

Towards the Development of Guidelines for Educational Evaluation of Augmented Reality Tools

Manoela M. O. da Silva*
Voxar Labs
Informatics Center
UFPE - Brazil

Rafael A. Roberto†
Voxar Labs
Informatics Center
UFPE - Brazil

Veronica Teichrieb‡
Voxar Labs
Informatics Center
UFPE - Brazil

Patricia S. Cavalcante§
EDUMATEC
Education Center
UFPE - Brazil

ABSTRACT

Augmented Reality (AR) is heralded to be a promising technology for education. Thus, it is important to properly evaluate it so practitioners feel more confident in its use. Considering the current lack of studies to evaluate AR based educational technology, this work aims to provide some guidelines in order to assist researchers when conducting educational evaluations. The proposed principles were based on both theoretical and practical research. The guidelines presented involve the use of multiple metrics, both quantitative and qualitative, high involvement of teachers and a comprehensive evaluation taking into account both formative and summative aspects. Two practical evaluation experiences were designed applying these principles and their findings are discussed in this paper.

Index Terms: K.3.m [Computing Milieux]: Computers and Education—Miscellaneous

1 INTRODUCTION

Technology is part of everyday life and it has been increasingly introduced in the classrooms. [17] explains that its arrival in schools implies a range of challenges to teachers, students and the pedagogical team since the technology brings a double challenge: to adapt schools to its advances and to guide the people involved to master critically this new media.

Augmented Reality (AR) consists of adding virtual elements to a real scene coherently so that ideally users cannot differentiate them from the real scene [2]. The coexistence of virtual and real environments allows learners to experience phenomena that otherwise would be impossible in the real world and, therefore, develop important abilities that cannot be evolved in other technology learning environments [30].

While AR offers new learning opportunities, it also creates new challenges for education in different domains, such as technological, learning and pedagogical issues [18]. Developments in AR technology have enabled many solutions to be developed for the educational field. Their impact in the educational setting is of utmost importance. Hence, evaluation of learning outcomes is an important step in order to unveil its potential in education.

Taking this scenario into account, this paper is oriented to provide additional data on the effects of AR in education and on the process of evaluating educational technology holistically, involving both the teachers and the students in the process. Therefore, it provides guidelines to perform educational evaluations of AR applications that were based on theoretical research. In order to assist researchers when conducting similar evaluations, this

*e-mail: mmos@cin.ufpe.br

†e-mail: rar3@cin.ufpe.br

‡e-mail: vt@cin.ufpe.br

§e-mail: patricia3smith@gmail.com

study also presents and discusses two semi-experiments whose evaluations were based on the principles proposed.

The present work is organized as follows: section 2 presents the related works, section 3 describes the evaluation guidelines, section 4 details the practical experiments performed and discuss the results obtained and, finally, section 5 concludes the paper.

2 RELATED WORK

It has been long since AR's potential in education is being investigated. AR can aid learning and make the overall learning process much more interesting and pleasant [19]. It dramatically shifts the location and timing of education and training [20]. Unlike other computer interfaces that draw users away from the real world and onto the screen, AR interfaces enhance the real world experience. [4] highlights some reasons why educational experiences afforded by AR are different:

1. Support of seamless interaction between real and virtual environments;
2. The use of a tangible interface metaphor for object manipulation;
3. The ability to transition smoothly between reality and virtuality.

Although AR has been studied for over forty years it has only been recently that researchers have begun to formally evaluate AR applications [11]. The authors point out that one reason for the lack of user evaluations in AR may be, among other factors, an absence of knowledge on how to properly evaluate AR experiences and design experiments. According to them, there seems to be a lack of understanding regarding the need of doing those types of studies and the right motivation for carrying them. If user evaluations are conducted out of incorrect motivation or if empirical methods are not properly applied, the reported results and findings are of limited value or can even be misleading.

When it comes to educational applications, it is very important to evaluate their impact on the learning experience and the feasibility of incorporating them into the classrooms. There may be many factors involved varying from cost to staff's acceptance. Evaluation of technology is an important step in design instruction. Hence, it is necessary to evaluate it properly so practitioners are more confident in its positive effects on the learning process. It is also relevant to consider the point of view of both teachers and learners since they might differ. For instance, [3] had shown the perceived usefulness and the perceived enjoyment as relevant factors for student's acceptance of an AR application, while the perceived ease of use was not a significant factor for student's acceptance.

