Using A/B Testing in MOOC Environments

Jan Renz
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Germany
jan.renz@hpi.de

Daniel.Hoffmann
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Germany
daniel.hoffmann@hpi.de

Thomas Staubitz
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Germany
thomas.staubitz@hpi.de

Christoph Meinel
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Germany
christoph.meinel@hpi.de

ABSTRACT

In recent years, Massive Open Online Courses (MOOCs) have become a phenomenon offering the possibility to teach thousands of participants simultaneously. In the same time the platforms used to deliver these courses are still in their fledgling stages. While course content and didactics of those massive courses are the primary key factors for the success of courses, still a smart platform may increase or decrease the learners experience and his learning outcome. The paper at hand proposes the usage of an A/B testing framework that is able to be used within an micro-service architecture to validate hypotheses about how learners use the platform and to enable data-driven decisions about new features and settings. To evaluate this framework three new features (Onboarding Tour, Reminder Mails and a Pinboard Digest) have been identified based on a user survey. They have been implemented and introduced on two large MOOC platforms and their influence on the learners behavior have been measured. Finally this paper proposes a data driven decision workflow for the introduction of new features and settings on e-learning platforms.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]; H.5 [**Information interfaces and presentation**]; K.3.1 [**Computer Uses in Education**]; J.4 [**Social and Behavioral Sciences**]

Keywords

MOOC, A/B Testing, microservice, E-Learning, Controlled Online Tests

LAK '16, April 25 - 29, 2016, Edinburgh, United Kingdom
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4190-5/16/04...$15.00
DOI: http://dx.doi.org/10.1145/2883851.2883876

1. INTRODUCTION

1.1 Controlled Online Tests

In the 18th century, a British naval captain wondered why sailors serving on the ships of the Mediterranean countries did not suffer from scurvy. On those ships, citrus fruits were part of the rations. So he ordered one half of his crew to eat limes (the treatment group), while the other half consumed the same rations they received before (the control group). Despite the displeasure of the crew, the experiment was successful. Without knowing the cause of the effect (that lack of vitamin C caused scurvy), he found out that limes prevented it [18]. This lead to citrus fruits being a part of the sailor's rations and a healthier crew on all ships.

In the late 1990s, Greg Linden, a software engineer at *Amazon*, developed a prototype showing product recommendations based on the current shopping cart content at checkout [13]. He was convinced that transferring the impulse buys, like candy at the checkout lane, from grocery stores to online shopping and improving them by personalization would increase the conversion rate and so lead to more income for the shop. While he received positive feedback from his co-workers, one of his bosses, a marketing senior vice-president strongly opposed his idea because he believed it would distract customers from checking out and therefore lead to a loss of revenue. So Linden was forbidden to work on it any further. Being convinced of the possible impact he did not follow this management decision, but instead launched a controlled online test. One group of customers saw the recommendations, the other did not. The senior vice-president was furious when he found out that the feature was launched. But it "won by such a wide margin that not having it live was costing Amazon a noticeable chunk of change", so he could not keep up his concerns. The feature was rolled out for all users short time later. Today testing is an essential part of amazons philosophy.

These two examples show how experimentation helps to validate hypotheses with data and how they may also contradict intuition and preconceptions. As pointed out by Thomke, "experimentation matters because it fuels the discovery and creation of knowledge and thereby leads to the development and improvement of products, processes, systems, and organizations." ([25])

Experimentation has long been costly and time-consuming as it would require special lab setups or paying agencies, but the web makes it possible to quickly and cost-efficiently evaluate new ideas using controlled experiments, also called A/B tests, split tests, or randomized experiments [9]. As stated in [8] it can be expected that small changes can have a big impact to key metrics. They also integrate well with agile methodologies, such as the ones described in *Lean Startup* by Eric Ries, which "is an approach for launching businesses and products, that relies on validated learning, scientific experimentation, and iterative product releases to shorten product development cycles, measure progress, and gain valuable customer feedback." ([11]). He states that no one, in despite of his expertise can fully anticipate the users' behaviour, so only by testing the best solutions for both the user and the provider can be determined.

MOOCs (used here as a synonym for scalable e-learning platforms) provide their service to thousands of learners, so they have a critical mass of users that enables the platform providers to run those controlled online experiments. Instead of using these tests to increase conversion rates or sales, the aim here is to identify instruments to optimize the learning experience and the learning outcome of those users.

1.2 openHPI

This work focuses on the MOOC platforms *openHPI* and *openSAP*. *openHPI* is a non-profit project provided by the Hasso Plattner Institute (HPI) in Potsdam, Germany for opening courses derived from the curriculum of the Institute for the general public.

The web university team of the chair of Internet and Web Technologies had previous experience with online learning research, having established the *tele-TASK* platform for recorded HPI lectures. They also provide a tele-recording system. But they have never been fully satisfied with the usage of the provided content.

In November 2012, the first MOOC in German language was held on openHPI, rendering HPI one of the first European MOOC providers. In 2014 an average of 7,000 participants have been enrolled at course end [17]. SAP, a well-known German software company, published their first MOOC on openSAP in May 2013. It targets professionals working with SAP products and is also used to educate SAP employees [22].

Both providers use the same underlying system, internally called *Xikolo* (Tsonga, a Bantu language, for school). Thus, the implementation of the A/B Testing framework and the changes to the user interface are equally applicable for openHPI and openSAP and have been applied in the academic context as well as in the enterprise learning context.

This paper describes the introduction of an A/B-Testing Framework to a micro-service based MOOC platform and the results obtained evaluating this service with different A/B tests. The remainder of the paper at hand is structured as follows:

- Section 2 gives an overview on the architecture of the A/B Testing Framework and the underlying Learning Analytics Engine.
- Section 4 describes how the possible test candidates have been identified.
- In Section 5 the three A/B tests conducted and their results are introduced and discussed.
- A conclusion and a discussion of future work can be found in Section 6.

2. A/B TESTING IN MICROSERVICE BASED LEARNING PLATFORMS

With the advent of MOOCs a large amount of educational data became available. There are two communities dealing with its analysis: *Learning Analytics* and *Educational Data Mining*. While they have many things in common, both are concerned about how to collect and analyze large-scale educational data for a better understanding of learning and learners, they have slightly different goals [23]. Learning Analytics aims at providing insights to teachers and learners, whereas Educational Data Mining rather focuses on automatic adaptation of the learning process with not necessarily any human interference.

