Evaluating Learning Analytics of Adaptive Learning Systems: A Work in Progress Systematic Review

Tobias Alexander Bang Tretow-Fish(✉) and Md. Saifuddin Khalid

Department of Applied Mathematics and Computer Science at the Technical University of Denmark, Kongens Lyngby, Denmark
compute@compute.dtu.dk
https://www.compute.dtu.dk/english

Abstract. There is currently no systematic overview of methods for evaluating Learning Analytics (LA) and Learning Analytics Dashboards (LAD) of Adaptive Learning Platforms (ALPs). 10 articles and 2 reviews are analyzed and synthesized. Focusing on the purposes of evaluation, methods used in the studies are grouped into five categories (C1-5): C1) evaluation of LA and LAD design and framework, C2) evaluation of performance with LA and LAD, C3) evaluation of adaptivity functions of the system, C4) evaluation of perceived value, and C5) Evaluation of pedagogical and didactic theory/context. While there is a relative high representation of evaluations in the C1-C4 categories of methods, which contribute to the design and development of the interaction and interface design features, the C5 category is not represented. The presence of pedagogical and didactical theory in the LA, LAD, and ALPs is lacking. Though traces of pedagogical theory is present none of the studies evaluates on its impact.

Keywords: Adaptive learning platforms · Learning analytics · Evaluation

1 Introduction

Adaptive Learning (AL) is not only a relatively new research area but also a multi-disciplinary field involving multiple synonymous and definitions. Adaptive learning, personalized learning, individualized learning, and customized learning are in some way interchangeable although adaptive learning is the most frequently used term of the four [13]. Various methods are applied for the design and evaluation of adaptive or personalized activities and contents of the digital learning platforms.

Existing reviews on Learning Analytic (LA), Learning Analytics Dashboards (LAD), and AL has not focused on the methods used to evaluate LA and LADs of Adaptive Learning Platforms (ALPs). For instance, the systematic literature

E. Brooks et al. (Eds.): DLI 2021, LNICST 435, pp. 37–52, 2022.
https://doi.org/10.1007/978-3-031-06675-7_3

review [7] presents six reviews on adaptive learning and seven reviews on learning analytics among others types of learning technologies but lacks focus on the methods for the evaluation of LA or LADs. The review [9] posed several questions on especially which methods have been employed for the evaluation of the systems. The review reports that the learners play an important role in the evaluation of intelligent tutoring systems, such as learners' experience when evaluating system usability. In the examined studies 5.66% of studies involving intelligent tutoring systems were evaluated only by learner experiences, while in combination with learner's performance, system's performance or both, learners' experiences have been used more frequently [9]. The review does not entail what methods were used for obtaining the learner experience or what types of usability tests were used.

The purpose of applying different methods of evaluation is therefore interesting to look into to get a better understanding of which perspectives are being evaluated from as well as how they are being evaluated.

This leads to the motivation for this systematic review. The motivation is to synthesize the evaluation methods applied in the design, development, and implementation of AL as they support pedagogical and learning related decisions for educators and students. Likewise, we want to examine how students' and educators' perceptions of LAD and LA are integrated in the evaluation methods. The study will contribute to the fields of usability engineering, user experience, and digital learning technology. The study on the methods of evaluating AL platforms is pivotal for improving the quality of learning experience and learning outcome, educators teaching experience and their adoption of the technology, and development process of companies and the implementation of the right evaluation methods.

The above-mentioned scope and motivation led us to devising the research question:

How to evaluate the Learning Analytics and Learning Analytics Dashboards of Adaptive Learning Platforms?

The desired outcomes is one set of methods for evaluating the functionalities and perceived experiences of the technological features, and the other set of methods on the evidence of improving learning outcomes, learning experience, and teaching quality. While the first contribute to the field of interaction design and the second contribute to the broader field of service design and innovation within the education and training domain.

2 Methods

Applying two different established methods, the protocol for the selection of papers and the protocol for the process of analysis and synthesis are conducted.

2.1 Selection of Papers: PRISMA

The selection of articles are conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [11], which includes four phases: identification, screening, eligibility, and included (See Fig. 1). Since the aim is to review evaluation methods used for LA and LAD on ALPs, various combinations of the following keywords are used: evaluation, adaptive learning, learning analytics, learning analytics dashboards, assessment, etc. The searches were restricted to peer-reviewed papers, published in English, Danish, and Norwegian (considering authors' language skills), from 2011 to the search date September 1, 2021. In consultation with a librarian and after testing different combinations of keywords, four databases were selected, and different combinations of the keywords returned the following: Scopus [n = 75], ACM [n = 144], ScienceDirect [n = 106], and Taylor & Francis [n = 38]. We envision further inclusion of databases such as IEEE Xplore, JSTOR, Routledge, Springer, and ERIC in our continued work.

