Developing an Evaluation Framework of Quality Indicators for Learning Analytics

Citation for published version (APA):

DOI:
10.1145/2723576.2723629

Document status and date:
Published: 01/01/2015

Document Version:
Early version, also known as pre-print

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
https://www.ou.nl/taverne-agreement

Take down policy
If you believe that this document breaches copyright please contact us at:
pure-support@ou.nl
providing details and we will investigate your claim.

Downloaded from https://research.ou.nl/ on date: 15 Sep. 2023
Developing an Evaluation Framework of Quality Indicators for Learning Analytics

Maren Scheffel
Open Universiteit Nederland
Valkenburgerweg 177
6419 AT Heerlen,
The Netherlands
maren.scheffel@ou.nl

Hendrik Drachsler
Open Universiteit Nederland
Valkenburgerweg 177
6419 AT Heerlen,
The Netherlands
hendrik.drachsler@ou.nl

Marcus Specht
Open Universiteit Nederland
Valkenburgerweg 177
6419 AT Heerlen,
The Netherlands
marcus.specht@ou.nl

ABSTRACT

This paper presents results from the continuous process of developing an evaluation framework of quality indicators for learning analytics (LA). Building on a previous study, a group concept mapping approach that uses multidimensional scaling and hierarchical clustering, the study presented here applies the framework to a collection of LA tools in order to evaluate the framework. Using the quantitative and qualitative results of this study, the first version of the framework was revisited so as to allow work towards an improved version of the evaluation framework of quality indicators for LA.

Categories and Subject Descriptors
I.6.4 [Computing Methodologies]: Model Validation and Analysis; J.1 [Computer Applications]: Administrative Data Processing—Education; K.3.1 [Computers and Education]: Computer Uses in Education

General Terms
Design, Standardization, Verification

Keywords
evaluation framework, assessment of learning analytics tools, quality indicators, group concept mapping

1. INTRODUCTION

Over the years that learning analytics (LA) have become more and more prominent, the number of tools and applications using such techniques as well as publications about them has grown rapidly. And although the added value of LA for learners as well as for educators has clearly been recognised in the last few years [12], research on the comparability of empirical LA studies and their tools is sparse. The comparison of LA approaches, i.e. their measures, algorithms, results, effects, etc., is hardly possible due to the lack of a comprehensive knowledge base about what makes a good, effective, efficient, useful LA tools in a given situation.

We therefore developed a framework of quality indicators (QIs) for LA to help standardise the evaluation of LA tools [10]. The framework comprises five criteria (Objectives, Learning Support, Learning Measures and Output, Data Aspects, and Organizational Aspects) with four QIs each (see Figure 1). In order to ensure an organically grown and accepted evaluation framework (EF), stakeholders active in the domain of LA have been involved in the development process of the framework using a group concept mapping (GCM) approach.

The aim of the evaluation study presented in this paper is to find out whether the framework developed is applicable to evaluate LA tools or whether it needs to be further adapted, changed, restructured or defined differently for another evaluation cycle. This paper is thus building on the results of the previous study [16], as well as drawing on the experience of Drachsler et al. [6] who have already successfully created an EF specifically for data competitions in TEL.

The rest of this paper is structured as follows: Section 2 presents the methodology to evaluate the framework by applying it to a number of LA tools, followed by the presentation of quantitative as well as qualitative study results. Section 3 then revisits the first framework version taking the previous results into account and presents ways to work towards an improvement of the framework for the next evaluation cycle. Section 4 concludes the paper.

2. FRAMEWORK EVALUATION STUDY

2.1 Methodology

For the evaluation of the framework two things had to be done: on the one hand the framework needed to be turned into an applicable tool itself and on the other hand a collection of LA tools to validate the framework against had to be compiled. As a first step, the framework’s criteria and QIs were therefore transformed into a questionnaire using Google Forms[1]. For every quality indicator the questionaire asked (1) whether that QI was present in/supported by a tool, (2) in what way that QI was present in/supported by the tool, and (3) how difficult or easy (on a scale of 1 (very difficult) to 5 (very easy)) it was to judge that QI. At the end of each criterion section participants were offered an open text box asking for

any additional comments.

To find suitable LA tool candidates the submissions to the previous Learning Analytics and Knowledge conferences as well as a number of existing tools from previous project partners were browsed. Eight prominent LA tools were then randomly selected to be used for the evaluation of the framework: Blackboard Learn 9.1 Retention Centre, CourseSignals, EnquiryBlogger, the LeMo project, SNAPP, StepUp!, Student Activity Meter, and Student Explorer. The study was conducted with members from the LACE project consortium and its associated partners. Each of the eight participants evaluated two of the eight tools, which in turn meant that each of the eight tools was evaluated twice.