For a better evaluation of learning data across different MOOCs, a general database schema was proposed by Veeramachaneni et al. called *MOOCdb* [26, 7]. The authors suggest developing a "shared standard set of features that could be extracted across courses and across platforms" ([26]). The schema includes three different modes named *observing*, *submitting*, *collaborating* and *feedback*.

Another approach is the Experience API (also known as xAPI or TinCan API) suggested by the Advanced Distributed Learning (ADL) Initiative [1]. It defines a way to store statements of experience, typically but not necessarily in a learning environment. A statement has at least three parts *actor*, *verb* and *object* representing subject, verb and object in a sentence. Additional properties can include references to resources like an UUID as *id*, a *result* denoting the outcome, contextual information in *context* or the time of the statement in *timestamp*.

In order to gather learning data on openHPI, a versatile and scalable solution called Lanalytics which allows to track user actions in a service-oriented environment [19] (for details on openHPI's distributed architecture see subsection 2.1) was implemented. The recorded actions can be stored in a variety of different formats (such as MOOCdb and Experience API) and data stores (such as PostgreSQL[1], a relational database, elasticsearch[2], a document database, and Neo4j[3], a graph database). This *LAnalytics* framework sets the foundation for further research in this work.

2.1 openHPI Architecture

openHPI is based on a micro-service architecture [15], which means there is no monolithic application, but multiple services, each with a defined responsibility [12]. The decision

[1] PostgreSQL: `http://www.postgresql.org`
[2] elasticsearch: `https://www.elastic.co`
[3] Neo4j: `http://neo4j.com`

to go for a Service Oriented Architecture was based on the learnings that resulted from employing and extending a monolithic application to run MOOCs in a previous version of the platform. Each service runs in its own process and handles only a small amount of data storage and business logic. This approach has a number of advantages. As the services are designed around capabilities, each service can use the technology that serves best the use case including different programming languages or DBMS that fit best [12]. Currently all but one service is implemented as a RubyOnRails application due to the existing developer qualification. Scaling in a micro-service architecture can be realized by distributing the services across servers, replicating only those needed. With a monolithic application, the complete application has to be replicated. Each service can be deployed independently, which makes it easier to continuously deploy new versions of the services [20]. In contrast to monolithic applications a fault in one service does not necessarily affect the whole application. Lastly, micro-services are relatively small and therefore easier to understand for a developer. Most of openHPI's developers are students and spend only a few hours per week actively developing. Therefore, this architecture not only minimizes the risk of breaking other parts of the software (by isolation), it also enables developers to become experts in a certain part of the app (exposed by one or more services).

While having many advantages, the presented architecture prohibits using one of the many available A/B-Testing solutions like the Ruby gems *split*[4] and *vanity*[5]. These libraries are designed to work within monolithic applications. Other existing solutions, such as *Optimizely* use JavaScript to alter the interface and to measure events. These solutions mostly target marketing driven A/B Tests with a simple set of metrics and changes (for example display a different pricetag or alternative landing page). But in our case many functionalities that might me relevant for A/B testing are not only part of the User Interface (UI). Instead they might include actions that happen in one of the underlying services or even asynchronous actions that are not UI related at all. This is where UI focused approaches will fail.

Additionally, the measured metric is not simply tracking conversions, but queries possibly complex data gathered by the Learning Analytics framework [24]. Furthermore the used metrics may consist of learning data. Keeping this data within the system and not sending it to a 3rd party tool avoids problems with data privacy. So a dedicated custom prototype was built to enable A/B testing in the Xikolo-framework.

2.2 Workflow

Each time a user accesses a page within the learning platform, the system detects if there are any tests currently running in the scope of the visited page by querying the *Grouping Service*. If there are tests running, the system needs to check if the user has one of this test features enabled. This check is handled by the *Account Service*. It will return an already given test group assignment or create a new one by applying a set of rules and deciding if the user should be in the test or the control group for each requested test. In Figure 1 the communication between the different parts of the system is shown in more detail. While this workflow generates additional requests, there was no measurable performance decrease of the front-end, as this calls could be run in parallel with other calls to the server backend. All code that is related to a function that is currently in A/B testing must be encapsulated in a code block. This code will only be executed if this user is part of the test group. This way of implemented features could later be used to make this feature being active or deactivated on a per platform or per user base using so called feature flippers, so this can be considered no extra work.

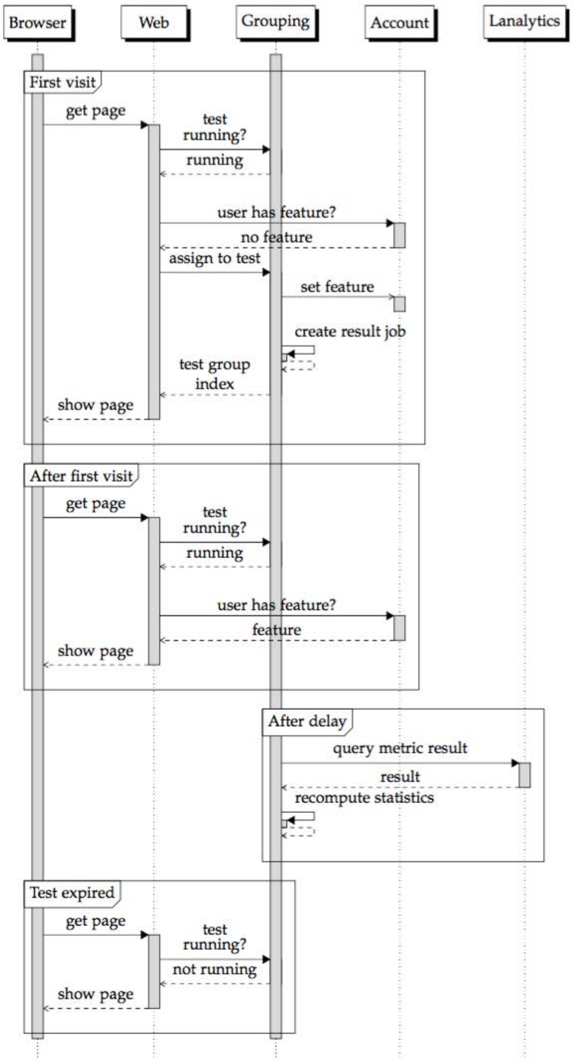


Figure 1: Abstract sequence diagram showing the communication between the browser and the services.