The exclusion criteria implemented in screening and eligibility stages are as follows: 1) A paper that does not mention LA or LAD in relation to ALP. 2) Papers with a focus on LA and LAD in other e-learning environments which do not meet the requirements of adaptivity for the learning platform. 3) Papers without empirical data examining LA or LAD on ALP. 4) For the conference proceedings, only included papers published as part of the main conference. Workshop papers and posters were excluded.

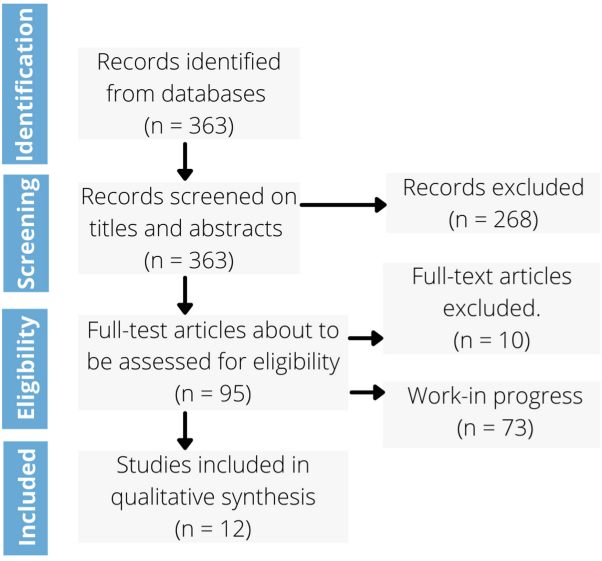

Fig. 1. PRISMA flow-chart

The two authors screened separate databases and only the papers selected by one author (n = 95) are included in this document. For this review, 10 articles and 2 reviews have been included for analysis and synthesis.

2.2 Constant Comparative Analysis Method

We applied the constant comparative analysis method for the analysis and synthesis [4]. The articles were encoded according to themes and then divided into categories. During this process, the coded sections were regularly compared to similar parts of texts containing the same codes. The intention was to create a connection between the texts and ensure the continuity of the codes' definitions [4].

Each included paper was read with the purpose of identifying methods, parameters, and purpose of evaluating LA and LAD. The data extracted from the papers are tabulated to synthesize: 1) The methods used when evaluating LA and LAD. 2) Parameters measured by the aforementioned methods to evaluate LA and LAD. 3) The purpose for the evaluation method applied. From the identified purposes a thematic analysis was initiated and categories were developed.

3 Analysis and Synthesis

In this section, we report the qualitative synthesis of the systematic review. The evaluation methods identified in the papers are summarized in the Table 1 are grouped into four categories. C1) Evaluation of LA and LAD design and framework - focusing on how LA and LAD is implemented on the platform. C2) Evaluation of performance with LA and LAD - focusing on user performance with LA and LAD statistics. C3) Evaluation of adaptivity - focusing on if and how the adaptivity functions of the system works. C4) Evaluation of perceived value - focusing on perceived value of students, educators, or users. C5) Evaluation of pedagogical and didactic theory/context - focusing on whether a pedagogical theory is the groundwork for the LA, LAD, or framework or if there are actionable pedagogical recommendations associated with the application.

Each category will have studies which are in depth described if their main focus aligns with the category. Several studies have multiple evaluations besides their main focus these papers will be mentioned in each category as evaluation features.

3.1 C1) Evaluation of la and LAD Design and Framework

Paper [1] focuses on evaluating LA and LAD design and framework whereas [8] only has evaluation of LA and LAD design and framework as a part of their study.

[1] propose EduAdapt, an architectural model for the adaptation of learning objects considering device characteristics, learning style and students' contextual

information. They develop an ontology (OntoAdapt) for recommending content to users. Particularly, for EduApadt the study wants to investigate if the use of ontology matches the learning objects adaptation scope.

The OntoAdapt [1] is an ontology which is evaluated in two phases. The first phase describes the development of the ontology and the second phase of applying it in a developed application. The first phase uses two strategies; scenarios and analyzing the quality and fidelity that OntoAdapt delivers against other ontologies. The scenarios identified different use cases from which they developed OntoAdapt. The analyzing of quality and fidelity of OntoAdapt compared to other ontologies was done with some evaluation metrics from full ontology evaluation (FOEval) (coverage, richness, and level of detail) and some provided by the software Protege (Annotations, Object property, Data property, Properties to the specific domain, Properties with specific range, Total number of classes, and Total number of subclass) to analyze the quality and the fidelity that OntoAdapt delivers in covering concepts on the associated subjects. These metrics were complemented with the tool Manchester-OWL Ontology Metric to validate and display statistics on OntoAdapts performance. These were used to calculate Attribute Richness (AR), Relation Richness (RR), Ontology Richness (OR), and Subclass Richness (SR).

