The Proof of the Pudding: Examining Validity and Reliability of the Evaluation Framework for Learning Analytics

Maren Scheffel[1], Hendrik Drachsler[1,2,3], Christian Toisoul[1], Stefaan Ternier[1], and Marcus Specht[1]

[1] Open Universiteit, Valkenburgerweg 177, 6419AT Heerlen, NL
maren.scheffel@ou.nl, hendrik.drachsler@ou.nl, christian.toisoul@ou.nl, stefaan.ternier@ou.nl, marcus.specht@ou.nl
[2] Goethe University Frankfurt, Germany
[3] German Institute for International Educational Research (DIPF), Germany

Abstract. While learning analytics (LA) is maturing from being a trend to being part of the institutional toolbox, the need for more empirical evidences about the effects for LA on the actual stakeholders, i.e. learners and teachers, is increasing. Within this paper we report about a further evaluation iteration of the Evaluation Framework for Learning Analytics (EFLA) that provides an efficient and effective measure to get insights into the application of LA in educational institutes. For this empirical study we have thus developed and implemented several LA widgets into a MOOC platform's dashboard and evaluated these widgets using the EFLA as well as the framework itself using principal component and reliability analysis. The results show that the EFLA is able to measure differences between widget versions. Furthermore, they indicate that the framework is highly reliable after slightly adapting its dimensions.

Keywords: learning analytics, evaluation, validity, reliability

1 Introduction

By using learning analytics (LA), i.e. by measuring, collecting, analysing and reporting the learners' data from a course in a useful and meaningful way, awareness and reflection about the learning and teaching processes can be stimulated [14, 11]. During the last few years the amount of LA-related research, publications and events has increased steadily [9]. Learning analytics, however, is not to be seen as pure 'number-crunching' on a strictly institutional level or as only being used to improve retention. Instead, it is about creating a holistic view on all learning and teaching processes involved [10]. Therefore, as LA should stimulate the self-regulating skills of the learners [16] and foster awareness and reflection processes for learners and teachers, it is recognised that a good way to present LA to users is through a visual representation [22]. Kim, Jo and Park [13] indicate that learners' achievement could be increased by allowing them access to a learning analytics dashboard, i.e. a collection of visualisations. They also point

out that LA visualisations should be carefully designed if interest in and usage of the dashboard and analytics is to be maintained by the main stakeholders, i.e. learners and teachers.

With the need for empirical studies growing and more and more discussions about the effect of learning analytics coming up [21, 8], a number of studies investigating the impact of LA dashboards have been published in the last few years. Lonn et al. [15] for example have shown that seeing their academic performance in a LA applications could affect students' interpretation of their data and their success. They stress that LA interventions need to be designed carefully with student goal perception in mind. Beheshita et al. [2] randomly assigned LA visualisations to students of a blended learning course and showed that it depended on the students' achievement goal orientation whether the effect of the visualisations on learning progress was positive or negative. They stress that students' achievement goal orientation and other individual differences need to be taken into account during the LA design process. Finally, Khan and Pardo [12] showed that depending on the students' information needs and the types of learning activities different kinds of LA dashboards and visualisations are needed for them to be effective. From all three studies it is thus clear that LA visualisations need to be embedded into the instructional design to have a positive effect.

An important aspect that thus needs to be kept in mind when using LA to address issues such as the ones mentioned above is the following: How can we make sure that the learning analytics are valid, reliable, understandable and supportive for the involved stakeholders? We have thus developed an evaluation instrument that allows a standardised approach to the evaluation of LA tools: the Evaluation Framework for Learning Analytics (EFLA) [19, 17]. The framework consists of four dimensions (Data, Awareness, Reflection, Impact) for learners and teachers.

Taking all of this into account, we designed and developed new versions for two widgets from the LA dashboard of the ECO MOOC platform and investigated in a lab experiment whether the current structure of the EFLA appropriately reflects the questionnaire's underlying components and whether the evaluation instrument can be used to measure changes between different versions of widgets. The lab setting was chosen as low numbers of teachers in the ECO environment would not give us sufficient input from that stakeholder group and because it allowed for a controlled experimental setting. We conducted our study with the following research questions in mind:

(RQ-A) Can the EFLA measure differences between iterations of a widget?
(RQ-B1) Do the four current EFLA dimensions validly represent the underlying components?
(RQ-B2) Do the items within the dimensions reliably measure the underlying component?

The next section describes the ECO platform's widgets and the evaluation instrument and elaborates on the method of analysis. After the presentation of results, the discussion section sets the results in relation to the research questions while the final section concludes the paper.

2 Method

2.1 Participants

Fifteen PhD candidates (eight women and seven men) and fifteen assistant, associate or full professors (seven women and eight men) from the Faculty of Psychology and Education of the Open University of the Netherlands voluntarily participated in the experiment. The PhD candidates were assigned the role of students while the post-docs were assigned the role of teachers during the experiment. All participants had at least basic knowledge about what LA is. Informed consent was obtained from all participants.