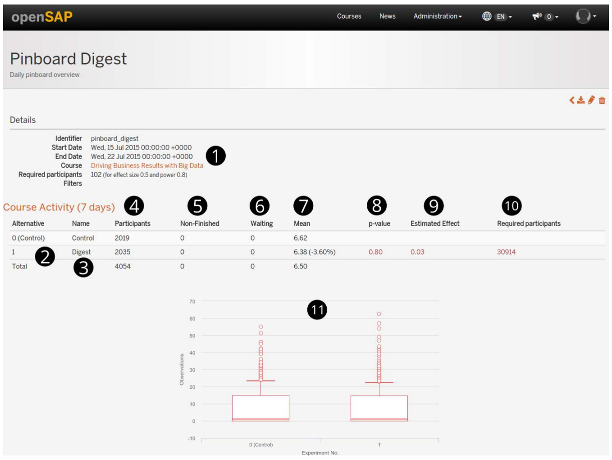
2.3 Administrators Dashboard

The creation of new AB tests with a certain complexity involves writing additional code and taking care that this code is well tested and rolled out, so this part can only be

[4] split, the Rack Based A/B testing framework: `https://github.com/splitrb/split`

[5] Vanity, Experiment Driven Development for Ruby: `https://github.com/assaf/vanity`

provided by the development team. The management of running A/B tests can be achieved using a newly introduced section within the backend of the learning software. There, administrators (or all users equipped with the needed set of permission) can enable, edit and view user tests. This includes not only the meta data of the user tests, but also the live test results. All those users can see the gathered data on a dashboard shown in item 2. For each metric the number of participants, the number of participants that did not yet finish the user test and the number of participants for whom the metric wait interval did not end yet is displayed. If a metric has been evaluated for some users in both groups the effect size is displayed, calculated as *Cohen's d* [3].



Pinboard Digest

Daily pinboard overview

Details

Identifier	pinboard_digest
Start Date	Wed, 15 Jul 2015 00:00:00 +0000
End Date	Wed, 22 Jul 2015 00:00:00 +0000
Course	Driving Business Results with Big Data
Required participants	102 (for effect size 0.5 and power 0.8)
Filters	

Course Activity (7 days)

Alternative	Name	Participants	Non-Finished	Waiting	Mean	p-value	Estimated Effect	Required participants
0 (Control)	Control	2019	0	0	6.62			
1	Digest	2035	0	0	6.38 (-3.60%)	0.80	0.03	30914
Total		4054	0	0	6.50			

Figure 2: Screenshot of the administrators dashboard of a user test showing *1*) general properties of the test, *2*) and for each metric the indices *3*) and names of the test groups, *4*) the number of participants, *5*) the number of participants that did not finish the test, *6*) the trials waiting for the metric result, *7*) the mean of the group, *8*) the p-value of statistical significance, *9*) the effect size, *10*) the required number of participants for a power of 0.8, *11*) box plots of the group results.

3. METRICS

Witte and Witte define *quantitative* data as "a set of observations where any single observation is a number that represents an amount or a count", whereas qualitative data is defined as "a set of observations where any single observation is a word, or a sentence, or a description, or a code that represents a category" ([28]).

Thus, quantitative data describes the intensity of a feature and is measured on a numerical scale. Qualitative data has a finite number of values and can sometimes be ordinally scaled. Qualitative usability studies observe *directly* how the user interacts with the technology, noting their behavior and attitudes, while quantitative studies *indirectly* gather numerical values about the interaction, mostly for a later mathematical analysis [21].

Each user test can have multiple metrics based on quantitative data, for example if the user enrolled in the course in question or the number of specific actions performed by the user in a given time frame. Most metrics require some time to pass in between the beginning of the user test (the user being assigned to one of the groups and presented with a certain functionality) and the measurement of the metrics. If the user test is course-specific, only actions concerning this course are queried. The amount of time relates on the metrics. Metrics that are based on the learning outcome might need a certain amount of self tests done by the users or the course to be ended. Other metrics that focus on user activity may need at least some days.

Most of the metrics query data is gathered by the LAnalytics service. This service processes messages sent in the services on certain events, for example if a user asks a new question, answers one or watches a video. This data is then sent and received using the *Msgr* gem[6], which builds on *RabbitMQ*[7]. The received events are then transformed and processed by several pipelines. While this is an asynchronous processing, usually all events are processed near real time. The LAnalytics service allows the usage of different storage engines, however all relevant events for the metrics for this tests are stored in an *elasticsearch* instance using the *Experience API* [1] standard. An Experience API statement consists of four parts: subject, verb and object, in this case user, verb and resource. The resource needs a UUID (Universally Unique Identifier) and can contain additional information for faster processing for example the question title. Additionally, the statement has a timestamp and a context, for example the course ID.

The following metrics are currently implemented and can be used within A/B tests:

3.1 Pinboard Posting Activity

The pinboard posting activity counts how often a user asks, answers and comments questions and discussions in the pinboard of a course.

Verbs: ASKED_QUESTION, ANSWERED_QUESTION, COMMENTED

3.2 Pinboard Watch Count

The pinboard watch count denotes the number of viewed questions and discussions of a user.

Verb: WATCHED_QUESTION

3.3 Pinboard Activity

This pinboard activity combines pinboard posting activity and pinboard watch count. Considering the different amounts of effort, a weighting is applied. The posting activity contributes with a ratio of 90%, while the watch count is weighted with 10%.

3.4 Question Response Time

The question response time denotes how long after a question was asked, the question is answered by a user. To compute this metric all Experience API statements with the verb ANSWERED_QUESTION are retrieved for a user, then the matching ASKED_QUESTION statement is queried and

[6] Msgr: `https://github.com/jgraichen/msgr`
[7] RabbitMQ: `https://www.rabbitmq.com/`

the average difference between this timestamps is computed. Since not all users answer questions in the specified time frame, empty values need to be allowed, but these values are removed before significance testing.

3.5 Visit Count

The visit count denotes how many items a user visited, including videos, selftests and text parts. This metric can be filtered by time and course. Verbs: VISITED

3.6 Video Visit Count

The video visit count denotes the number of visited videos per user. This metric can be filtered by time, video and course.

Verb: VISITED
Filter: content_type == video

3.7 Course Activity

The course activity summarizes the aforementioned metrics to measure the overall activity of a user in a course. The pinboard activity is weighted with 50%, while the visit count is included without weight.

3.8 Course Points

After the end of a course the number of points are persisted and the quantiles of the users' points are calculated. For each enrollment a *completed* event is emitted, which is received and processed by the LAnalytics Service. The course points metric returns the number of points a user received in a specified course.