The second phase, were the testing of the ontology. A mobile application prototype for Apple iOS mobile devices was developed and they used the prototype in an undergraduate course called Ubiquitous and Mobile Computing with 20 learners who used the Adapt application during 1 month. The study applied a survey using the Felder and Silverman index of learning styles with 44 items on a 5 point Likert scale on four dimensions (Active/Reflective, Sensing/Intuitive, Visual/Verbal, and Sequential/Global). Afterwards a pretest on the EduAdapt was preformed with 20 Learning Objectives (LOs). A post test was then offered after 1 month and to complete a survey on EduAdapt. The survey was based on the work of a two-tier test and a usability evaluation and was compounded of 10 statements, the students had to rate using a 5 point Likert scale to measure the level of user satisfaction. These results were evaluated on reliability with the Cronbach alpha approach and Wilcoxon-Mann-Whitney test to assess whether samples have the same distribution.

As one of the results in this expansive study "we can highlight as the main scientific contribution the proposal of a model for learning objects adaptation that employs inferences and rules in an ontology considering various contexts, including the student's learning style" [1, p. 83].

Besides this paper, an additional paper touch upon the evaluation of LA and LAD design and framework in their study. [8] presents a framework to frame user requirements of an adaptive system.

3.2 C2) Evaluation of Performance with la and LAD

Two papers focused mainly on evaluating user performance with LA and LAD and one additional paper mentioned the measurement of performance through LA and LAD but not as part of the main scope of its study.

Table 1. Review results

Author	Category	Evaluated unit	Methods	Parameters	Purpose
Di Mascio et al. (2013) [3]	C3, C4	Adaptive learning system TERENCE	Heuristic evaluation, expert reviewing, cognitive walk-through, observations, think-aloud and verbal protocols, controlled experiments, simulation and system performance indicators	Users' attitudes towards the system, users' performance and system performance	The qualitative methods (Heuristic, expert reviewing, cognitive walk-through evaluations etc.) are used to evaluate design choices for the system. Whereas, the simulations and system performance indicators are also used to evaluate the design choices but from a usability perspective on system performance
Bresó et al. (2016) [2]	C3, C4	Mechanism that adapts to stamina/mood	Surveys and simulations	Adaptability and variability	The simulation method was used to evaluated on the amount of possible outputs from the system. The surveys combined with a pilot case evaluated on the perceived levels of both variability and adaptability
Tlili et al. (2019) [14]	C3, C4	Method for modelling to learners personalities	Survey and LA student personality scores	Learner personality	LA was used to estimate learners' personalities and surveys were used to evaluate the validity of the personality models
Wei-Chih et al. (2015) [5]	C2, C3, C4	Adaptive learning algorithm	Surveys, pre- and post-tests, performance scores, and user satisfaction scores	Learner satisfaction scores and learning effectiveness	How does the algorithm performs (student performance) as well as what the learning satisfaction with the LAD were
Nye et al. (2021) [10]	C4	MentorPal, adaptive framework for virtual mentors	Formative user testing interviews, log data, pre- and post-surveys for career attitudes, and post-survey for usability	Feasibility for virtual mentoring	The LAD is evaluated on statistics which were used to check for the model quality, this was verified against users' subjective quality assurance testing

(*continued*)

Table 1. (*continued*)

Author	Category	Evaluated unit	Methods	Parameters	Purpose
Abech et al. (2016) [1]	C1, C4	Ontology model for LA and LAD	FOEval, user feedback, surveys, measurements of survey reliability, user scenarios, competence questions and usage patterns	Learners, learning objects, devices, context, context awareness, coverage, richness, and level of detail	The evaluations of the ontology goes through two phases. First phases evaluations are used for developing the ontology. Second phase evaluations are applied to compare it with other ontologies and to evaluate how the ontology performs in a learning context
Mavroudi et al. (2016) [8]	C1, C4	Teacher-led design on envisioned adaptive system	Evaluation questionnaire and Qualitative Comparative Analysis	Teacher perceptions and expectations	This paper proposes a methodology to frame requirements to a number of critical success factors in meeting the users' expectations of the system
Khawaja et al. (2014) [6]	C3, C4	Adaptive multitouch tabletop interaction application	Subjective ratings, Linguistic Inquiry and Word Count, and Advanced Text Analyzer	User's experienced cognitive load, Language Complexity Measures, and Linguistic Category Features	A method for a none-intrusive, non-manipulative adaptive learning design which adapts to users' cognitive load
Santos et al. (2015) [12]	C2, C3, C4	Ambient Intelligence Context-aware Affective Recommender Platform	Tutor Oriented Recommendations Modeling for Educational Systems methodology, user-centered design methods, data mining techniques, interviews, and SUS questionnaire	Learners' affective state, educators' tacit experiences, learner physiological signals	Exploration of ambient intelligence providing sensory-oriented feedback. How or if this improves personalized support through a recommender system
Zhang et al. (2021) [15]	C2, C4	Student-centered online one-to-one tutoring system	Pre- and post test of students' academic performance and system log files	Students' learning performance (academic), teachers' performance (attracting students)	Does such a system have a practical value