2.2 Materials

The Learning Analytics Widgets. Massive Open Online Courses (MOOCs) have the potential to provide education at a low cost for a wide and diverse public [6]. The European project ECO (Elearning Communication Open-Data)[4] has therefore created a platform that gives free access to MOOCs based on Open Educational Resources. A learning analytics dashboard containing several visualisations is part of the ECO platform to support the ECO users. The visualisations are based on interaction data of the users with the platform and with the MOOCs, e.g. launching a course, accessing pages, watching videos, posting in a forum, uploading homework, etc. All users of the portal, i.e. the students as well as the teachers of the MOOCs, see the same visualisations.

Two of the existing ECO LA visualisations were chosen for the experiment: the Activity Widget and the Resources Widget. The Activity Widget shows how active the learners are in a MOOC according to the number of actions done in that MOOC. The Resources Widget shows what types of resources are present in this course and how often all users together have accessed the various resources in the MOOC (see Appendix A at *bit.ly/EFLApudding* for the screenshots and more detailed descriptions of all widget versions).

The second version of the Activity Widget again shows the total activity per user. Additionally a user's own position is highlighted. Users can choose between two types of clustering: the Median with quartiles and an artificial intelligence algorithm. Both create four clusters in reference to Cobo et al.'s four activity types [5]. In order to protect the users' privacy, none of the users are able to identify who the other users are in the visualisation as the ECO LA dashboard does not distinguish between students and teachers of the course. The updated version of the Resources Widget compares a user's MOOC path with the ideal path of the course and the paths of other participants. A user can see which activities he has accessed and which ones not. Teachers could use this tool to identify if learners are using the MOOC as planned by discovering if activities are accessed too early, too late, or not at all. Students could compare themselves to other users and to the model line. Again, in order to protect the users' privacy, none of the other users are identifiable.

[4] https://ecolearning.eu

Table 1. Dimensions and items of the learner and the teacher section of the EFLA.

		EFLA items for learners/teachers
Data:	**D1**	For this LA tool it is clear what data is being collected.
	D2	For this LA tool it is clear why the data is being collected.
	D3	For this LA tool it is clear who has access to the data.
Awareness:	**A1**	This LA tool makes me aware of my/my students' current learning situation.
	A2	This LA tool makes me forecast my/my students' possible future learning situation given my/their (un)changed behaviour.
Reflection:	**R1**	This LA tool stimulates me to reflect on my past learning/teaching behaviour.
	R2	This LA tool stimulates me to adapt my learning/teaching behaviour if necessary.
Impact:	**I1**	This LA tool increases my motivation to study/teach.
	I2	This LA tool stimulates me to study/teach more efficiently.
	I3	This LA tool stimulates me to study/teach more effectively.

The Evaluation Framework. An institution's need for reflection on how ready they are to implement LA solutions is addressed by the Learning Analytics Readiness Instrument (LARI) [1]. While LARI has been shown to be an effective instrument to evaluate institutional readiness, there is no standardised instrument so far to evaluate the LA tools once implemented. However, more and more LA tools are being designed, developed and implemented. In order to close this gap, we have therefore developed the Evaluation Framework for Learning Analytics (EFLA). Inspired by the System Usability Scale (SUS), a "reliable, low-cost usability scale that can be used for global assessments of system usability" [3], the EFLA aims to provide similar facilities for the LA domain. Using the subjective assessments by their users is a quick and simple way to get a general indication of the overall quality of a tool in comparison to other tools or other versions of the same tool as Brooke [3] points out.

The first version was constructed through a group concept mapping (GCM) study with experts from the LA community and consisted of five dimensions (Objectives, Learning Support, Learning Measures and Output, Data Aspects, and Organisational Aspects) with four items each [19]. After a small evaluation study with LA experts [17] as well as a revisit of the GCM data and a thorough look at related literature, the second EFLA version was developed. Split into two parts, one for learners and one for teachers, the framework now consisted of four dimensions (Data, Awareness, Reflection and Impact) with three items each. This version was turned into an applicable tool, i.e. a questionnaire for students and teachers, and then used in an online course [18]. Based on a subsequent evaluation of the EFLA-2, the third version was created. While the dimensions stayed the same, the items were slightly reduced and further refined. Table 1 shows version 3 of the EFLA that was used in this study. All items are rated on a scale from 1 for no agreement to 10 for high agreement.