Verbs: COURSE_COMPLETED

3.9 Micro Survey

Not all interface changes can be evaluated with an objective metric, for example design changes. For these cases a qualitative feedback metric is used. It allows for fast evaluation by prompting users to rate whether they like the displayed version. In contrast to the other metrics, this one is just a concept and is not yet implemented. For this metric every users would be asked to rate a functionality or a design. Then the ratings provided by test and control group can be compared.

4. IDENTIFYING TEST CANDIDATES

To utilize the power of an A/B Testing framework, possible test candidates must be identified and selected.

4.1 Dropout and Absence in MOOCs

Since MOOCs can be joined freely and impose no commitment on the user, there is a high number of students who do not visit the course after enrollment, stop visiting it after a while, or leave it completely. The reported dropout rate on *Coursera* is 91% to 93% [10] and on *openHPI*[8] it is between 77 and 82% [17, 16]. So the number of registrations should be seen as an indicator of interest rather than the ambition to finish the course. Halawa et al. [6] claim that not only complete dropout is a problem, but also periods of absence which have an impact on the user's performance. While 66% of all students of the analyzed course with an absence of less than two weeks entered the final exam and scored 71% on average, only 13% of the students that were absent longer than one month took the final exam with a mean score of 46%.

Several recent works addressed this issue. One countermeasure is to make the course content available for every interested person. Only if wanting to take an assignment or to contribute to the forums a registration is necessary. This way people that just want to take a look at the content but are not interested in taking the course are filtered out from the participants.

Yang et al. [29] point out that higher social engagement corresponds with lower dropout, because it "promotes commitment and therefore lower attrition". This was also shown by Grünewald et al. [5] in an analysis of the first two *openHPI* courses. However, one half of the participants did not actively participate in forum discussions. *openHPI* programming courses have higher completion rates than other courses. An average of 31% received a certificate in the two programming courses, while the average completion rate in 2014 was 19.2% [17]. The courses provide an interactive programming environment. Exercises have predefined test cases, against which students can try their code against. This higher engagement of learners might be a reason for the larger completion rate.

4.2 User Experience Survey

For a prior investigation of how users perceive their experience on openHPI, we conducted a survey. It was announced via an email to all users and on openHPI's social media channels. From March 25, 2015 to May 25, 2015, all users have been asked for their opinion about their user experience on and the usability of the platform. The survey contained questions about existing functionalities, but also about unfinished or unpublished functionalities and functionalities not available on the platforms, but maybe available on other MOOC platforms. The survey yielded 512 responses of which 161 were incomplete. For the following evaluation only the complete responses are considered. 61% of the participants are older than 40 years and 63% are male. 71.6% of all participants are satisfied with the overall usability of openHPI (a rating of 4 or 5 on a scale from 1 to 5). 73% were satisfied with the learnability, 73.1% with the video player and 71.9% with the tests. Only the discussions deviate from these results. They have a satisfaction rate of 61.5%. Additionally, when asked whether the discussions support their learning process, only 36.1% agreed. Regarding gamification, 29.7% rated the importance of gamified elements for them with 4 or 5. 34.9% agreed that gamification elements would influence their course participation in a positive way.

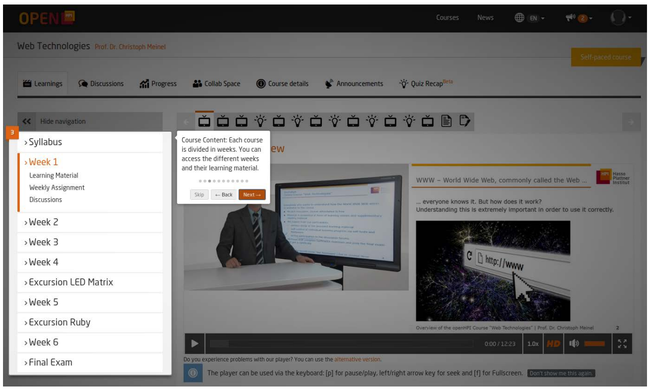
In conclusion, the overall perception of the usability of the openHPI platform is at a high level, but the discussions are not as helpful as intended. The didactical concept and expectation and the user perception diverge. This gap can be closed using the experimentation framework and should be addressed when optimizing the learning outcome.

5. CONCLUDED TESTS

[8] openHPI: `https://open.hpi.de`

Based on the survey results three tests have been selected to evaluate the introduced A/B testing framework based on the expected impact. The selection was based on the predicted user acceptance and the expected impact in combination with the amount of work needed to implement these new features. As the learners in MOOCs are connected via the forum, it is also important to respect this fact while choosing possible test candidates, as this could lead to confusion or jealousy. While all of these tests required to implement prototypes of these features none of these functionalities were so essential or prominent that not having it may lead to disappointed users. Some of the tests featured additional explanatory text, explaining that the user is part of a test group. One possible test candidate featuring gamification elements which are really prominent on the platform was not chosen for this reason. As we run several platforms, a possible test strategy is to roll it out on one instance of the platform only and then "normalize" the metrics. All tests could be easily stopped or deactivated by the user, disabling the tested feature for that user.

5.1 Onboarding



Course Content: Each course is divided in weeks. You can access the different weeks and their learning material.

... everyone knows it. But how does it work? Understanding this is extremely important in order to use it correctly.

Do you experience problems with our player? You can use the alternative version.

The player can be used via the keyboard: [p] for pause/play, left/right arrow key for seek and [f] for Fullscreen.

Figure 3: Third step of the tour explaining the course navigation

It may be useful for inexperienced users to get an overview about the system's functionality which could lead to a more intense usage of the MOOC platform. A tour was created that lets them visit the most important pages and explains the features in eleven steps: It starts automatically after a user enrolls for their first course and highlights a single part of the currently displayed page while providing some additional explanatory test as shown in Fig. 3. The first steps explain the course area containing the week and item navigation. Then the user is forwarded to the pinboard page and difference between questions and discussions (questions are posted with a specific problem in mind and have answers, discussions want to debate certain course contents and have only comments) are explained. Afterwards, the progress page is opened and the progress layout and how to retrieve certificates after course end is explained. The tour was implemented using *intro.js*[9] in a version modified to support multiple pages.

To validate the hypothesis, we conducted a user test using the aforementioned testing framework.

The course activity metric (subsection 3.7) and the pinboard activity metrics (subsection 3.3) were used to validate the impact of the alternative group.