A recent survey reviewed applications intended to complement traditional curriculum materials for K-12 [26]. In this work, the authors performed a qualitative analysis on the design aspects and evaluation for AR Learning Environments (ARLES). The focus of the survey was to investigate ARLES designed for kindergarten, primary and/or secondary school. Its aim was to explore learning

theories as basis for effective learning experiences. Concerning the evaluation aspects, the authors discovered that aside from the performance of students in pre-tests and post-tests, other aspects of the learning experience such as motivation and satisfaction were usually observed in the evaluations performed in the literature.

[9] analyzed 14 papers on a systematic review on how researchers evaluate AR based educational applications. Her results confirm that most of the applications evaluated are designed to K-12 students and five of them are directed to undergraduate and master students. Additionally, only two studies work with very young learners aged up to seven-eight years old [5]. This age range brings specific challenges to the evaluation process since these students are usually not formally evaluated through tests. They are rather evaluated holistically concerning the abilities and competencies they achieve throughout time. There is also a lack of studies regarding early literacy development using AR systems [9].

It is well known the important role teachers play in technology adoption in the classrooms, however, many studies evaluating educational technologies show no involvement [6] or almost no involvement of the teachers in the evaluation [23]. In those latter ones, teachers sometimes help in order to carry out the evaluations. Some researchers involve the school board in order to help in the selection of contents to be worked with [14]. Others take into account local curriculum in order to develop content for the tool [25]. Only one study considered teacher's experiences as an important part of the educational evaluation. [16], for instance, investigated teacher's experiences. One of the measures was teacher's judgments of usability and value of technologies related to field trip instruction. They also collected feedback from teacher participants through a post-interview. Prior to the field trip, two of the teachers used learning quests during the class, while the third teacher used them as one of the "stations" during the activities on the day prior to the field trip. During the field trip, the teacher led a discussion about the data collected by students.

It is difficult to find studies that evaluate solely educational aspects. Different aspects are also taken into account when evaluating the tools, such as efficiency, satisfaction, motivation, aspects concerning usability and user's attitudes among others. Indeed, eight studies combine different methodologies in order to evaluate AR tools, although, only two of them combined multiple metrics to evaluate educational evaluation [9].

3 GUIDELINES FOR EDUCATIONAL EVALUATION

[12] point out the usefulness of mixed methods in the research design. These authors highlight the importance of employing both quantitative and qualitative metrics as a way of compensating the weakness of each method. [7] explains that the advantages of using multimethod approaches in social research are manifold. The authors highlight two of them:

1. While single observation in fields such as physics and chemistry usually yield sufficient and unambiguous information, it provides a limited view of the complexity of human behavior and interactions;
2. Exclusive reliance on one method may bias or distort the researcher's picture of a particular reality he/she is investigating.

In order to effectively evaluate new educational technology, it is important to effectively integrate them in the schools. [10] points out two premises for effective integration and implementation of technology for K-12 classrooms, that are:

1. The teacher must act as an instructional designer, planning the use of technology to support learning;

2. Schools must support teachers in this role.

It is important for researchers and developers to have an understand on how teachers will integrate new technologies into their lessons since this will shape student's learning opportunities. [13] stress the need to involve teachers in the process of adopting new technology so the activities are integrated to their lesson plan and meaningful to the students. For instance, activity theory [21] shows that activities are culturally mediated and inserted into a given context that includes the mediation of artifacts, of the community, of its rules and its division of labor. In the process of transforming the activity of teaching into learning, there is a whole complex of mediations involving the curriculum, the educational rules, teacher's training, artifacts to name a few. This complex scenario needs to be taken into account in order for researchers to understand the changes caused by the introduction of a new artifact and the changes needed to expand and adjust the system.

[8] explains that the evaluation of a piece of technology in isolation will tend to focus on various aspects of the technology itself, such as screen design and text layout. On the other hand, the evaluation of a courseware within the course itself will allow for examination of other factors that will lead to successful integration of the product within the course. Some of these aspects are:

- Educational setting;
- Aims and objectives of the course;
- Teaching approach;
- Learning strategies;
- Assessment methods;
- Implementation strategy.

Formative evaluations as stated by Scriven are typically conducted during the development or improvement of a program, person or product and it is conducted with the intent to improve [27]. On the other hand, summative evaluation is typically quantitative, using numeric scores or letter grades to assess learner achievement. Thus, a comprehensive evaluation involving both types of assessment is advisable in order to have a better overview of the process and its outcome.