2.3 Procedure

All participants were invited to an individual face-to-face session for the experiment. At the beginning of each session, every participant received an introduction to the experiment and was asked to give their informed consent to take part in the study. Following an experimental script, each participant first received some introductory information about the ECO platform and its LA dashboard before getting detailed explanations about the four LA widgets while being shown a screenshot of the corresponding widget. For the two updated widget versions a live demo was also provided. After each widget explanation, participants were asked to evaluate the widget using the EFLA while assuming either the role of a student (all PhD candidates) or a teacher (all post docs). At the end of each EFLA survey participants had the option to add comments. When all four widgets had been evaluated, participants were asked to supply some demographic information (gender and age range) and were given a final opportunity to enter comments about the experiment. Once all data was collected from the participants, several statistical analyses were calculated using IBM's SPSS Statistics and graphs showing the average evaluation of each EFLA item for the different widgets from both stakeholder groups were created. The statistical analyses included t-tests for the widget evaluation and principal component analysis as well as reliability analysis for the EFLA evaluation.

3 Results

3.1 Widget Evaluation

Figure 1 shows the average scores of the ten EFLA items from students and teachers for both versions of the widgets. On average students and teachers gave better ratings to the second versions of both widgets. The only item students rated lower in an updated widget version is D1 for the Resources Widget. The items that teachers rated lower in an updated widget version are D3 and R2 for the Activity Widget and also D1 for the Resources Widget. While the original versions of the widgets received higher ratings from the teachers, the updated widget versions received higher ratings from the students.

Conducting paired sample t-tests for the ten EFLA items allowed us to see whether the differences in ratings between the two versions of the widgets were significant or not (see Appendix B at *bit.ly/EFLApudding* for detailed results tables). For the student participants there are several EFLA items where the difference between the ratings of the widgets' two versions is significant. The second version of the Activity Widget received significantly higher ratings for the items A1 ($p = .019$), R1 ($p = .044$), R2 ($p = .008$) and I2 ($p = .022$) while the Resources Widget received significantly higher ratings for all items (p ranges between .000 and .048) except D1. In case of the teachers, each widget only has one item where the difference between the two versions is significant: for item I2 of the Activity Widget $t(14) = -2.942, p = .011$ and for item A2 of the Resources Widget $t(14) = -2.839, p = .013$.

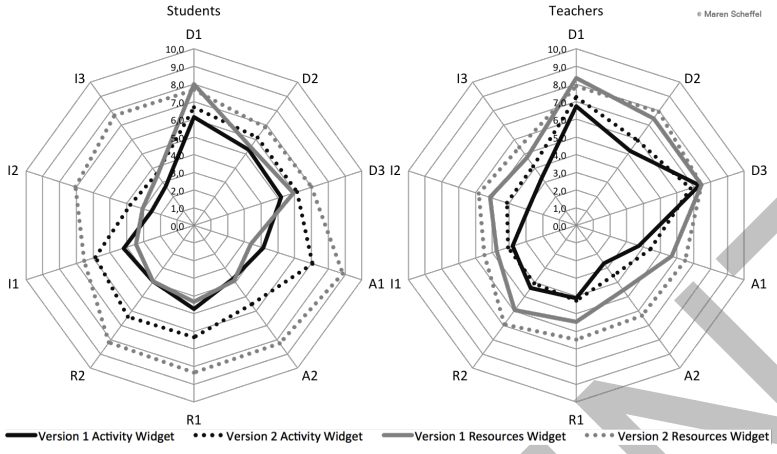

Fig. 1. Average scores of the EFLA items for students (left) and teachers (right) for both versions of both widgets.

Table 2. Descriptive statistics of all EFLA items from all widgets combined for students (left) and teachers (right).

	students						teachers					
	N	Min.	Max.	Mean	St.D.	Var.	N	Min.	Max.	Mean	St.D.	Var.
D1	60	1	10	7.12	2.450	6.003	60	3	10	7.55	1.908	3.642
D2	60	1	10	5.93	2.968	8.809	60	2	10	6.63	2.091	4.372
D3	60	1	10	6.07	3.194	10.199	60	2	10	7.27	3.162	9.995
A1	60	1	10	5.87	3.105	9.643	60	1	10	5.07	2.642	6.979
A2	60	1	10	5.35	2.839	8.062	60	1	9	4.27	2.421	5.860
R1	60	1	10	5.93	2.711	7.351	60	1	10	5.08	2.438	5.942
R2	60	1	10	5.62	2.853	8.139	60	1	9	5.33	2.319	5.379
I1	60	1	10	5.02	2.902	8.423	60	1	8	4.52	2.259	5.101
I2	60	1	10	4.12	2.811	7.901	60	1	10	4.48	2.411	5.813
I3	60	1	10	4.38	2.946	8.681	60	1	9	4.42	2.309	5.332

3.2 EFLA Evaluation

Every participant completed the EFLA survey for both versions of the two LA widgets which gives us a total N of 120 for each EFLA item (60 per stakeholder group, 30 per widget, 15 per widget version). All statistical analyses were conducted separately for the students' and teachers' data due to the different semantics, i.e. different wording leading to different meaning, of the ten EFLA items. The highest N within one analysis is thus 60.