5.1.1 Alternatives

After enrollment the groups saw:

Group 0: a confirmation that they are enrolled.

Group 1: a welcome message and a tour guiding them through the course area.

5.1.2 Setup

The test ran for a week starting on May 20, 2015 17:20 targeting users who enrolled for their first course on openHPI. It started after enrollment and ended immediately for the control group and after skipping or finishing the tour for the treatment group.

The control group comprised 172 participants, the alternative 119 (plus 16 that did not finish the tour).

All metrics were evaluated after one week.

5.1.3 Results

The results (Table 1, Table 2) show that an onboarding tour increases the number of visits of learning items (34.5 % for videos, 27.9 % for all items). However, the difference is not significant, p<0.05.

Table 1: Onboarding: Results for visit count

Name	Participants	Mean	Change	p
Control	172	11.49		
Tour	119	14.70	+27.93%	0.15
Total	291	12.80		

Table 2: Onboarding: Results for video visit count

Name	Participants	Mean	Change	p
Control	172	4.01		
Tour	119	5.39	+34.48%	0.11
Total	291	4.58		

The change in pinboard activity is negative (-7%, Table 3).

Table 3: Onboarding: Results for pinboard activity

Name	Participants	Mean	Change	p
Control	172	0.27		
Tour	119	0.25	-6.99%	0.55
Total	291	0.26		

5.2 Reminder Mails

A possible measure to prevent that learners drop out of the course over time is to send reminder emails after a certain period of inactivity. Since on most platforms of the openHPI ecosystem new contents are published on Mondays, we chose a period of four days to reach participants before the weekend, when they have more time to work on the course.

[9] intro.js: `https://github.com/usablica/intro.js`

The emails are sent by the Notification Service, which is responsible of receiving updates sent in other services and forwarding them as a notification on the web or as an email. The regular email for this test is initiated using a daily routine at 2:00 UTC. It queries all enrollments of users that have not visited their course for the last four days and have visited less than 90% of the course content. The latter restriction prevents users that nearly finished a course, but are not interested in some parts of it to repeatedly receive reminders. Depending on the group, the popular questions of the course's discussions of the four last days and unwatched videos are queried for each enrollment.

5.2.1 *Alternatives*

The test is designed in a multivariate manner and comprises three alternative groups. The groups are sent:

Group 0: no email

Group 1: an email reminding them to visit the course again

Group 2: an email reminding them to visit the course again including an extract of the latest activity in discussions

Group 3: an email reminding them to visit the course again including videos they did not see yet

Group 4: an email reminding them to visit the course again including an extract of the latest activity in discussions and videos they did not see yet

An exemplary email as a user part of group 4 would receive is show in Figure 4. As with all notifications, the users can opt-out of these emails.

5.2.2 *Setup*

The test ran for two weeks starting on July 6, 2015, 22:00 UTC, targeting all users enrolled in *Web Technologies* on openHPI. A trial started when the first reminder email was sent and ended upon successful delivery.

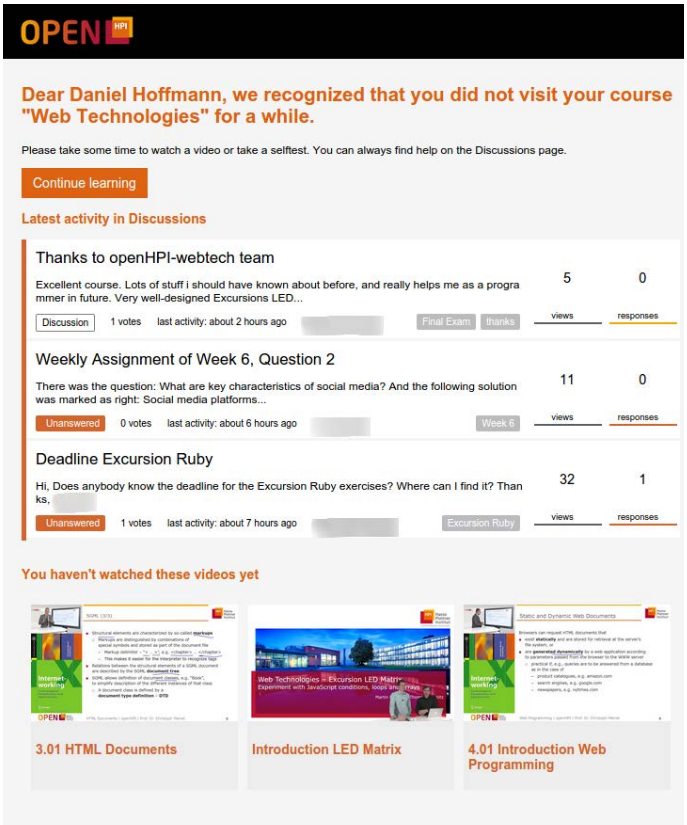
The control group comprised 1830 participants, the alternatives 1831, 1833, 1834, and 1868 summing up to 9196 participants in total.

All metrics were evaluated after one week. Only the course point metric was evaluated after the certificates were published.

5.2.3 *Results*

The results show that sending a reminder email increases the overall course activity (Table 4). However, only emails containing videos show a statistically significant change of 37.4% with a p-value of 0.02 $<$ 0.05 for videos and 43.3% (p-value 0.009 $<$ 0.05) for videos and questions.

The same effect can be seen for the visit count (Table 5) with an increase of 38.3% and a p-value of 0.018 $<$ 0.05 for videos, and 43.5% (p-value 0.009 $<$ 0.05) for videos and questions and even stronger for the video visit count (Table 6) with a gain of 60.4% and a p-value of 0.004 $<$ 0.05 for videos, and 73.1% (p-value 0.002 $<$ 0.05) for videos and questions.



Dear Daniel Hoffmann, we recognized that you did not visit your course "Web Technologies" for a while.

Please take some time to watch a video or take a selftest. You can always find help on the Discussions page.

Continue learning

Latest activity in Discussions

Thanks to openHPI-webtech team

Excellent course. Lots of stuff i should have known about before, and really helps me as a programmer in future. Very well-designed Excursions LED...

Discussion 1 votes last activity: about 2 hours ago [redacted] Final Exam thanks 5 views 0 responses

Weekly Assignment of Week 6, Question 2

There was the question: What are key characteristics of social media? And the following solution was marked as right: Social media platforms...