4 PRACTICAL EXPERIENCE

The guidelines were generated based on theoretical research. Nevertheless, it is important to have a practical experience in order to validate these principles. Thus, a semi-experiment was designed following the principles pointed out above. The goal of this study was to evaluate the impact of the ARBlocks [24] in the literacy progress of grade 1 students from a public school in Brazil. Two first grade groups taught by the same teacher participated in the study. Each class has approximately 20 students with ages between 6 and 7 years old. The teacher conducts a regular evaluation in order to assess student's literacy skills. Based on this evaluation, the tool was applied in the group with lowest scores twice a week for four weeks. The other was our control group. The complete methodology of this study can be seen in Figure 1.

Through this diagram, it is possible to observe the use of multiple metrics: post-test, observation, post-interview and the third and fourth annual evaluation. They are both quantitative (first one) and qualitative (the other three). Moreover, the annual evaluation was designed by the teacher in which she applies a literacy skill assessment approximately every two months in order to monitor student's progress. On this evaluation, she classifies them according to the psychogenesis of written language theory [28]. Thus, these evaluations encompass both formative and summative evaluations

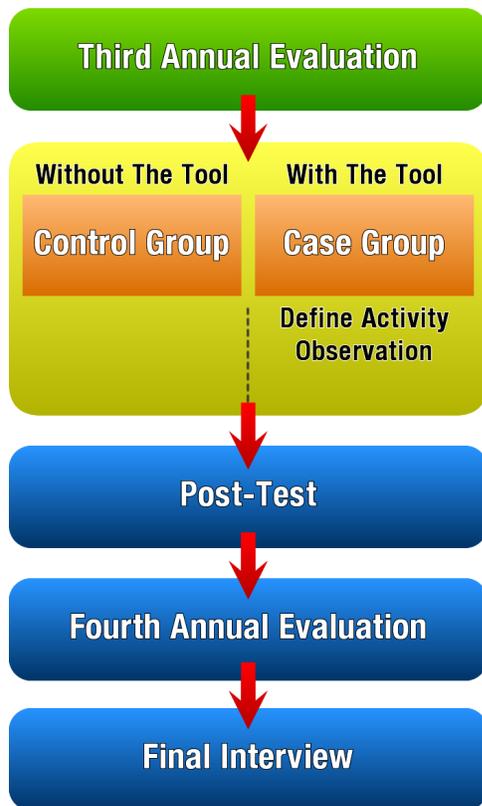


Figure 1: Diagram with the methodology design of the first study. Third and fourth annual evaluation correspond to students' academic achievement.

since rather than just a punctual measurement, student's progress throughout the process was taken into account. The teacher was involved in the entire process of integration of AR technology into her lessons. She designed the activities based on her classroom needs and programmers developed them.

The activities proposed by the teacher involved mostly reading skills and phonemic awareness since this is one of the first steps to reach reading competence [1]. The sections were organized so that each child could interact with the AR tool every visit. The class was divided in three groups of six to seven participants each, depending on the number of students attending the class. While one group was using the system, the others were engaged in other educational activities related to the same topic. This classroom arrangement was proposed by the teacher in order to manage the fact that it was not possible to work with all students at once using the tool. Each section lasted about 30 to 45 minutes depending on the activity. In [28], it is available the full description of the activities and results of the experiment.

After the period using the AR tool, a post-test was applied with both the case and control groups. The test was elaborated with the help of the teacher, which was important to guarantee students familiarity with it. This post-test was divided in two parts, one with questions concerning contents worked with the help of the tool and the other about contents worked traditionally without the help of the tool. The assessment carried out by the teacher was also taken into account in order to provide researchers with a better overview of students throughout the process. As a qualitative evaluation, we used observations and conducted a semi-structured interview about her considerations regarding the tool's use [28].

Although preliminary, based on the data collected from the

different metrics used, the AR tool helped to motivate students and to foster the development of their literacy skills. The methodology used provided some positive aspects. For instance, the fact that the tool was integrated to the lesson plan was highlighted as important by the teacher. She pointed out that it was different from what is usually done in the informatics laboratory. This aspect seemed to have raised teacher's engagement to use the tool as well as reflected in student's results.

Additionally, the use of multiple metrics allowed researchers to understand details of interaction that were important. For instance, student's motivation to work with the tool, teacher's impression of the work and changes needed in order to use it (e.g: classroom arrangement) to name a few. Besides, the use of her own evaluation combined with the other metrics, namely, the interview and observations, revealed that the student's progress was beyond the expected.

