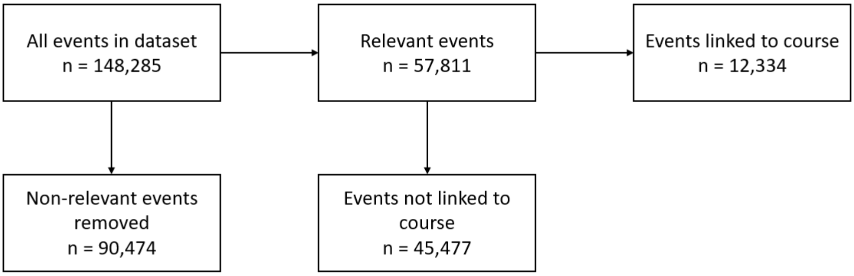
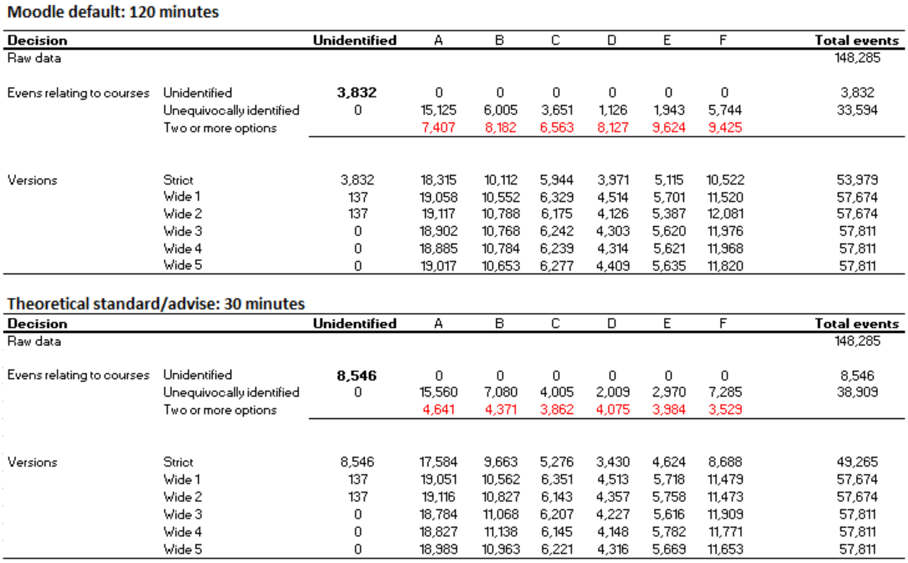


Figure 6: Number of events during and after data processing.

As a result of the option between 30 and 120 minutes, we see in Table 2 that in the 120 minute default 3,832 events take place within study periods in which there is no link to any course at all. Just by shortening the end of the study session to 30 minutes, the number of not directly identifiable events more than doubles to 8,546 events. Shorter periods in the 30 minutes version leads to more unequivocally identifiable events; all known events in these periods belong to one and the same course. In the 120-minute default, it is just the opposite: the number of events pertaining to two or more different courses within a study period nearly doubles compared to the 30 minutes variant. Independent of the selected version, there are 137 not directly identifiable events we could not solve by carrying the last observation forward till the next observed (wide 1) or just taking the most frequent course in a study session (wide 2).

Table 2: Assigning events to courses in order to estimate time-spent-on course.



Moodle default: 120 minutes

Decision		**Unidentified**	A	B	C	D	E	F	**Total events**
Raw data									148,285
Evens relating to courses	Unidentified	**3,832**	0	0	0	0	0	0	3,832
	Unequivocally identified	0	15,125	6,005	3,651	1,126	1,943	5,744	33,594
	Two or more options		7,407	8,182	6,563	8,127	9,624	9,425	
Versions	Strict	3,832	18,315	10,112	5,944	3,971	5,115	10,522	53,979
	Wide 1	137	19,058	10,552	6,329	4,514	5,701	11,520	57,674
	Wide 2	137	19,117	10,788	6,175	4,126	5,387	12,081	57,674
	Wide 3	0	18,902	10,768	6,242	4,303	5,620	11,976	57,811
	Wide 4	0	18,885	10,784	6,239	4,314	5,621	11,968	57,811
	Wide 5	0	19,017	10,653	6,277	4,409	5,635	11,820	57,811