Table 2 shows the descriptive statistics, i.e. N, minimum value, maximum value, mean, standard deviation and variance, for all ten EFLA items for the students (left) and the teachers (right). Two values seem to be slightly different from the rest: the variance of EFLA item D3 for students as well as for teachers is noticeably higher than all other variance values.

First Analysis. Before conducting the principal component analysis (PCA) we first looked at the factorability of the ten EFLA items for students and teachers. For the students' EFLA only few correlations were below .3 and all ten items correlated at least .6 with at least two other items. Additionally, the Kaiser-Meyer-Olkin measure of sampling adequacy was .836, i.e. above the recommended value of .6, and Bartlett's test of sphericity was $\chi^2(45) = 462.515, p < .000$. All diagonals of the anti-image correlation matrix were above .7. For the teachers' EFLA there were also few correlations below .3 and nine items correlated at least .4 with at least two other items (only D3 did not). Additionally, the Kaiser-Meyer-Olkin measure of sampling adequacy was .848, i.e above the recommended value of .6, and Bartlett's test of sphericity was $\chi^2(45) = 405.841, p < .000$. Nine diagonals of the anti-image correlation matrix were above .7 (except D3 where it was .486). Due to these results, none of the items were discarded at this point and we continued with the PCA using Varimax rotation in order to identify the factors underlying the EFLA. As we had structured the EFLA with four dimensions in mind (Data, Awareness, Reflection, Impact), the solution with four components was examined first, followed by those with three and with two components (see Appendix C at *bit.ly/EFLApudding* for details of all analyses).

First Principal Component Analysis – Students. For the students' four-components solution all communalities were above .8 except I1 which was .749. Together the four components explained 85.824% of the variance (80.805 for the three components with primary loadings). All items in the four-components solution (rotated matrix) had a primary loading of .6 or above. However, only three of the four components contained primary loads. Component 1 was clearly formed by items I1, I2 and I3, component 2 consisted of items A1, A2 and R1 and component 3 was clearly formed by items D1, D2 and D3. Item R2 had two possible primary loads (.636 and .634) and could be part of either component 1 or component 2. Looking at the three-components solution for the students' data, the communalities were all above .736. The three components cumulatively explained 81.427% of the variance. Also, the distinction between the components was clearer than in the four-components solution: component 1 contained items I1, I2 and I3, component 2 contained items A1, A2, R1 and R3 and component 3 contained items D1, D2 and D3. Again all items had a primary loading of .6 or above. The two-components solution for the students' data had communality values above .7 except for A2 (.660) and I1 (.672). Cumulatively the two components explained 75.238% of the variance. This solution had primary loadings for nine items above .8 and one item at .796 with component 1 containing A1, A2, R1, R2, I1, I2 and I3 and component 2 containing the items D1, D2 and D3.

To sum up, the three-components solution seems to be the best result as all components contain primary loads (the four-components solution does not) and as it explains more variance than the two-components solution.

First Principal Component Analysis – Teachers. The PCA of the teachers' data provided somewhat less clearly structured solutions. In the four-components solution all communalities were above .7. Together the four components explained 83.866% of the variance. All items had a primary loading of at least .6. Component 1 contained items R1, R2, I1, I2 and I3, while component 2

contained items D2, A1 and A2. Items D1 and D3 each formed their own component. The Data items thus did not form one component but were spread over three different ones. The three-components solution for the teachers' data had communality values of at least .7 for all values except for D2 (.589) and I3 (.691). Cumulatively 77.409% of variance were explained by the three components. This solution had one clear component containing items R1, R2, I1, I2 and I3 with all primary loadings above .7. D1, D2 and A1 formed one component, as did D3 and A2, all with primary loadings above .5. Both A1 and A2, however, had rather high cross-loads: while A1 had a primary load of .677 in component 2 (together with D1 and D2) it had a cross-load of .580 for component 3 (where it would join A2 and D3). A2 (primary load of .586) on the other hand also had a high cross-load of .551 in component 1 (where it would join R1, R2, I1, I2 and I3). Finally, in the two-components solution for the teachers' data, the communalities were above .6 except for D1 (.489), D2 (.526) and A1 (515). The two components explained 68.146% of the variance. Component 1 contained D2, A1, A2, R1, R2, I1, I2 and I3 (all with primary loads above .6), while the second component was comprised of items D1 and D3. Again, the Data items did not form one clear component. Item D1 (primary load of .503 in component 2), however, had a rather high cross-load of .486 in component 1 and could thus possibly be positioned there leaving D3 to form its own component.

To sum up, the three-components solution seems to be the best result as all components have at least two primary loads (the four-components solution does not) and as it explains more variance than the two-components solution.