[5] developed a new algorithm, called the competency-based guided-learning algorithm (CBGLA). The study aimed to develop a CBGLA-based learning system that includes personalized learning paths which guide learners in achieving the learning objectives. The purposes of the guided-learning functions are to

accelerate and streamline the learning process. The system was tested on six third-year college students of electrical engineering before the experiment was conducted on 59 third-year college students of electrical engineering [Experimental group = 29, control group = 30]. To test the effectiveness of CGBLA a quasi-experimental research method was employed using a non-equivalent test design. The statistical mean, independent sample t-test, and one-way analysis of covariance (ANCOVA) was used to investigate the participants' learning effectiveness, satisfaction, and three dimensions of system validity through achievement of learning objectives, required learning time, and learning effectiveness. Learner satisfaction was investigated using a 16-item survey of five-point Likert scale covering three dimensions: interface design, design of adaptive guided-learning mechanism, and the perception of CBL. "The results of system validity experiments were significantly positive. This paper also conducted learning experiments to analyze learning effectiveness. Results showed that students learned more effectively under the guidance of the CBGL system than under the instruction of a teacher. [...] However, students expressed a lower degree of satisfaction when surveyed about their perception of CBL" [5, p. 124].

[15] presents the Student-Centered Online One-to-one Tutoring system (SCOOT), which deals with the cost of one-to-one tutoring. SCOOT is presented as a supplementary service where students can ask questions outside school to expand the flexibility of posing questions. The tutoring sessions with SCOOT is organized in four essential components: organization of teachers, student inquiry, and pair matching mechanism and the tutoring session. In SCOOT, teachers and students are able to communicate online through screen sharing, sending text and pictures, and speech. The teachers' interface of the application needs the teachers to log into the system and mark themselves as available for synchronous live conversations. The students' interface of the application require the students to log on and interact with the available teachers. These tutoring sessions are initiated by the students. The study seeks to evaluate the efficiency of SCOOT as well as examining how students' prior knowledge and the superficial patterns of tutoring sessions affected their learning. The evaluation include integrate students' learning performance and behavior log files instead of running between-subject experiments. The study ran for 50 days with a pre-test before and post-test afterwards. To get an in-depth understanding of how tutoring sessions affected students' learning, 40 tutoring sessions were randomly selected based on the criteria inferred and the sessions were manually labeled in detail with the coding schema developed from Chi's Interactive Constructive Active Passive (ICAP) framework. The participants consisted of 810 students in Grade 7 and 64 mathematics teachers and tutoring sessions which had a length of less than 1 min were omitted. Pretest performance combined with system usage factors was used to predict posttest performance by linear regression, this was done with Waikato Environment for Knowledge Analysis (WEKA). Common descriptive statistical analysis and Pearson correlation coefficient were computed. "The results suggested that system interventions are needed at both the student and teacher sides to facilitate good-quality tutoring interactions; otherwise, SCOOT

may further increase the difference between high- and low- achieving students" [15, p. 17].

Apart from the two above-mentioned papers [12] evaluates the effectiveness of supporting the learning process by e.g. giving affective and sensory input to help calm the user in a stressful learning context, and whether the input was helpful in the students' performance.

3.3 C3) Evaluation of Adaptivity

Four papers evaluated on the adaptivity of LA, LAD, or framework. Two additional papers mentioned adaptivity in their studies but did not present it as their main focus.

For defining personality in adaptive learning systems, [14] devised an evidence-based personality model by mapping students' participation in 15 functionalities of iMoodle learning management system against Big Five Inventory (BFI) dimensions (i.e. Extraversion, Agreeableness, Consciousness, Neuroticism, and Openness). The devised method is defined as an LA approach for defining the learners' personality. Based on 50 students data, Chi-square test is used as an assessment criterion to compare between the assessed personality levels from the results of LA approach and that assessed from the BFI results. Since the study is exploratory and little information is previously known, the obtained experimental data from their pilot experiment was validated using three methods namely, Chi-square, 10-fold cross-validation and Cohen's Kappa. The study concluded that the "LA approach with Bayesian network can model learners' personalities with an acceptable precision and a fair agreement compared to BFI for only three personality dimensions, namely, extraversion, openness and neuroticism" [14, p. 12].