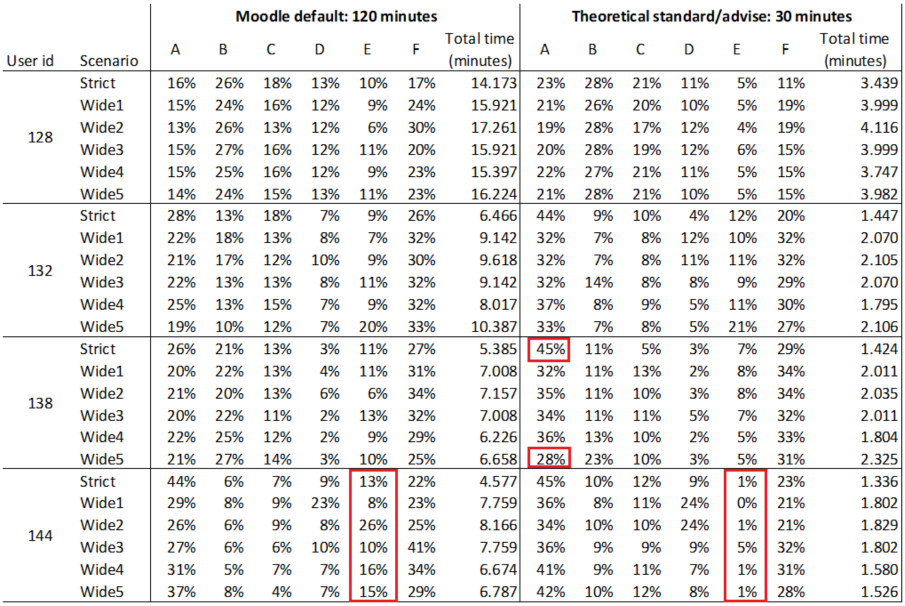
Theoretical standard/advise: 30 minutes

Decision		**Unidentified**	A	B	C	D	E	F	**Total events**
Raw data									148,285
Evens relating to courses	Unidentified	**8,546**	0	0	0	0	0	0	8,546
	Unequivocally identified	0	15,560	7,080	4,005	2,009	2,970	7,285	38,909
	Two or more options		4,641	4,371	3,862	4,075	3,984	3,529	
Versions	Strict	8,546	17,584	9,663	5,276	3,430	4,624	8,688	49,265
	Wide 1	137	19,051	10,562	6,351	4,513	5,718	11,479	57,674
	Wide 2	137	19,116	10,827	6,143	4,357	5,758	11,473	57,674
	Wide 3	0	18,784	11,068	6,207	4,227	5,616	11,909	57,811
	Wide 4	0	18,827	11,138	6,145	4,148	5,782	11,771	57,811
	Wide 5	0	18,989	10,963	6,221	4,316	5,669	11,653	57,811

4.2 Identifying time-spent-on tasks and courses

After cleaning the data and imputing the missing values, we have 12 datasets and can calculate the number of activities on each course based on the various data sets. At first glance it seems that only differences between the strict and the wide scenarios are noteworthy. The solutions within the five wide approaches do not differ that much. That is erroneous: the number of events in Table 2 are presented over all students together. What we really want to know is the number of events – and more importantly – time-spent-on by each individual student. Both measures vary enormously according to the chosen dataset. We can now also calculate the time-spent-on each course by each individual student as shown in Table 3 for just four students.

Table 3: Relative amount of time spent on courses for four different students.