First Reliability Analysis. In order to see how reliable the scales are and to check whether any of the items should be excluded, we calculated the reliability values, i.e. Cronbach's Alpha, for several item combinations based on the PCA results: the four EFLA dimensions Data, Awareness, Reflection and Impact individually (D,A,R,I), the combination of the Awareness and Reflection items (A+R), the combination of the Awareness, Reflection and Impact items (A+R+I), and the combination of the Reflection and Impact items (R+I). Only one scale, i.e. the teachers' three Data items on their own, received a low reliability score (.397). All other scales had a reliability score of .8 or higher. For two scales a substantial increase ($> .05$) in Cronbach's Alpha could be achieved by eliminating an item. For the students' EFLA eliminating item I1 in the Impact-items-only scale would result in a Cronbach's Alpha of .954 while an elimination of item D3 in the Data-items-only scale of the teachers' EFLA would result in a Cronbach's Alpha of .574.

As the items D3 and I1 seemed to cause problems and hindered a clear component solution, we decided to delete them and to re-do the analysis with the remaining eight items D1, D2, A1, A2, R1, R2, I2 and I3.

Second Analysis. Before doing the PCA, we again looked at the factorability of the EFLA items. For the students' data there were again few correlations between the items that were below .3 and all items correlated at least .6 with at least one other item. The Kaiser-Meyer-Olkin measure of sampling adequacy was .799 (which is above the recommended value of .6) and Bartlett's test of sphericity was

Table 3. PCA using Varimax rotation for four, three and two components for students' EFLA (primary loads are light grey) and teachers' EFLA (primary loads are grey).

	four components								three components						two components			
	students				teachers				students			teachers			students		teachers	
	1	2	3	4	1	2	3	4	1	2	3	1	2	3	1	2	1	2
D1	.070	.048	.936	.301	.210	.120	.934	.176	.077	.008	.903	.171	.100	.928	.054	.904	.103	.940
D2	.162	.151	.377	.878	.292	.205	.199	.895	.156	.184	.862	.451	.262	.524	.226	.864	.466	.572
A1	.840	.300	.049	.220	.220	.881	.297	.135	.845	.289	.184	.202	.878	.329	.839	.197	.566	.443
A2	.849	.266	-.013	.157	.481	.767	-.120	.216	.853	.254	.096	.506	.786	-.029	.824	.109	.823	.091
R1	.780	.436	.246	-.153	.874	.218	.067	.240	.798	.386	.087	.893	.236	.144	.863	.100	.876	.213
R2	.717	.540	.063	.123	.834	.226	.152	.348	.731	.523	.131	.869	.247	.265	.896	.142	.849	.333
I2	.380	.891	.052	.112	.869	.230	.167	.199	.405	.881	.115	.872	.241	.221	.864	.121	.852	.289
I3	.419	.863	.049	.126	.785	.339	.246	-.001	.443	.853	.122	.741	.332	.220	.876	.129	.783	.293

$\chi^2(28) = 359.650, p < .000$. All diagonals of the anti-image correlation matrix were above .7 (except for D1 which was .526). The teachers' data also showed few correlations below .3 and, except for D1 and D2 which correlated at .4 with three other items, all other items correlated at .6 with at least one other item. Additionally, the Kaiser-Meyer-Olkin measure of sampling adequacy was .826 and Bartlett's test of sphericity was $\chi^2(28) = 338.879, p < .000$. All diagonals of the anti-image correlation matrix were above .7.

Second Principal Component Analysis – Students. Table 3 shows the results of the PCA using Varimax rotation for these different settings. For the students' four-components solution all communalities were above .8. Together the four components explained 89.975% of the variance. All items in the four-components solution had a primary loading of .7 or above. Component 1 was clearly formed by items A1, A2, R1 and R2, component 2 consisted of items I2 and I3, component 3 only contained D1 and component 4 only contained D2. Looking at the three-components solution for the students' data, the communalities were all above .793. The three components cumulatively explained 84.559% of the variance. Again, component 1 was clearly formed by items A1, A2, R1 and R2 and component 2 consisted of items I2 and I3. Component 3 was made up of D1 and D2. All primary loadings were above .7. The two-components solution for the students' data had communality values above .7 except for A2 (.691). Cumulatively the two components explained 77.195% of the variance. This solution had primary loadings for all items above .8 with component 1 containing A1, A2, R1, R2, I2 and I3 and component 2 containing the items D1 and D2.

To sum up, the three-components solution seems to be the best result as all components have at least two primary loads (the four-components solution does not) and as it explains more variance than the two-components solution.