Unanswered 0 votes last activity: about 6 hours ago [redacted] Week 6 11 views 0 responses

Deadline Excursion Ruby

Hi, Does anybody know the deadline for the Excursion Ruby exercises? Where can I find it? Thanks, [redacted]

Unanswered 1 votes last activity: about 7 hours ago [redacted] Excursion Ruby 32 views 1 responses

You haven't watched these videos yet

3.01 HTML Documents

Introduction LED Matrix

4.01 Introduction Web Programming

OPEN HPI
Web Technologies – Excursion LED Matrix
Experiment with JavaScript conditions, loops and arrays
SGML (3/3)
Static and Dynamic Web Documents
Internet-working

Figure 4: Sample reminder email for group 4

Table 4: Reminder Emails: Results for course activity

Name	Participants	Mean	Change	p
Control	1830	1.12		
Text	1831	1.34	19.92%	0.109
Questions	1833	1.2	6.96%	0.337
Videos	1834	1.53	37.36%	**0.02**
Q. and V.	1868	1.6	43.32%	**0.009**
Total	9196	1.36		

Table 5: Reminder Emails: Results for visit count

Name	Participants	Mean	Change	p
Control	1830	1.1		
Text	1831	1.33	20.63%	0.102
Questions	1833	1.19	7.45%	0.327
Videos	1834	1.53	38.26%	**0.018**
Q. and V.	1868	1.58	43.46%	**0.009**
Total	9196	1.35		

The results for pinboard activity (Table 7) are surprising, as they reveal a decrease of discussions visits for all alternatives. The decrease is less for the alternatives showing questions, but the results still indicate that reminder emails have no impact, if not a negative impact (that was not tested), on the number of visits in the discussions. A possible explanation could be that users that saw the recommended content of

Table 6: Reminder Emails: Results for video visit count

Name	Participants	Mean	Change	p
Control	1830	0.47		
Text	1831	0.6	29.0%	0.07
Questions	1833	0.59	26.87%	0.11
Videos	1834	0.75	60.38%	**0.004**
Q. and V.	1868	0.81	73.08%	**0.002**
Total	9196	0.64		

the pinboard could realise that there is no content that would motivate them to visit the pinboard, while otherwise they may just have browser there and then explored some interesting threads.

Table 7: Reminder Emails: Results for pinboard watch count

Name	Participants	Mean	Change	p
Control	1830	0.12		
Text	1831	0.07	-46.22%	0.915
Questions	1833	0.08	-38.22%	0.846
Videos	1834	0.06	-51.23%	0.932
Q. and V.	1868	0.1	-14.33%	0.635
Total	9196	0.09		

Another outcome we did not expect was that the emails have no positive effect on the total points achieved in the course. As Table 8 shows, the means of all alternatives are inferior to that of the control group.

Table 8: Reminder Emails: Results for course points

Name	Participants	Mean	Change	p
Control	1830	9.57		
Text	1831	8.43	-11.88%	0.899
Questions	1833	7.87	-17.71%	0.977
Videos	1834	8.71	-8.97%	0.83
Q. and V.	1868	9.51	-0.63%	0.526
Total	9196	8.82		

5.3 Pinboard Digest Mails

In the pinboard students can discuss course contents and ask questions about advanced topics to help them to understand the taught content better. As outlined in subsection 4.1 social interaction also prevents dropouts. However, in the user experience survey on *openHPI* (see subsection 4.2), the statement “Did the discussions support your learning process?” received only a rating of 2.93 on a scale from 1 to 5. This test evaluated if a daily pinboard overview email fuels the participation in the discussions of a course. Such an email contains questions the user has not yet seen, but received much attention from others. It also includes unanswered questions to reduce the time until a question is answered.

These emails are also sent each day at 4:00 UTC. For each enrollment in an active course with an open forum it is determined if the user visited the course in the last two weeks. This avoids disturbing users that are dropped out to receive these emails. Reminder emails are designed for those cases. A background job queries the pinboard service for unseen questions for that user and unanswered questions in the course.

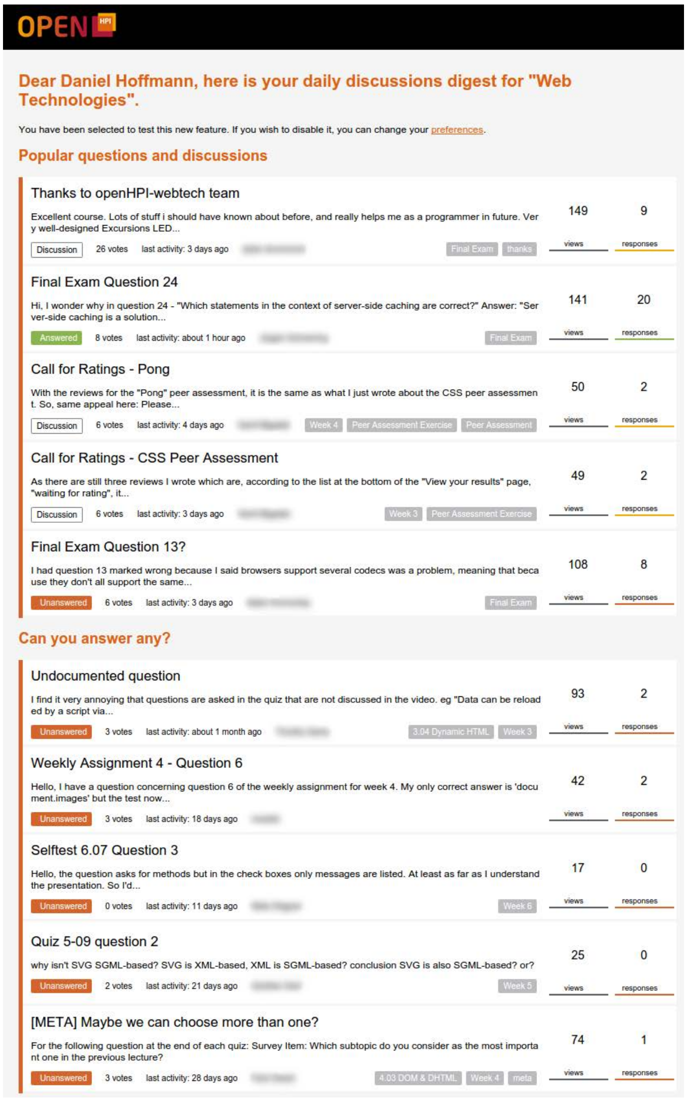
5.3.1 Alternatives

The groups are sent:

Group 0: no email

Group 1: an email including five unseen questions with the most activity and five unanswered questions with the demand to answer them

An exemplary email is show in Figure 5. Similar to the reminder emails, users can disable these emails.