Due to the nature of the study, i.e. the evaluation of the framework of quality indicators for LA, outcomes dealing with individual tools are not addressed. Instead the focus is entirely on the setup and applicability of the framework’s criteria and quality indicators.

2.2 Quantitative Results

To get an overview of the results for all QIs Table 1 shows how many yes, no and not applicable every QI received. The highest scoring instance for yes, no and not applicable are highlighted. Table 2 summarises the rating values of all QIs and also lists their average rating. The highest average rating is achieved by the QI awareness, i.e. 4.3, while the lowest average is achieved by efficiency, i.e. 2.6. These two indicators are also the ones with the lowest (awareness) and highest (efficiency) non-applicability.

The data shows that the QIs of the first criterion, i.e. Objectives, are often present in/supported by the tools analysed. Also, the amount of non-applicability of these indicators is rather low compared to that of other criteria. The QI of awareness has the highest score of yes, followed closely by that of behavioural change. Non-applicability of QIs is quite low in this criterion which in reverse means that they are applicable and thus suitable indicators when evaluation LA tools. Motivation seems to be the most controversial QI as it has the most diverse results. Looking at the ratings for the Objectives criterion this view is supported as most study participants found it easy or very easy to judge the QIs of this criterion.

The non-applicability of the QIs in the criterion Learning Support is similarly low as that of the Objectives criterion. Although they are applicable, however, they are not present in/supported by the tools as often as the QIs of the first criterion. Especially recommendation and activity classification seem not to be as common in LA tools. The ratings for the QIs in the Learning Support criterion are not as ten-dentious as the previous ones. There are still many easy and very easy ratings. However, the number of difficult and very difficult ratings is notably higher. Especially activity classification was deemed a difficult to evaluate QI by the study participants.

Looking at the ratings for the QIs in the Learning Measures and Output criterion we can see that they are almost evenly spread over the scale. No clear tendency of either difficulty or ease can be identified. Also, the non-applicability of the QIs is quite a bit higher than that in the first two criteria. In 50% of the cases efficiency was not applicable while effectiveness was not applicable in 38% of the analysed cases. All QIs in this criterion, however, have rather low (or even none) no values. It thus seems that indicators here tend to be either present in/supported by the LA tools or not applicable.
Table 1: Presence (yes/no) or non-applicability of quality indicators in a tool

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Yes</th>
<th>No</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>awareness</td>
<td>15</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>reflection</td>
<td>12</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>motivation</td>
<td>9</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>behavioural change</td>
<td>14</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>perceived usefulness</td>
<td>14</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>recommendation</td>
<td>8</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>activity classification</td>
<td>6</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>det. of students at risk</td>
<td>12</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>comparability</td>
<td>12</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>effectiveness</td>
<td>9</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>efficiency</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>helpfulness</td>
<td>14</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>transparency</td>
<td>9</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>data standards</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>data ownership</td>
<td>1</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>privacy</td>
<td>9</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>availability</td>
<td>7</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>implementation</td>
<td>6</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>training of stakeholders</td>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>organisational change</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

The QIs with the most no values are those of the Data Aspects criterion, i.e. they are often not present in supported by the analysed LA tools. Non-applicability is on a medium level of about a third for this criterion while the yes values vary from low to medium levels. Study participants tended to be rather positively confident when rating the QIs of this criterion. Although there are hardly any very easy ratings, the number of easy ratings is quite high.

The most clear and obvious rating tendency was given to the QIs of the Organisational Aspects criterion. In three quarters of the cases the QIs have either been rated as easy or very easy to judge by the study participants. The non-applicability of the QIs is the highest for this criterion while yes and no values vary.

2.3 Qualitative Results

Apart from collecting quantitative feedback about the quality indicators, study participants were also offered the opportunity to describe the application of the QI and to add comments.

Generally, participants thought that it was rather easy to judge the QIs of the Objectives criterion. The resources they used to evaluate the tool often provided information about whether it supported awareness, reflection, motivation and behavioural change. One issue raised by participants was the distinction between a tool intending to foster something and actually being successful in doing so. Based on the fact that in many cases only the actual user of a tool can assess whether awareness, reflection, motivation or behavioural change was fostered, they suggest to ask whether a tool intends to do something when evaluating it. Another issue raised was that the main user type of a tool should be identified before evaluating it as some tools might cater to learners, other to teachers, etc. A third issue mentioned by the participants was that of direct or indirect fostering (or better the intention to do so) of the indicators.