Second Principal Component Analysis – Teachers. The PCA of the teachers' data provided the following results. In the four-components solution all communalities were above .792. Together the four components explained 89.644% of the variance. All items had a primary loading of at least .7. Component 1 contained items R1, R2, I2 and I3, while component 2 contained items

Table 4. Reliability statistics and scale statistics of different item groups for students' EFLA (left) and teachers' EFLA (right)

	s t u d e n t s					**t e a c h e r s**				
items	N	Cron.α	Mean	Var.	St.D.	N	Cron.α	Mean	Var.	St.D.
D	2	.745	13.05	23.608	4.859	2	.574	14.18	11.237	3.352
A	2	.852	11.22	30.851	5.554	2	.814	9.33	21.650	4.653
R	2	.890	11.55	27.913	5.283	2	.945	10.42	21.468	4.633
I	2	.954	8.50	31.712	5.631	2	.881	8.90	19.922	4.463
A+R	4	.916	22.77	105.945	10.293	4	.870	19.75	69.513	8.337
A+R+I	6	.936	31.27	226.029	15.034	6	.916	28.65	149.214	12.215
R+I	4	.925	20.05	104.794	10.237	4	.935	19.32	75.135	8.668

A1 and A2. Items D1 and D2 each formed their own component. The three-components solution for the teachers' data had communality values of at least .7 for all items except for D2 (.547). Cumulatively 82.201% of variance were explained by the three components. This solution had one clear component containing items R1, R2, I2 and I3 with all primary loadings above .7. A1 and A2 formed component 2, and D1 and D2 formed component 3, all with primary loadings above .7 except for D2 (.524). Finally, in the two-components solution for the teachers' data, the communalities were either just below or well above .7 except for D2 (.545) and A1 (.517). The two components explained 72.445% of the variance. Component 1 contained items A1, A2, R1, R2, I2 and I3, all with primary loads above .7 except for A1 (.566), while the second component was comprised of items D1 (.940) and D2 (.572).

To sum up, the three-components solution seems to be the best result as all components have at least two primary loads (the four-components solution does not) and as it explains more variance than the two-components solution.

Second Reliability Analysis. Again, we calculated reliability values, i.e. Cronbach's Alpha, for several item combinations: the four EFLA dimensions Data, Awareness, Reflection and Impact individually (D,A,R,I), the combination of the Awareness and Reflection items (A+R), the combination of the Awareness, Reflection and Impact items (A+R+I), and the combination of the Reflection and Impact items (R+I). Table 4 gives an overview of these analyses for the students' as well as the teachers' EFLA. Only one scale, i.e. the teachers' Data items on their own, receives a noticeably lower reliability score (.574). All other scales have a reliability score of .7 or higher. For none of the scales a substantial increase ($> .05$) in Cronbach's Alpha could be achieved by eliminating an item.

4 Discussion

4.1 Widget Evaluation

The evaluation of the widgets using the EFLA questionnaire shows that there are indeed significant differences in evaluation results between the different widget versions. RQ-A can thus be answered with "yes". However, the differences

are not significant for all items of all widgets from both stakeholders. Students really seemed to appreciate the second versions of the widgets much more than the first versions. Especially the Resources Widget received significantly higher evaluation results for its second version. Taking into account the open comments from the questionnaire as well as the questions and comments uttered during the experiment by both stakeholder groups, these results are not really surprising. The teacher participants were much more hesitant and held back by the lab setting of the experiment while the student participants could easily put themselves in the mindset of an online course participant. Another factor that is likely to play a role in influencing the teachers' widget evaluations is that due to the ECO platform's not distinguishing between the user types of learners and teachers, the personalisation aspect of the widgets' second versions was rather pointless for the teachers. That is, they might feel disregarded.

4.2 EFLA Evaluation

Although none of the items were discarded before conducting the first PCA, the descriptive statistics (variance) as well as the factorability check (correlations and anti-image correlations for the teachers' data) hinted at possible issues with item D3. We began the first PCA assuming that EFLA consisted of four distinct dimensions. For the students' data, however, only three components had primary loadings in the four-components solution thus indicating that there are only three underlying components to EFLA. This was also supported by the other two solutions (the variance explained was higher for the three-components solution compared to the two-components solution).

The first analysis of the teachers' data also showed that a four-components solution did not best represent the data. It also became apparent that D1, D2 and D3 and to some extent A1 and A2 seemed to be problematic for the teachers. Their PCA results for those items were much less clear than those of the students. This had already been foreshadowed during the experiment. The teacher participants asked considerably more questions than the student participants and voiced uncertainty about how to answer some of the questions. This insecurity about the items is likely to be reflected in their answers resulting in partially inconclusive PCA results. The students did not seem to have such issues with the items and their results are thus more confident and possibly more credible.