Dear Daniel Hoffmann, here is your daily discussions digest for "Web Technologies".

You have been selected to test this new feature. If you wish to disable it, you can change your preferences.

Popular questions and discussions

Thanks to openHPI-webtech team

Excellent course. Lots of stuff i should have known about before, and really helps me as a programmer in future. Very well-designed Excursions LED...

Discussion 26 votes last activity: 3 days ago [redacted] Final Exam thanks | 149 views | 9 responses

Final Exam Question 24

Hi, I wonder why in question 24 - "Which statements in the context of server-side caching are correct?" Answer: "Server-side caching is a solution...

Answered 8 votes last activity: about 1 hour ago [redacted] Final Exam | 141 views | 20 responses

Call for Ratings - Pong

With the reviews for the "Pong" peer assessment, it is the same as what I just wrote about the CSS peer assessment. So, same appeal here: Please...

Discussion 6 votes last activity: 4 days ago [redacted] Week 4 Peer Assessment Exercise Peer Assessment | 50 views | 2 responses

Call for Ratings - CSS Peer Assessment

As there are still three reviews I wrote which are, according to the list at the bottom of the "View your results" page, "waiting for rating", it...

Discussion 6 votes last activity: 3 days ago [redacted] Week 3 Peer Assessment Exercise | 49 views | 2 responses

Final Exam Question 13?

I had question 13 marked wrong because I said browsers support several codecs was a problem, meaning that because they don't all support the same...

Unanswered 6 votes last activity: 3 days ago [redacted] Final Exam | 108 views | 8 responses

Can you answer any?

Undocumented question

I find it very annoying that questions are asked in the quiz that are not discussed in the video. eg "Data can be reloaded by a script via...

Unanswered 3 votes last activity: about 1 month ago [redacted] 3.04 Dynamic HTML Week 3 | 93 views | 2 responses

Weekly Assignment 4 - Question 6

Hello, I have a question concerning question 6 of the weekly assignment for week 4. My only correct answer is 'document.images' but the test now...

Unanswered 3 votes last activity: 18 days ago [redacted] | 42 views | 2 responses

Selftest 6.07 Question 3

Hello, the question asks for methods but in the check boxes only messages are listed. At least as far as I understand the presentation. So I'd...

Unanswered 0 votes last activity: 11 days ago [redacted] Week 6 | 17 views | 0 responses

Quiz 5-09 question 2

why isn't SVG SGML-based? SVG is XML-based, XML is SGML-based? conclusion SVG is also SGML-based? or?

Unanswered 2 votes last activity: 21 days ago [redacted] Week 5 | 25 views | 0 responses

[META] Maybe we can choose more than one?

For the following question at the end of each quiz: Survey Item: Which subtopic do you consider as the most important one in the previous lecture?

Unanswered 3 votes last activity: 28 days ago [redacted] 4.03 DOM & DHTML Week 4 meta | 74 views | 1 responses

Figure 5: Sample pinboard digest email for group 1

5.3.2 Setup

The test ran for one week starting on July 15, 2015, 00:00 UTC, targeting all users enrolled in *Driving Business Results* *with Big Data* on *openSAP*. A trial started when the first pinboard digest email was sent and ended upon successful delivery. The control group comprised 2019 participants, the alternative 2035 summing up to 4054 participants. All metrics were evaluated after one week.

5.3.3 Results

The results show that sending daily discussions overviews increases the pinboard activity of the affected users (Table 9) by 64% which is a statistically significant improvement (p-value 0.021 < 0.05). They also raise the posting activity in the discussions (Table 10) by 140%, which is statistically significant as well (p-value 0.03 < 0.05).

Table 9: Pinboard Digest Mail: Results for pinboard activity

Name	Participants	Mean	Change	p
Control	2019	0.06		
Digest	2035	0.1	63.98%	0.021
Total	4054	0.08		

Table 10: Pinboard Digest Mail: Results for pinboard posting activity

Name	Participants	Mean	Change	p
Control	2019	0.02		
Digest	2035	0.04	139.77%	0.03
Total	4054	0.03		

6. CONCLUSION AND FUTURE WORK

The introduction of an A/B Testing Framework gives platform developers, architects and providers the needed toolset to be aware of the impact of new instruments and features before introducing them to all users. This lays the foundation for constant improvement of the user experience and learning outcome on openHPI and MOOC platforms in general. The possibility to have a live view on the current data situation was very helpful and empowered agile decisions like extending tests or run them on other instances of the portal as well. While the evaluation of the new framework was successful, the majority of the test runs have been successful.

It is not uncommon for A/B-tests to fail. In fact, reported success rates hover around 10 and 12.5% [4, 2], the others show little change between the groups. But in the case of the onboarding test the margin was rather large but not significant. This test should be repeated over a longer time span to examine if the results are similar and whether a larger sample size causes statistical significance.

The reminder email test confirmed our hypothesis that reminding students after a certain period of absence increases their course activity. Showing videos in the email was the decisive factor for statistical significance, which indicates that users are more content driven, not social driven. We also received positive feedback from users who appreciated the feature. However, the test also yielded surprising results. The email decreased the pinboard visits regardless of the fact whether it included questions or not. It also did not affect the course results of the learners in a positive way. A possible explanation could be that it was performed in the last two weeks of the course running for six weeks. Users that dropped out before might either be not determined to complete the course in the first place or were deterred by the workload needed to catch up. A test running over the full length of a course could show if the results can be reproduced.

The test concerning pinboard digest emails verified our assumption that it increases the participation in the discussions. The alternative succeeded by the large margin of 140% more questions, answers and comments.

Some of the results confirm our hypotheses and some contradict our intuition. Hence, the tool justifies its place in the openHPI ecosystem. It allows to decide which features should be included backed with data. Disadvantageous ones are not published permanently and only those with a positive impact on the learning behavior or experience are included.

Based on this results next steps could be taken to improve this framework. This includes introducing pre- and post tests, as well as other actions to allow better interpretation as suggested in [14].