As for limitations, we point out the fact that there was no instrument to capture the teacher's impressions along the process which may be problematic since in the final interview researchers may lose important information forgotten along the way. Finally, authors understand that the limited time spent during the study, one month, was not enough to allow students and teacher to get completely used to the tool. Therefore, in this case, the novelty effect of the technology may still be considered a confounding.

Based on the lessons learned during the first study, we have performed a second one using the same AR tool through a semi-experiment in an English language school located in Brazil. The tool was used in four different groups. Two "Pre-kinder 2" (PK2), age range 6-7 years old, and two "Kids 1" (K1), age range 8-9 years old. These groups were taught by two different teachers. Classes in this school last one hour and fifteen minutes twice a week. PK2 case group has 12 students and the control group has 10. K1 case group has 9 students and K1 control group has 8 students. Statistical measures were also employed. Language course curricula are usually designed for a semester, which provides a better overview of the process. The semi-experiment was conducted for approximately three months, which represents half of the course level. Additionally, authors made sure that the groups involved had similar knowledge levels [29]. The methodology of the final study can be observed in Figure 2.

Similar to first methodology, it is possible to observe the use of multiple metrics: school's evaluations, post-test, observations, interviews and the research diary. These metrics are also both quantitative (first two) and qualitative (last three). The school evaluation is holistic and occurs in 2 moments, the middle and final term, which is also a good indicator of students' previous knowledge. However, it is important to point out that the school evaluation also have a qualitative component. Teachers always write some comments regarding student's development in each report card. Thus, this study also encompasses both formative and summative evaluations. The formative evaluation is represented by the qualitative metrics which intend to understand the process of integration of AR into this scenario as well as to promote the improvement of the process. On the other hand, the summative evaluation is represented by the post-test and the school's evaluations.

Students started to use the tool after the middle term evaluation and worked with it until the final term evaluation. The teachers, as in the pilot study, were involved in all the process. This study also had a new instrument of research: the research diary, which allows researchers to get information about day to day activities and explore those information in a subsequent interview [15, 22]. Additionally, a pre-interview was also conducted in order to present the tool to the teachers, understand their needs concerning the system, explain all the process to them and the role they will play in it.

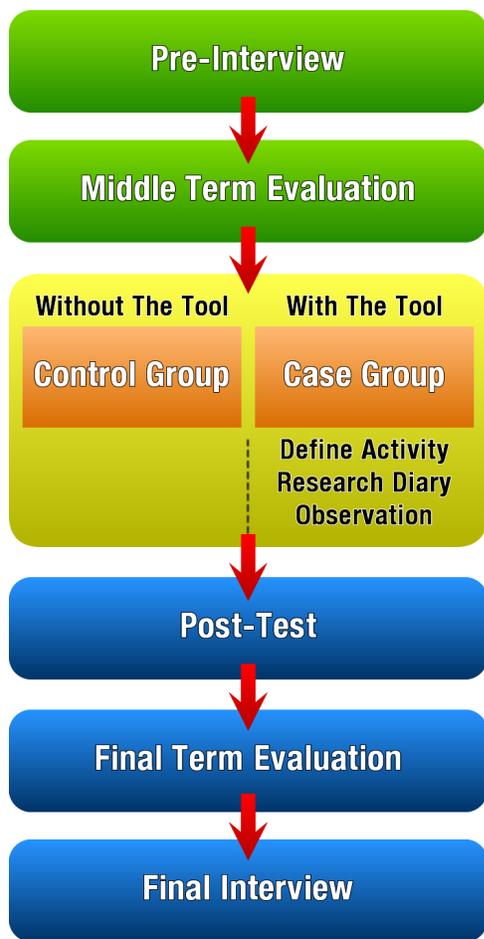


Figure 2: Diagram with the methodology design of the second study. The middle and the final term evaluation correspond to students' academic achievement.

Both PK2 and K1 teachers elaborated activities based on what they were studying at the moment. ARBlocks sessions lasted 15 to 25 minutes each. PK2 group had sessions twice a week while K1 group had sessions once a week due to time constraints. During the experiment, it was observed that most of the quantitative results were not statistically significant due to the small sample size of the groups. Taking into account the difficulty to have a suitable number of students, the use of multiple metrics is a viable alternative to counteract for this aspect. Thus, it was possible to combine the quantitative evidence with the qualitative data in order to have a proper evaluation. [29] provides a complete discussion of the results from this experiment.