To evaluate adaptability and variability of content from both simulations and user feedback, [2] presents the Personal Health System as a part of Help4Mood which supports users in not relapsing into depression, thereby learning to live with their condition. The Personal Health System is a developed tool that adapts its content to users stamina or mood. The design of the Personal Health System has been performed by adopting a user centred design methodology, which was done by involving a set of users, clinicians and caregivers. The evaluation is done through two methodologies one is producing simulated data and the second is collecting user feedback on the system. The simulation data was produced with two categories of scenarios in mind. The scenarios were designed on clinical requirements and were restrictive and flexible scenarios. Restrictive scenarios had a high number of constraints in the relative order and dependencies between tasks. They also had a high number of constraints in the periodicity and priority of the tasks. The flexible scenario in contrast had a minor number of constraints. The evaluation space corresponds to the multivariate combination of answers to all questions that might make sense given the context of the user. The simulations were done on 19 tasks and 31 subtasks and a task could be formed by one or more subtasks. There were 20.000 simulations of interactive sessions (restrictive n = 10.000 and flexible n = 10.000). In this study, adaptability was

defined as how much the produced content of a session can change in relation to current and past information and inferred about the users' condition. Variability was defined as how the content order is offered depending both on user actions during the interactions and restrictions defined by clinicians. The second methodology encompassed two tests where users used the system and afterwards answered an 11-item survey with 3-point Likert scale on their perceived usefulness of different functionalities of the system. Two of these items referred to adaptability and variability. The paper concluded that "We can ensure that our framework provides a sufficient degree of adaptive and varied sessions, allowing the personalisation of the interactive sessions in order to improve the user experience" [2, p. 90].

In a study by [6], user's experienced cognitive load is examined to help improve performance in complex, time-critical situations by dynamically deploying more appropriate output strategies to reduce cognitive load. This is done through linguistic behavioral features as indices of user's cognitive load. A pilot study was conducted on a paper mock-up with two teams of four participants consisting of experts from fire management work roles. Their feedback improve the task design as the interaction design. The study examined a session where 44 participants (11 teams of four operators) participants strategically managed fire fighting tasks as a team. Participants interacted with a multi-touch tabletop screen that displays the fire management tasks and related information. All participants had general knowledge about firefighting, but none had ever participated in any actual firefighting, training fire fighting exercises, or used any fire management system before. Task design was set up with three different levels of task complexity or cognitive load. The levels were low, medium, and high (in the analysis combining low and medium to a single low category) Data consisted of participants voices which were recorded with wireless close-talk microphones recorded with the audio recording tool WaveSurfer and two video cameras which were used to record the operators' interactions. Further data consisted of logs of interactions with the touch table including operators' touch positions and dragging behavior as well as a survey on the self-rated perception of task difficulty as individuals and as a team on two separate 9-point Likert scales. The survey also contained an open question for general comments on task complexity, use of policy documents, and any communication issues. The analysis were done on observations, data transcription, feature extraction, and statistical analyses of the linguistic features with Linguistic Inquiry, Word Count, and Advanced Text Analyzer to investigate the variations in their behavior under different task load levels. In conclusion the paper states that: "An interaction system that is able to analyze users' speech and linguistic patterns to determine their current cognitive load could dynamically adapt its response to minimise the users' extraneous cognitive load and help them maintain task performance" [6, p. 362].

[12] presents an Ambient Intelligence Context-aware Affective Recommender Platform (AICARP) that applies Tutor Oriented Recommendations Modeling for Educational Systems (TORMES) elicitation methodology to sense changes in learners' affective state. AICARP delivers interactive context-aware affective