For the QIs of the Learning Support criterion participants also stressed that taking the user type into account when evaluating a tool is important. Participants also mentioned that there are two types of indicators in this criterion. While usefulness can be deemed an intended goal of a tool, the QIs of recommendation, activity classification and detection of students at risk are features / functionalities of a tool. Although both types of indicators are valid to be used to evaluate a LA tool, an EF should benefit from using only one type of indicator per criterion. It was also noted that for some indicators it might not suffice to say whether a tool does something or not in order for it to be deemed a good tool, e.g. too many recommendations might be worse than no recommendations. The QI that caused most trouble to the study participants is that of activity classification. Participants found it rather difficult to judge this QI as they did not fully understand what it meant while participants of the GCM study most likely had a clear concept in mind, i.e. that LA tools "know" what their users are doing automatically. It was therefore suggested to rephrase or redefine the indicator.

The criterion Learning Measures and Output was an overall difficult one to judge for the participants. They not only had difficulties judging some of the indicators but that the criterion title increased this difficulty even more. They were unsure whether to relate an indicator to the measures or the output of a tool, to the processes or the tool itself and thus suggested to define a better, clearer name and concept for this criterion. The indicator comparability was quite difficult for participants to apply as they were not completely sure about what was to be comparable. In the GCM study the LA experts had identified comparability in relation to the measures and outcomes of an analytics tool, e.g. that effects of one tool could be compared to those of another tool. From the responses of this study's participants, however, it is clear that some participants assumed the comparability to be for users within one tool. This misunderstanding clearly needs to be addressed by a better definition of the indica-

Table 2: Overview of 1(very difficult)-to-5(very easy) scale ratings plus average rating for all QIs

<table>
<thead>
<tr>
<th>Indicator</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>awareness</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>reflection</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>motivation</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td>behavioural change</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>perceived usefulness</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>recommendation</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>activity classification</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>det. of students at risk</td>
<td>-</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>comparability</td>
<td>-</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>3.3</td>
</tr>
<tr>
<td>effectiveness</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2.8</td>
</tr>
<tr>
<td>efficiency</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>2.6</td>
</tr>
<tr>
<td>helpfulness</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>3.2</td>
</tr>
<tr>
<td>transparency</td>
<td>-</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>3.2</td>
</tr>
<tr>
<td>data standards</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3.3</td>
</tr>
<tr>
<td>data ownership</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2.9</td>
</tr>
<tr>
<td>privacy</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>3.7</td>
</tr>
<tr>
<td>availability</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>4.0</td>
</tr>
<tr>
<td>implementation</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>3.9</td>
</tr>
<tr>
<td>training of stakeholders</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>3.9</td>
</tr>
<tr>
<td>organisational change</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>12</td>
<td>1</td>
<td>3.6</td>
</tr>
</tbody>
</table>
tor and possibly a rephrasing. For the indicators effectiveness and efficiency it was suggested to distinguish between the intention of a tool and the fulfillment of that intention. Also, participants would have liked to see clear definitions in order to better distinguish them from another. They also suggested to clearly indicate the type of user of a tool, too, when applying these indicators. The same applies to the indicator helpfulness. They also suggested to clearly distinguish this QI from the one about perceived usefulness by giving a clear definition to both.

In the GCM study the LA experts had identified the QIs dealing with Data Aspects as the most important and as the most feasible ones. This time, however, the QIs of this criterion were often either not supported by a tool or not applicable. The main reason given for either saying no or not applicable was that they had not used the tool themselves but had to rely on the resources describing the tools. It was thus suggested to add an I don’t know-option. Here, the indicator data ownership was deemed the most difficult to rate. Some participants were not able to fully grasp and apply the concept to a given tool and therefore suggested a more detailed definition of the indicator. Again, they would have liked to see the type of user in focus mentioned when doing the evaluation. For the QIs transparency and privacy the issue of differentiation and a clearer definition was raised. It was also mentioned that in the case of transparency, two types could be present in a tool: a tool supports transparency if users know what data about them is collected and stored but also if one user can see information about other users.

The criterion on Organisational Aspects was by far the easiest to rate for the participants. It is also the one with the most not applicable values. Many participants reported that this was due to many of the tools being prototypical implementations that had only been used within one course or as a small test bed study. Another reason given was the lack of information provided by the resources, used for the evaluation about anything related to Organisational Aspects and not being able to use the tool. The difference between the QIs of availability and implementation was not clear to a number of participants. They thus suggested to either define the QIs more clearly or merge them into one.