		Moodle default: 120 minutes							Theoretical standard/advise: 30 minutes						
User id	Scenario	A	B	C	D	E	F	Total time (minutes)	A	B	C	D	E	F	Total time (minutes)
128	Strict	16%	26%	18%	13%	10%	17%	14.173	23%	28%	21%	11%	5%	11%	3.439
	Wide1	15%	24%	16%	12%	9%	24%	15.921	21%	26%	20%	10%	5%	19%	3.999
	Wide2	13%	26%	13%	12%	6%	30%	17.261	19%	28%	17%	12%	4%	19%	4.116
	Wide3	15%	27%	16%	12%	11%	20%	15.921	20%	28%	19%	12%	6%	15%	3.999
	Wide4	15%	25%	16%	12%	9%	23%	15.397	22%	27%	21%	11%	5%	15%	3.747
	Wide5	14%	24%	15%	13%	11%	23%	16.224	21%	28%	21%	10%	5%	15%	3.982
132	Strict	28%	13%	18%	7%	9%	26%	6.466	44%	9%	10%	4%	12%	20%	1.447
	Wide1	22%	18%	13%	8%	7%	32%	9.142	32%	7%	8%	12%	10%	32%	2.070
	Wide2	21%	17%	12%	10%	9%	30%	9.618	32%	7%	8%	11%	11%	32%	2.105
	Wide3	22%	13%	13%	8%	11%	32%	9.142	32%	14%	8%	8%	9%	29%	2.070
	Wide4	25%	13%	15%	7%	9%	32%	8.017	37%	8%	9%	5%	11%	30%	1.795
	Wide5	19%	10%	12%	7%	20%	33%	10.387	33%	7%	8%	5%	21%	27%	2.106
138	Strict	26%	21%	13%	3%	11%	27%	5.385	45%	11%	5%	3%	7%	29%	1.424
	Wide1	20%	22%	13%	4%	11%	31%	7.008	32%	11%	13%	2%	8%	34%	2.011
	Wide2	21%	20%	13%	6%	6%	34%	7.157	35%	11%	10%	3%	8%	34%	2.035
	Wide3	20%	22%	11%	2%	13%	32%	7.008	34%	11%	11%	5%	7%	32%	2.011
	Wide4	22%	25%	12%	2%	9%	29%	6.226	36%	13%	10%	2%	5%	33%	1.804
	Wide5	21%	27%	14%	3%	10%	25%	6.658	28%	23%	10%	3%	5%	31%	2.325
144	Strict	44%	6%	7%	9%	13%	22%	4.577	45%	10%	12%	9%	1%	23%	1.336
	Wide1	29%	8%	9%	23%	8%	23%	7.759	36%	8%	11%	24%	0%	21%	1.802
	Wide2	26%	6%	9%	8%	26%	25%	8.166	34%	10%	10%	24%	1%	21%	1.829
	Wide3	27%	6%	6%	10%	10%	41%	7.759	36%	9%	9%	9%	5%	32%	1.802
	Wide4	31%	5%	7%	7%	16%	34%	6.674	41%	9%	11%	7%	1%	31%	1.580
	Wide5	37%	8%	4%	7%	15%	29%	6.787	42%	10%	12%	8%	1%	28%	1.526

Compared to the theoretical standard/advise of 30 minutes, students spend about 3 to 4 times as much time on the total of six courses under the Moodle default of 120 minutes. Considering Moodle's default session ending time of 120 minutes, students spent about 3 to 4 times as much time on their courses compared to the total time-on-task when using the theoratical standard ending time of 30 minutes. This is in line with the assumptions used – 120 minutes is four times as long as 30 minutes. However, if we look at the relative time students spent on specific courses between the two versions or within the used scenarios of a version, the link between assumptions used and relative time becomes foggy and blurred.

In the 30 minutes version, all students seem to spend relatively more time on course A and less on course B, compared to the Moodle default of 120 minutes. However, student 144 spends also relatively less time on course E. If we compare over the scenarios within the separate versions, we sometimes see huge differences between strict and several wide scenarios. For instance, in the 30 minutes version, student 138 spends 45% of his time in the strict version to course A, in wide 5 this is reduced to a mere 28%. In the Moodle default, the relative time-spent-on in these sets is more or less the same (26% versus 21%).

As our results show, it is difficult to see a common pattern in these figures, indicating different assumptions lead to different dashboard figures. Concluding, we observe that time-spent-on as a key variable for the quality of learning stays without reach for teachers as a basis to act upon and interfere with a particular student: it just depends and variates with the assumptions made and the truth is hard to find.

5 Discussion and conclusion

In this paper, we have shown that the choices made during the cleaning process of student data can have large impact on the outcome of the subsequent analysis. Estimating time-on-task is one example of a learning (outcome) measure which is affected by data cleaning, but also other metrics used in learning analytics research might be influenced, e.g., the use of (metacognitive) tools or the number of discussion board postings. With the emerge of full-scale and multimodal learning analytics – requiring the aggregation of data from multiple sources –the effects of data cleaning on the analysis' outcomes are even more amplified. We are not in search of a holy grail for student data cleaning (which probably does not exist at all), but the goal of this study is to make both practitioners and academics aware of these - often unspoken - choices and their effect.