educational recommendations in an interactive way through complementary sensory communication channels. The recommendations are given to make users adjust breathing, stress etc. To evaluate the TORMES methodology, problem scenarios were used to identify the necessary requirements or user goals while taking into account the context of use elicited in the previous activity. Problem scenarios were used to develop solution scenarios that solved or avoided the problems posed by delivering interactive recommendations. To specify these solutions the recommendation modeling work with five dimensions; recommended action (what), recommendation rules (when and who), justification of the recommendation (why), recommendation format (how and where), and recommendation attributes (which). Evaluation of the scenarios were carried out by applying the user-centered design method Wizard of Oz. In this empirical study, a psychoeducational expert with experience in supporting learners face-to-face and online acted as the Wizard. Video of the participant and affective data (pulse, skin temperature, skin resistance, and skin conductance) was visualized to the wizard who in turn generated the associated recommend action (e.g. the green LED and the buzzer playing a pure tone). The study had six participants one of them being visually impaired. Before the study began participants completed the General Self-Efficacy Scale (GSE), the Big Five Inventory (BFI), and the Positive and Negative Affect Schedule (PANAS). As part of the study participants had to complete two tasks. Each of the task involved speaking for 5 min, while being recorded with the webcam. Before talking, participants had 1 min to think about what to say. Data consisted of AICARP system data (the previous mentioned physiological data), recordings from a webcam (facial expressions and voice), recordings from a video-camera (body movements), and time-stamped notes by an observer. The impact of the elicited interactive recommendation on the learner was evaluated at the end of the experiment by means of a questionnaire and an interview. The questionnaire was the System Usability Scale (SUS) 10-item 5 point Likert scale. The interview consisted of five open questions with the goal of understanding participants' opinions of their interaction with the system regarding perception, intrusiveness, and utility. Chi-square test was conducted to determine whether there were independence between the usability of the system and the effectiveness of the recommendations perceived by the participants. Answers given in the open questions were coded categorically. To verify these categories chi-square tests were again applied. The results cannot be applied as representative due to a very low sample size. The study concludes that "[...] this research opens a new avenue in related literature which focuses on managing the recommendation opportunities that an ambient intelligent scenario can provide to tackle affective issues during the language learning process when preparing for the oral examination of a second language learning course" [12, p. 50].

In addition to these papers, a number of papers mentions evaluation methods that evaluates on adaptivity. [3] presents usability associated with adaptivity and [5] effectiveness of adaptivity.

3.4 C4) Evaluation of Perceived Value

2 papers had their main focus on evaluating from users' perceived value or evaluation of users' perceived value. All of the rest 8 studies reviewed in this paper had in one way or the other included users' perceived value as a feature of their studies.

[3] describes the development of the TERENCE system's Graphical User Interface (GUI) prototypes through evidence based and user-centered design where they identified users' requirements and context of use by using users and domain experts. The first group were learners ($n \approx 170$), which here is described as 7–8 year old primary-school students who are poor comprehending and hard of hearing or deaf. The second group were educators ($n \approx 10$), who were primary school teachers, support teachers, and parents of learners. The last group were experts ($n \approx 10$) who were psychologists and linguists, who designed and developed the learning material. In evaluating the ALS TERENCE the GUI was assessed and it was assessed on two levels. The Learner GUI and the Expert/Educator GUI. Three evaluations were done and the two first ones were done with experts. The purpose for the expert evaluation were to assess whether the learning material were adequate for the learners and to evaluate the usability of prototypes, in particular whether the interfaces followed standard visual design guidelines, whether the interfaces supported the user's next step to achieve the task, and whether the interfaces provided appropriate feedback. The prototypes were evaluated using heuristic evaluation, expert review and cognitive walk through. More evaluations were conducted with end users. The purpose was to provide indications related to the pedagogical effectiveness of the prototypes and to evaluate their usability. The methods were: observational, think-aloud, verbal protocols, and controlled experiment. The paper informs about upcoming analyses of a large-scale evaluation with 900 end users. The initial findings and analysis is not included here. Their findings informs an expansion of usability testing to also include timing and focus of users' participation as well as system performance during the execution of users' tasks [3, p. 5-7].

[10] presents a virtual mentor system called MentorPal. In this empirical study the system gives career advice to high school students (n = 31) attending STEM internships who considering STEM careers. They participated in 3 sessions with MentorPal where they completed a pre-survey, interacted with MentorPal for 25–30 min, and then completed a post-survey. Researchers unobtrusively observed the students during usage and were available to help if needed. The STEM career advice were given with focus on STEM careers in the Navy. The system works as follows; the student asks one of four recorded virtually represented mentors by Free Text, Speech Input, or Topic Buttons about the mentor's career to get a better understanding of the career's alignment with the students interests and goals. MentorPal responds with the most suited answer. The primary pedagogical technique encouraged during recording of the mentors was the use of anecdotes and narrative. Development of MentorPal was done with three parameters in focus: Conversational Flow, Video-Chat Authenticity, and Low Cost. MentorPals performance was evaluated through pre- and post

surveys. Usability was evaluated with Unified Theory of Acceptance and Use of Technology constructs (UTAUT) survey on a 6 point Likert scale with 6 items. To evaluate change in attitude towards specific careers a survey was generate from variants of the CAPA Career Confidence Inventory and the CAPA Interest Inventory which resulted in respectively 50 items based on the approximately 400 CAPA items on a 5 point Likert scale. The results were tested and evaluated through traditional classifier statistics these were used to check for the direction of increases or decreases for model quality with 5-fold cross-validation accuracy scores, this was verified against subjective quality assurance testing. Their findings were limited on both sample size, sample diversity, and impact but one of the clear conclusions were that: "A panel of four mentors (even one hypothetically optimized through hindsight) is insufficient to cover either the main career interests or diversity representation of even 31 students. So, future research should investigate how students respond to self-reported or automatically personalized panels drawn from a larger set of mentors representing broader career choices and backgrounds" [10, p. 39].