The reliability analysis confirmed that several items might hinder a clear component solution. Two items, D3 and I1, had to be discarded. The fact that it was precisely those two items that were problematic is reasonable if we look at the actual questions behind those items. D3 says "For this LA tool it is clear who has access to the data". In comparison to this item, D1 and D2 much more clearly address the micro level of the immediately involved learners and teachers themselves [11] which is what EFLA is about. Both of those items are much more connected to the user's personal point of view whereas D3 could be (mis)interpreted so as to cover the whole learning environment instead of an individual LA tool despite the statement saying "For this LA tool...". Additionally, in order to interpret a visualisation it is important to know what data it is

based on and why (i.e. what the purpose is) but to know who else has access to the data does not affect the interpretation. Instead, it is more an issue of an institution's LA policy than an individual visualisation to make sure that privacy and transparency regulations are in place and transparently communicated.

Already during the experiment, student as well as teacher participants mentioned that they had difficulties answering item I1 due to its generality. The item says "This LA tool increases my motivation to study/teach". Whereas I2 and I3 cover the specific aspects of efficiency and effectiveness, item I1 covers motivation in general. Many participants said that their being motivated by a visualisation very much depended on the contents of the widget. For example, if a student sees that he is the lowest performing student, he might not be motivated to study by such a visualisation, while the opposite might be true if he sees himself in the top-performing group. On other days, the same student might feel very motivated to study when seeing that he is lagging behind. General motivation is thus too context-dependent to receive a reliable rating for one visualisation.

The second PCA without the two discarded items confirmed the previous indication that there are three underlying components for the EFLA items. In this solution each component was loaded by at least two items and explained more of the variance than the two-components solution. There is, however, a difference in how the items are spread across the components. For the students' data, D1 and D2 form one component, A1, A2, R1 and R2 form a second one and I2 and I3 form a third. The teachers' data resulted in one component containing D1 and D2, a second one containing A1 and A2 and another one containing R1, R2, I2 and I3. Even though some of the items of the student and teacher EFLA are semantically different, the two EFLA versions are still to be seen as two sides of the same coin.

Thus, in order to decide which of the three-components solutions to use for the next version of the EFLA, we took several aspects into account. First, the teacher participants of our study voiced more insecurities than the student participants did which leads us to put more confident in the students' results. Second, the reliability results for the students' data showed higher Cronbach's Alpha values than those of the teachers and the explained variance was higher for the students' three component solution. And third, supporting awareness and reflection processes in users in order to impact the learning or teaching processes is an important aim of LA. Awareness and reflection go hand in hand, with the former being a prerequisite of the latter [4, 7, 20].

Based on this, the new version of EFLA now consists of three dimensions: Data, Awareness & Reflection, Impact. The Data dimension contains items D1 and D2 and the Impact dimension contains items I2 and I3. Finally, the Awareness & Reflection dimension contains the four items A1, A2, R1 and R2 (see Appendix D at *bit.ly/EFLApudding* for the full framework structure).

RQ-B1 thus has to be answered with "no" as the assumed four-components structure did not turn out to be the best solution. However, the three-components solution we settled on does provide a fairly similar EFLA structuring to the one we envisioned as the items were not completely re-arranged within new clusters

but two of the original dimensions were combined into one. RQ-B2 also has to be answered with "no" as not all ten EFLA items turned out to reliably measure their component. However, eight of the items did and will thus constitute the new EFLA.

5 Conclusion

This paper presented the results of an empirical lab study where we developed and implemented several widgets for a MOOC platform's LA dashboard and evaluated them using the Evaluation Framework for Learning Analytics (EFLA). We also evaluated said framework using principal component analysis and reliability analysis. The results of the widget analysis showed that the EFLA can indeed be used to measure differences between different widget iterations. The results of the EFLA analysis show that there are three underlying dimensions in the EFLA instead of four and that not all items in version 3 of the EFLA reliably measured these dimensions. A new and improved fourth version of the EFLA has thus been created that can be used to validly and reliably evaluate LA tools. All items are to be rated on a scale from 1 for 'strongly disagree' to 10 for 'strongly agree'. In order to calculate a LA tool's EFLA score, i.e. a number between 0 and 100, the following steps are needed per stakeholder group: (1) calculate the average value for each item based on the answers given for that item, (2) calculate the average value for each dimension based on the average of its items, (3) calculate the dimensional scores by rounding the result of $((x-1)/9)*100$ where x is the average value of a dimension, and (4) calculate the overall EFLA score by taking the average of the three dimensional scores.

The learning analytics community now has the opportunity to verify the EFLA's applicability and benefit, i.e. the proof of the pudding is now in the eating. The framework has been published as open access and the framework's template flyer as well as an interactive spreadsheet to automatically calculate the EFLA scores and create visualisations of the scores are available for download via the LACE website at *http://www.laceproject.eu/evaluation-framework-for-la/*.