6.1 New feature release cycle

Thanks to the presented framework feature decisions for *openHPI* and *openSAP* can be based on empirical data. The workflow of releasing new features now looks as follows. After the implementation of the feature it is reviewed internally in the development team. Depending on the importance of the feature, it is also reviewed by the customer / product owner. If the reviews are positive the feature is deployed, but deactivated using *feature flippers*. Metrics are defined to assess the feature. Then one or more user tests are performed. After a predefined time span the test results are reviewed. If there is an improvement in the metrics and there is no negative impact (for example comunicative pollution) the feature is rolled out for all users, since it evidentially improves the learning behavior of the participants. If no change is detected, it is rolled out but deactivated by default. Users that want to use the feature can activate it. If the test reveals that the new feature performs worse than the control, it is not rolled out.

This new workflow allows for fast evaluation of features. Rather than deciding by intuition which features should be added, they are tested beforehand if they are beneficial for the users. Only those that perform better are activated for all users.

6.2 Evaluation of used metrics and introduction of negative metrics

The used metrics could be evaluated based on the measuremnt of this metrics in courses. For the tests where the activity was increased but the course results stayed the same, additional metrics should be introduced to assure that there is no negative impact of the introduced test. Therefore several metrics that are based on ”negative” events should be introduced. As during the time of evaluation of the new framework such events were not recorded by the learning analytics service, these metrics are not yet implemented. One possible metric is the amount of course or platform unenrollments. Another possible metric is the amount of users unsubscribing from notification mails. This metric could help to indicate users being annoyed by too many mails received from the platform.

6.2.1 Marketing or content driven style A/B tests

All tests run in the context of the evaluation of this new framework are based on new functionalities. Still, it could be also used to run A/B test evaluating smaller, more content-driven or UI driven changes within the platform. This work could be started on the course detail pages. These pages can be compared to classical landing pages, therefore it can be predicted that they have a huge unused potential. An enrollment rate per page visitor rate could be used as a metric. However, given the simple requirements of tests like this could also be run by using Optimizely or other externally provided testing tools. As Willems states in [27] other MOOC platforms like edX allows optional content modules for a given sub set of learners. Also Khan Academy allows A/B testing on certain exercise types. All of these AB tests may be supported by the platform, but do not include functional additions. These types of tests involve additional creation of content, therefore they are hard to realize given the bottleneck of content creation, however the presented A/B testing framework could be used for tests like this.

6.2.2 Community contributed test cased

Another idea of gathering test cases is to collect them from users and the research community. Therefore the authors are open for suggestions of tests. It is planned to at least run three community contributed tests within the next month.

7. REFERENCES

[1] Advanced Distributed Learning (ADL). Experience API v1.0.1, 2014.

[2] P. Chopra. Appsumo reveals its A/B testing secret: only 1 out of 8 tests produce results - VWO Blog, 2011.

[3] J. Cohen. *Statistical Power Analysis for the Behaviorial Sciences*. Lawrence Erlbaum Associates, Inc, 1977.

[4] M. A. Farakh. Most of your AB-tests will fail, 2013.

[5] F. Grunewald, E. Mazandarani, C. Meinel, R. Teusner, M. Totschnig, and C. Willems. openHPI - A case-study on the emergence of two learning communities. In *2013 IEEE Global Engineering Education Conference (EDUCON)*, pages 1323–1331. IEEE, Mar. 2013.

[6] S. Halawa, D. Greene, and J. Mitchell. Dropout Prediction in MOOCs using Learner Activity Features. *eLearning Papers*, 37(March):1–10, 2014.

[7] S. Halawa and U.-m. O. Reilly. MOOCdb : Developing Standards and Systems to support MOOC Data Science Massachusetts Institute of Technology. Technical report, 2014.

[8] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1857–1866, New York, NY, USA, 2014. ACM.

[9] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, volume 2007, page 959, New York, New York, USA, 2007. ACM Press.

[10] D. Koller. MOOCs on the Move: How Coursera Is Disrupting the Traditional Classroom, 2012.

[11] J. H. Kolodziej. *The lean team.*, volume 37. 2011.

[12] J. Lewis and M. Fowler. Microservices, 2014.

[13] G. Linden. Geeking with Greg: Early Amazon: Shopping cart recommendations, 2006.

[14] M. W. Lipsey, K. Puzio, C. Yun, M. A. Hebert, K. Steinka-Fry, M. W. Cole, M. Roberts, K. S. Anthony, and M. D. Busick. Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*, 2012.

[15] C. Meinel, M. Totschnig, and C. Willems. openHPI: Evolution of a MOOC Platform from LMS to SOA. *Proceedings of the 5th International Conference on Computer Supported Education*, pages 593–598, 2013.

[16] C. Meinel, C. Willems, J. Renz, and T. Staubitz. Reflections on Enrollment Numbers and Success Rates at the openHPI MOOC Platform. In *Proceedings of the European MOOC Stakeholder Summit 2014*, pages 101–106, Lausanne, Switzerland, 2014.

[17] OpenHPI. Blick nach vorn – und zurück!, 2015.

[18] M. Q. Patton, P. H. Rossi, H. E. Freeman, and S. R. Wright. Evaluation: A Systematic Approach., 1981.

[19] J. Renz, T. S. Suarez, Gerardo Navarro, and C. Meinel. Enabling schema agnostic learning analytics in a service-oriented mooc platform. L@S '16.

[20] C. Richardson. Microservices Architecture pattern, 2014.

[21] C. Rohrer. When to Use Which User Experience Research Methods, 2008.

[22] Sap Se. About openSAP, 2014.

[23] G. Siemens and R. S. J. D. Baker. Learning analytics and educational data mining. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, page 252, New York, New York, USA, 2012. ACM Press.

[24] G. N. Suarez. *Enabling Learning Analytics in a Service-Oriented MOOC Platform.* Master's thesis, Hasso Plattner Institute, 2015.

[25] S. Thomke. *Experimentation Matters.* 2003.

[26] K. Veeramachaneni and U.-m. O'Reilly. Developing data standards and technology enablers for MOOC data science. *MOOC Research Initiative Conference*, (October 2013):1–8, 2013.

[27] J. J. Williams, K. Ostrow, X. Xiong, E. Glassman, J. Kim, S. G. Maldonado, N. Li, J. Reich, and N. Heffernan. Using and designing platforms for in vivo educational experiments. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, pages 409–412, New York, NY, USA, 2015. ACM.

[28] R. S. Witte and J. Witte, S. *Statistics.* John Wiley & Sons, 9 edition, 2009.

[29] D. Yang, T. Sinha, D. Adamson, and C. Rose. "Turn on, Tune in, Drop out": Anticipating student dropouts in Massive Open Online Courses. *Proceedings of the NIPS Workshop on Data Driven Education*, pages 1–8, 2013.