In addition, several papers mentions the evaluation of perceived value. [1,2,12,14,15] present the evaluation of perceived value as a method for further informing performance of LA. [2,12] uses evaluation of perceived value to evaluate adaptability and variability and to assess the usability of the LA, [5] presents it to assess satisfaction levels of LA, [6] estimate perceived level of cognitive load, [15] assesses the practical value of the LA, and [1,8] developing the application.

3.5 C5) Evaluation of Pedagogical and Didactic Theory/Context

Three papers mentioned pedagogical theory as a contextual factor for their studies. None of the studies evaluated on how pedagogical or didactic theory was evaluated upon in either LA, LAD, or frameworks. The three papers that mentioned pedagogical were: [3] who had a second iteration of expert evaluation which consisted of 10 learning experts. As they applied the TERRENCE system to their prototype they included a pedagogical direction described as the pedagogical stimulation plan. The results from the user evaluation consisting of approx 170 users assessed whether the pedagogical effectiveness of the prototypes, the evaluation of its usability, and whether expectations to the pedagogical stimulation plan was met. This was done through observational, think-aloud, verbal protocols, and controlled experiment. [1] reviewed other works on ontology which had a pedagogical approach. This was compared to their own ontology's adaption to learning styles but their own ontology was not assessed on any pedagogical parameters. [5] used competency-based learning to develop their CBGLA algorithm but their study does not mention how CBGLA could or should be implemented in a pedagogical context neither how CBGLA resulted in the development of users' competencies.

4 Conclusion and Discussion

In this work in progress systematic literature review, we identified 12 relevant papers, synthesized 10 empirical papers, and covered two reviews as part of the introduction for establishing the scope of the paper. The methods that directly or indirectly contribute to the evaluation of LA or LAD of ALPs are grouped into five categories: C1) evaluation of LA and LAD design and framework, C2) evaluation of performance with LA and LAD, C3) evaluation of adaptivity functions of the system, C4) evaluation of perceived value, and C5) Evaluation of pedagogical and didactic theory/context. Figure 2 shows the number of papers covering the methods under the five categories as the central focus of their study.

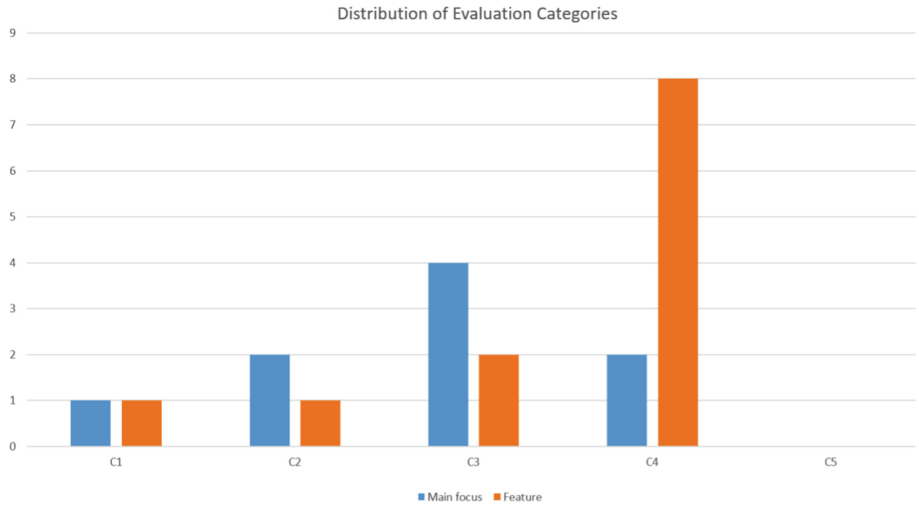

Fig. 2. Distribution of evaluation categories

Pedagogical and didactic theory/context (C5) as a theme occurred in multiple papers but none of the papers covered the evaluation of impact of an LA or LADs of ALP. LA and LAD are rarely examined in an educational context as a learning tool which informs either students and educators on making informed pedagogical or didactic choices framed by a pedagogical or didactic theory.

We experienced the lack of pedagogical theories and concepts such as motivation, engagement, gamification, and nudging to mention a few. For future studies, we raise the question, how do we improve learning and teaching quality with LA if there are no learning theory attached to the data collection and presentation? And how can LA and LAD lead to better learning or teaching if there are no actions associated with the data rather than just a presentation of learning objectives' difficulty, time spent on the platform or active users. Pedagogy and didactics needs to be connected with LA and LAD of ALPs to support teachers

and students as they focus on cognitive and meta-cognitive impact, behavioral change, and social learning activities.