3. WORKING TOWARDS AN IMPROVED EVALUATION FRAMEWORK VERSION

The results of this framework evaluation study allow us to identify several issues with the framework that need to be addressed in order to work towards an improved EF for the next evaluation cycle. Some of them might be cleared up fairly easily while others will need to be examined carefully so as to ensure actual improvement. The issues identified can be divided into the following categories: (1) concept definitions, (2) differentiations, (3) framework structure, and (4) questionnaire adaption.

The first category, concept definitions, relates to any case where it was expressed that either a criterion or a QI needs to be rephrased or defined more clearly in order to be properly applied to a tool evaluation. One criterion and three QIs where this is the case were particularly mentioned: Learning Measures and Output, activity classification, comparability, and data ownership. Renaming, and thus redefining, a whole criterion also influences how the QIs of that criterion are interpreted. When constructing the next version of the framework, this will have to be taken into account. Although only these four elements of the framework have explicitly been mentioned, all other criteria and QIs should also undergo a careful definition inspection and pilot testing to avoid similar issues in the next evaluation cycle.

The issues of the second category, differentiations, are closely related to those of the first category. Participants identified some QIs, or better pairs of QIs, that needed to be defined more clearly and supported by some distinct example so as to be able to properly distinguish between them. Otherwise users of the framework might misunderstand them and thus distort the results of a tool evaluation. The QIs mentioned by the study participants are usefulness vs. helpfulness, effectiveness vs. efficiency, transparency vs. privacy, and availability vs. implementation.

The third category, framework structure, deals with the issue of inter-criterion homogeneity of QI types. It was suggested to ensure that the types of indicators within one criterion are the same in order to improve the applicability of the whole criterion. Generally, indicators should tend to be concept rather than feature driven. Participants identified this issue in the criterion Learning Support but all other criteria should be inspected as well so as to avoid this issue from appearing again in the next evaluation cycle.

The fourth category, questionnaire adaption, comprises issues that need to be addressed when setting up the next version of the framework’s questionnaire or better the next practical, applicable and executable version of the framework (as this might not be the same questionnaire format as it was this time). Several aspects were noted that would highly improve the applicability of the indicators. For many QIs the answers would differ depending on the user type addressed. This should thus be clarified for each questionnaire, possibly leading to specific instances of the EF being needed for different stakeholders. Questions for the QIs should best ask about the intention of a tool as this is something that can be answered much more easily than a tool’s actual impact on a user. This is especially true if the evaluator has no access to the tool but has to work with descriptive resources. The third issue related to questionnaire adaption is the possible addition of answer options. Several participants of the evaluation study remarked that they would have liked to see an I don’t know-option as too much-option as information for some QIs might be too sparse.

An issue that is not related to any of the categories and that cannot be improved by us is the sparsity of information provided in the resources about LA tools. While addressing the issues mentioned above will make it easier for externals to evaluate a tool, the most complete evaluations will be those of the actual users or creators of a tool. In those cases where users or creators apply the framework to their own tool, however, the results might be biased which has to be taken into account as well.

4. CONCLUSION

In this paper we first presented the findings of a group concept mapping study to empirically identify criteria and quality indicators for LA tools to form a framework. We then conducted a second study with members of the LACE project to apply the framework to a number of tools in order to evaluate it. With the feedback from the participants we were able to identify problematic issues and have collected suggestions how to overcome the issues and improve
the framework.

Figure 2 shows which criteria and quality indicators have been identified with category 1 (solid), category 2 (dashed) or no issues (dotted). The outcomes of the evaluation study will be carefully analysed and discussed within the LACE consortium to develop an improved version of the framework. Apart from the theoretical framework set up, the structure of the related evaluation instrument will also be improved as different stakeholders might require different versions of the instrument. The improved framework as well as its implementation will then form the basis of another evaluation cycle. The results of the tool analyses of these studies will be fed into an Evidence Hub, a knowledge base of evidence created and curated by LACE that captures evidence for the effectiveness and the relative desirability of the outcomes resulting from use of various tools and techniques.

5. ACKNOWLEDGMENTS

This work was partly funded by the LACE project (GA No. 619424) under the FP7 programme of the European Commission.

6. REFERENCES


9http://evidence.laceproject.eu
Figure 2: Indicators and criteria with category 1 (solid) and category 2 (dashed) and no issues (dotted)