References

1. Arnold, K.E., Lonn, S., Pistilli, M.D.: An Exercise in Institutional Reflection: The Learning Analytics Readiness Instrument (LARI). In: Proc. of the 4th Int. Conf. on Learning Analytics and Knowledge. pp. 163–167. LAK '14, ACM, New York, NY, USA (2014)
2. Beheshitha, S., Hatala, M., Gašević, D., Joksimovic, S.: The Role of Achievement Goal Orientations When Studying Effect of Learning Analytics Visualizations. In: Proc. of the 6th Int. Conf. on Learning Analytics and Knowledge. pp. 54–63. LAK '16, ACM, New York, NY, USA (2016)
3. Brooke, J.: SUS: A quick and dirty usability scale. In: Jordan, P.W., Weerdmeester, B., Thomas, A., Mclelland, I.L. (eds.) Usability evaluation in industry. Taylor and Francis, London (1996)
4. Butler, D., Winne, P.: Feedback and self-regulated learning: a theoretical synthesis. Review of Educational Research 65(3), 245–281 (1995)

5. Cobo, A., Rocha, R., Rodriguez-Hoyos, C.: Evaluation of the Interactivity of Students in Virtual Learning Environments Using a Multicriteria Approach and Data Mining. Behaviour & Information Technology 33(10), 1000–1012 (2014)
6. Drachsler, H., Kalz, M.: The MOOC and Learning Analytics Innovation Cycle (MOLAC): A Reflective Summary of Ongoing Research and its Challenges. Journal of Computer Assisted Learning 32(3), 281–290 (2016)
7. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors 37(1), 32–64 (1995)
8. Ferguson, R., Clow, D.: Learning Analytics Community Exchange: Evidence Hub. In: Proc. of the 6th Int. Conf. on Learning Analytics and Knowledge. pp. 520–521. LAK '16, ACM, New York, NY, USA (2016)
9. Gašević, D., Dawson, S., Mirriahi, N., Long, P.: Learning Analytics – A Growing Field and Community Engagement. Journal of Learning Analytics 2(1), 1–6 (2015)
10. Gašević, D., Dawson, S., Siemens, G.: Let's not forget: Learning analytics are about learning. TechTrends 59(1), 64–71 (2015)
11. Greller, W., Drachsler, H.: Translating Learning into Numbers: A Generic Framework for Learning Analytics. Educational Technology & Society 15(3), 42–57 (2012)
12. Khan, I., Pardo, A.: Data2U: Scalable Real Time Student Feedback in Active Learning Environments. In: Proc. of the 6th Int. Conf. on Learning Analytics and Knowledge. pp. 249–253. LAK '16, ACM, New York, NY, USA (2016)
13. Kim, J., Jo, I.H., Park, Y.: Effects of Learning Analytics Dashboard: Analyzing the Relations Among Dashboard Utilization, Satisfaction, and Learning Achievement. Asia Pacific Education Review 17(1), 13–24 (2016)
14. Long, P., Siemens, G.: Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE Review 46(5), 31–40 (Sep 2011)
15. Lonn, S., Aguilar, S., Teasley, S.: Investigating Student Motivation in the Context of a Learning Analytics Intervention During a Summer Bridge Program. Computers in Human Behavior 47, 90–97 (2015)
16. Persico, D., Pozzi, F.: Informing learning design with learning analytics to improve teacher inquiry. British Journal of Educational Technology 46(2), 230–248 (2014)
17. Scheffel, M., Drachlser, H., Specht, M.: Developing an Evaluation Framework of Quality Indicators for Learning Analytics. In: Proc. of the 5th Int. Conf. on Learning Analytics and Knowledge. pp. 16–20. LAK '15, ACM, New York, NY, USA (2015)
18. Scheffel, M., Drachsler, H., Kreijns, K., de Kraker, J., Specht, M.: Widget, Widget As You Lead, I Am Performing Well Indeed!: Using Results from an Exploratory Offline Study to Inform an Empirical Online Study About a Learning Analytics Widget in a Collaborative Learning Environment. In: Proc. of the 7th Int. Conf. on Learning Analytics and Knowledge. pp. 289–298. LAK '17, ACM, New York, NY, USA (2017)
19. Scheffel, M., Drachsler, H., Stoyanov, S., Specht, M.: Quality Indicators for Learning Analytics. Educational Technology & Society 17(4), 117–132 (2014)
20. Schön, D.: The reflective practitioner: How professionals think in action. Temple Smith, London, UK (1983)
21. Siemens, G., Dawson, S., Lynch, G.: Improving the quality and productivity of the higher education sector – policy and strategy for system-level deployment of learning analytics. Discussion paper for the Australian Government, Society for Learning Analytics Research (SoLAR) (2013)
22. Verbert, K., Govaerts, S., Duval, E., Santos, J.L., Assche, F., Parra, G., Klerkx, J.: Learning Dashboards: An Overview and Future Research Opportunities. Personal Ubiquitous Computing 18(6), 1499–1514 (2014)