Broadly, we see assessments with ontologies, frameworks, methodologies, experimental designs, mathematical models, and LA statistics which are almost all the building blocks of a LA. Only evaluations of the visualization and the pedagogical elements are not present.

References

1. Abech, M., et al.: A model for learning objects adaptation in light of mobile and context-aware computing. Pers. Ubiquit. Comput. **20**(2), 167–184 (2016). ISSN: 1617–4909. https://doi.org/10.1007/s00779-016-0902-3. http://moodle.com. http://link.springer.com/10.1007/s00779-016-0902-3
2. Bresó, A., et al.: A novel approach to improve the planning of adaptive and interactive sessions for the treatment of major depression. Int. J. Hum. Comput. Stud. **87**, 80–91 (2016). ISSN: 10959300. https://doi.org/10.1016/j.ijhcs.2015.11.003
3. Di Mascio, T., et al.: Design choices: affected by user feedback? Affected by system performances? Lessons learned from the TERENCE project. In: ACM International Conference Proceeding Series, pp. 16–19, September 2013. https://doi.org/10.1145/2499149.2499171
4. Hewitt-Taylor, J.: Use of constant comparative analysis in qualitative research. Nurs. Stand. (Royal College of Nursing (Great Britain): 1987) **15**, 39–42 (2001). https://doi.org/10.7748/ns2001.07.15.42.39.c3052
5. Hsu, W.-C., et al.: A competency-based guided-learning algorithm applied on adaptively guiding e-learning. Interact. Learn. Environ. **23**(1), 106–125 (2015). ISSN: 1744–5191. https://doi.org/10.1080/10494820.2012.745432. https://www.tandfonline.com/action/journalInformation?journalCode=nile
6. Asif Khawaja, M., et al.: Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. Int. J. Hum. Comput. Interact. **30**(5), 343–368 (2014). ISSN: 1532–7590. https://doi.org/10.1080/10447318.2013.860579. https://www.tandfonline.com/action/journalInformation?journalCode=hihc
7. Martin, F., Dennen, V.P., Bonk, C.J.: A synthesis of systematic review research on emerging learning environments and technologies. Educ. Technol. Res. Dev. **68**(4), 1613–1633 (2020). ISSN: 1042–1629. https://doi.org/10.1007/s11423-020-09812-2. https://link.springer.com/10.1007/s11423-020-09812-2
8. Mavroudi, A., et al.: Teacher-led design of an adaptive learning environment. Interact. Learn. Environ. **24**(8), 1996–2010 (2016). ISSN: 1049–4820. https://doi.org/10.1080/10494820.2015.1073747. https://www.tandfonline.com/action/journalInformation?journalCode=nile. https://www.tandfonline.com/doi/full/10.1080/10494820.2015.1073747
9. Mousavinasab, E., et al.: Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Interact. Learn. Environ. **29**(1), 142–163 (2021). https://doi.org/10.1080/10494820.2018.1558257
10. Nye, B.D., et al.: Feasibility and usability of MentorPal, a framework for rapid development of virtual mentors. J. Res. Technol. Educ. **53**(1), 21–43 (2021). https://doi.org/10.1080/15391523.2020.1771640. https://www.tandfonline.com/action/journalInformation?journalCode=ujrt20

11. Page, M.J., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ **372**, n71 (2021). ISSN: 1756–1833. https://doi.org/10.1136/bmj.n71. https://www.bmj.com/lookup/doi/10.1136/bmj.n71
12. Santos, O.C., Boticario, J.G., Rodriguez-Sanchez, M.C.: New review of hypermedia and multimedia toward interactive context-aware affective educational recommendations in computer-assisted language learning toward interactive context-aware affective educational recommendations in computer-assisted language learning (2015). ISSN: 1361–4568. https://doi.org/10.1080/13614568.2015.1058428. https://www.tandfonline.com/action/journalInformation?journalCode=tham20
13. Shemshack, A., Spector, J.M.: A systematic literature review of personalized learning terms. Smart Learn. Environ. **7**(1), 1–20 (2020). https://doi.org/10.1186/s40561-020-00140-9
14. Tlili, A., et al.: Automatic modeling learner's personality using learning analytics approach in an intelligent Moodle learning platform. Interact. Learn. Environ. (2019). ISSN: 17445191. https://doi.org/10.1080/10494820.2019.1636084. https://www.tandfonline.com/action/journalInformation?journalCode=nile20
15. Zhang, L., et al.: Evaluation of a student-centered online one-to-one tutoring system. Interact. Learn. Environ. **0**(0), 1–19 (2021). https://doi.org/10.1080/10494820.2021.1958234. https://www.tandfonline.com/action/journalInformation?journalCode=nile