The use of the AR tool for a longer period of time allowed it to be better integrated into the classroom routine. It counteracted for the novelty effect of the technology, which can result in misleading conclusions. Nevertheless, this prolonged use of the tool may require research instruments that will capture data throughout the process. Therefore, the research diary helped to understand the impact of the tool in the initial sessions since in the post-interview, the teacher focused more in her observations of the last ones.

The technology should be seen by the teacher as something that adds value to what they do rather than something that will add up to their workload. This work also attested the need to have high involvement and engagement of the teachers in the evaluation process. Authors are aware that this may not be fully achieved since

this is a demanding task.

The planning process is also crucial concerning new technology adoption. Teachers need to have full support for planning their lessons for technology use in order to take advantage of their unique characteristics rather than just adapting their activities which may be seen as extra work.

5 CONCLUSION

Based on the research carried out, it was possible to see that there is still a lack of evaluations for educational AR applications. Therefore, this work intended to provide some insight for researchers by presenting guidelines to assist them when conducting this type of studies. These guidelines proposed concern the following aspects:

1. **Use of multiple metrics:** based on the literature review and our practical insight, the authors advocate for the use of multiple metrics both quantitative and qualitative in order to have a better overview of the technology inserted in the teaching context as well as its effects;
2. **Comprehensive evaluation:** although it is not always possible to have a longitudinal evaluation, it is recommended to have a comprehension of more than punctual assessments but rather understand its effect in student's development in a longer term. In our evaluation, it was used the teacher's regular evaluation. This evaluation was complemented with the post-test applied after the use of the tool along with other metrics, such as observations, interviews and a research diary (final study only);
3. **Role of teacher:** it is widely recognized that teachers play a major role in technology adoption in the schools. Especially with young children, teacher's role is more evident since students at younger ages do not have formal exams but are rather evaluated continuously in their development progress. Therefore, it is important for researchers to understand teacher's routines, goals and hopefully the criteria they have to select resources for their classroom. Thus, it is advisable to consider teacher's point of view when adopting, using and evaluating technology integration in their classroom. Moreover, it is important to have tools that are flexible enough in order to facilitate teachers' input of content. It was evident in the field research that flexibility in terms of interaction is also recommended. Additionally, when involving teachers, it is important that they acknowledge the value of a given technology since any process of adoption of a new tool demands a high level of engagement from them, which may be time consuming. Although defying, we believe this involvement may bring important insight for both researchers and developers.

As for limitations, during our experiments, it was not possible to have a larger sample of students from the same groups, which would have reinforced our conclusions and assured statistic significance of the results. Future works might apply those guidelines in different contexts and with different age groups in order to expand them and provide practical advice on how engaging better with teachers and to better understand their classroom. However, authors are aware that different settings would bring forth different needs.

ACKNOWLEDGEMENTS

Authors would like to thank CNPq for partially funding this research (processes 134114/2013-2 and 140898/2014-0), as well as the schools, teachers and students who participated in this study.