5.1 Recommendations

Based on our research, we present the following recommendations: (1) provide users of learning analytics tools (students, teachers et cetera) with the insight what assumptions and corresponding choices were made during the data cleaning process. This helps them to better understand the results and visualizations of the data analysis; (2) provide users with the opportunity to see other versions based on different assumptions of the data set as well; (3) to make scientific work better reproducible and comparable, researchers should elaborate on the cleaning of their data. In the current literature, researchers often almost immediately jump from raw data to results without saying anything on the choices made, although some exceptions exist, cf. Bos & Brand-Gruwel (2016); Chen, Chen, & Xing (2015); Kovanović et al. (2016); (4) involve domain experts in the cleaning process. Data experts working on the data sets without knowing the exact context the data was collected in, might use erroneous assumptions to clean the data. By consulting domain experts before the data handling, the resulting data might be better suit the learning context (Müller & Guido, 2016); (5) stakeholders should feel responsible, support the choices made, and be transparent about them.

If we want students, colleagues and other professionals to work with our analysis, results or dashboard functionality, we should be open and give them a detailed report of the decisions made. As a rule of thumb, we should state and explain explicitly how we have dealt with the issues at hand in such way the user can understand it (Van Belle, 2011).

5.2 Future work

Now we have different data sets, we might want to research in what ways to inform end users about the data cleaning process. That is, how can we inform users – students, teachers et cetera – what assumptions were made, what steps were taken, what user preferences are, and what the effects on the analysis outcome are. We propose the use of focus groups to identify (critical) success factors for awareness creation about data cleaning and its consequences.

References

Agudo-Peregrina, Á F., Iglesias-Pradas, S., Conde-González, M. Á, & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. Computers in Human Behavior, 31, 542-550.

Baguley, T. (2012). Serious stats: A guide to advanced statistics for the behavioral sciences. Palgrave Macmillan.

Bos, N., & Brand-Gruwel, S. (2016). Student differences in regulation strategies and their use of learning resources: Implications for educational design. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, 344-353.

Brink, H., Richards, J., & Fetherolf, M. (2016). Real-world machine learning. Manning Publications Co.

Chen, B., Chen, X., & Xing, W. (2015). Twitter archeology of learning analytics and knowledge conferences. In Proceedings of the Fifth International Conference on

Learning Analytics and Knowledge, 340-349.

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 134-138.

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. IEEE Transactions on Learning Technologies,

Daniel, B. K. (2017). Big data in higher education: The big picture. Big data and learning analytics in higher education (pp. 19-28) Springer.

Enders, C. K. (2010). Applied missing data analysis. Guilford Press.

Gibson, D. C., & Ifenthaler, D. (2017). Preparing the next generation of education researchers for big data in higher education. Big data and learning analytics in higher education (pp. 29-42) Springer.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences Sage.

Knobbout, J., & van der Stappen, E. (2018). Where Is the Learning in Learning Analytics? In European Conference on Technology Enhanced Learning (pp. 88-100).

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Penetrating the black box of time-on-task estimation. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, 184-193.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, 15-24.

Marzano, R. J. (2003). What works in schools: Translating research into action ASCD.

Müller, A. C., & Guido, S. (2016). Introduction to machine learning with python: A guide for data scientists " O'Reilly Media, Inc.".

Papamitsiou, Z. K., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology & Society, 17(4), 49-64.

Rienties, B., Toetenel, L., & Bryan, A. (2015). Scaling up learning design: Impact of learning design activities on lms behavior and performance. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, 315-319.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368-384.

Scheerens, J., & Bosker, R. (1997). The foundations of educational effectiveness Pergamon.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S. P., Shum, S., Ferguson, R., Baker, R. (2011). Open learning analytics: An integrated & modularized platform.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 30.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics Allyn & Bacon/Pearson Education.

Van Belle, G. (2011). Statistical rules of thumb. John Wiley & Sons.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data " O'Reilly Media, Inc.

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, 145-149.

Yin, R.K. (2014) Case study research and applications: Design and methods. Thousand Oaks: Sage publications