REFERENCES

- [1] M. J. Adams, B. R. Foorman, I. Lundberg, and T. Beeler. *Phonemic awareness in young children: A classroom curriculum*. PH Brookes, 1998.
- [2] R. T. Azuma et al. A survey of augmented reality. *Presence*, 6(4):355–385, 1997.
- [3] A. Balog and C. Pribeanu. The Role of Perceived Enjoyment in the Students' Acceptance of an Augmented Reality Teaching Platform: a Structural Equation Modelling Approach. *Studies in Informatics and Control*, 19(3):319–330, sep 2010.
- [4] M. Billinghurst and A. Dünser. Augmented Reality in the Classroom. *Computer*, 45(7):56–63, July 2012.
- [5] S. Cai, X. Wang, and F.-K. Chiang. A case study of augmented reality simulation system application in a chemistry course. *Computers in Human Behavior*, 37:31 – 40, 2014.
- [6] K.-E. Chang, C.-T. Chang, H.-T. Hou, Y.-T. Sung, H.-L. Chao, and C.-M. Lee. Development and behavioral pattern analysis of a mobile guide system with augmented reality for painting appreciation instruction in an art museum. *Computers & Education*, 71:185 – 197, 2014.
- [7] L. Cohen, L. Manion, and K. Morrison. *Research Methods in Education*. Taylor & Francis ebook collection. Taylor & Francis, 2003.
- [8] P. Crompton. Evaluation: A practical guide to methods. *Learning Technology Dissemination Initiative*, 66, 1996.
- [9] M. M. O. da Silva. Evaluation of the Use of Augmented Reality Tools in the Education Field. Master's thesis, Universidade Federal de Pernambuco, Brazil, 2015.
- [10] S. Dexter. eTIPS-Educational technology integration and implementation principles. *Designing instruction for technology-enhanced learning*, pages 56–70, 2002.
- [11] A. Dünser, R. Grasset, and M. Billinghurst. A Survey of Evaluation Techniques Used in Augmented Reality Studies. Technical report, 2008.
- [12] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian. *Guide to Advanced Empirical Software Engineering*, chapter Selecting Empirical Methods for Software Engineering Research, pages 285–311. Springer London, London, 2008.
- [13] A. Fitzpatrick. Analytical survey information and communication technologies in the teaching and learning of foreign languages: State-of-the-art, needs and perspectives. unesco institute for information technologies in education. Technical report, Moscow, Russia, 2004.
- [14] D. Fonseca, N. Mart, E. Redondo, I. Navarro, and A. Snchez. Relationship between student profile, tool use, participation, and academic performance with the use of augmented reality technology for visualized architecture models. *Computers in Human Behavior*, 31:434 – 445, 2014.
- [15] C. S. Jacelon and K. Imperio. Participant diaries as a source of data in research with older adults. *Qualitative health research*, 15(7):991–997, 2005.
- [16] A. M. Kamarainen, S. Metcalf, T. Grotzer, A. Browne, D. Mazzuca, M. S. Tutwiler, and C. Dede. EcoMOBILE: Integrating augmented reality and probeware with environmental education field trips. *Computers & Education*, 68:545 – 556, 2013.
- [17] V. M. Kensi. *Educação e Tecnologias: O Novo Ritmo da Informação*. Papirus, 2006.
- [18] L. Kerawalla, R. Luckin, S. Seljeftot, and A. Woolard. “Making it real”: Exploring the potential of augmented reality for teaching primary school science. *Virtual Real.*, 10(3):163–174, Nov. 2006.
- [19] N. N. Kostaras and M. N. Xenos. Assessing the usability of augmented reality systems. In *Proceedings of the 13th Panhellenic Conference on Informatics*, pages 197–201, 2009.
- [20] K. Lee. Augmented reality in education and training. *TechTrends*, 56(2):13–21, 2012.
- [21] A. N. Leont'ev. *Activity, consciousness, and personality*. Prentice-Hall, Englewood Cliffs, N.J., 1978.
- [22] M. Ortlipp. Keeping and using reflective journals in the qualitative research process. *The qualitative report*, 13(4):695–705, 2008.
- [23] E. Redondo, D. Fonseca, A. Snchez, and I. Navarro. New strategies using handheld augmented reality and mobile learning-teaching methodologies, in architecture and building engineering degrees. *Procedia Computer Science*, 25:52 – 61, 2013. 2013 International Conference on Virtual and Augmented Reality in Education.
- [24] R. Roberto, D. Freitas, F. Simoes, and V. Teichrieb. A dynamic blocks platform based on projective augmented reality and tangible interfaces for educational activities. *Journal on Interactive Systems, SBC*, 4(2):8–18, 2013.
- [25] G. Salvador-Herranz, D. Perez-Lopez, M. Ortega, E. Soto, M. Alcaniz, and M. Contero. Manipulating virtual objects with your hands: A case study on applying desktop augmented reality at the primary school. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 31–39, Jan 2013.
- [26] M. Santos, A. Chen, T. Taketomi, G. Yamamoto, J. Miyazaki, and H. Kato. Augmented reality learning experiences: Survey of prototype design and evaluation. *Learning Technologies, IEEE Transactions on*, 7(1):38–56, Jan 2014.
- [27] M. Scriven. Beyond formative and summative evaluation. *Evaluation and Education: A Quarter Century*, 1991.
- [28] M. Silva, R. Roberto, and V. Teichrieb. Evaluating educational system based on projective augmented reality. In *Anais do SBIE 2013*, pages 10 pp.–, november 2013.
- [29] M. Silva, R. Roberto, and V. Teichrieb. Evaluation of augmented reality technology in the english language field. In *Anais do SBIE 2015*, pages 577–586, november 2015.
- [30] H.-K. Wu, S. W.-Y. Lee, H.-Y. Chang, and J.-C. Liang. Current status, opportunities and challenges of augmented reality in education. *Computers & Education*, 62(0):41 – 49, 2013.