Augmented Reality Applications for K-12 Education: A Systematic Review from the Usability and User Experience Perspective

Authors

Effie Lai-Chong LAW*

School of Informatics, University of Leicester, UK <lcl9@leiceter.ac.uk>

Matthias HEINTZ

School of Informatics, University of Leicester, UK <mmh21@leiceter.ac.uk>

Abstract

In the past two decades, we have witnessed soaring efforts in applying Augmented Reality (AR) technology in education. Several systematic literature reviews (SLRs) were conducted to study AR educational applications (AREAs) and associated methodologies, primarily from the pedagogical rather than from the human-computer interaction (HCI) perspective. These reviews vary in goal, scale, scope, technique, outcome and quality. To bridge the gaps identified in these SLRs, ours is to meet fourfold objectives: to ground the analysis deeper in the usability and user experience (UX) core concepts and methods; to study the learning effect and usability/UX of AREAs and their relations by learner age; to reflect on the prevailing SLR process and propose improvement; to draw implications for the future development of AREAs. Our searches in four databases returned 714 papers of which 42, together with 7 from three existing SLRs, were included in the final analysis. Several intriguing findings have been identified: (i) the insufficient grounding in usability/UX frameworks indicates that there seems a disconnection between the HCI and technology-enhanced learning community; (ii) a lack of innovative AR-specific usability/UX evaluation methods and the continuing reliance on questionnaire may hamper the advances of AREAs; (iii) the learner age seems not a significant factor in determining the perceived usability and UX or the learning effect of AREAs; (iv) a limited number of studies at home suggests the missed opportunity of mobilizing parents to support children to deploy AREAs in different settings; (v) the number of AREAs for children with special needs remains disappointedly low; (vi) the threat of predatory journals to the quality of bibliometric sources amplifies the need for a robust approach to the quality assessment for SLR and transparency of interim results. Implications of these issues for future research and practice on AREAs are drawn.

Keywords: Augmented reality, Education, Usability, User Experience, Systematic Review

1 INTRODUCTION

Augmented Reality (AR) is a form of technology that superimposes 3D virtual objects or content in a real-world environment to create a sense of mixed reality [4, 64]. In the recent decade, AR technology has become increasingly sophisticated, advancing from conventional fiducial markers and location-based GPS (e.g. Pokéman GO) to sophisticated depth cameras (e.g. Google Glass, Hololens) to create richer interaction experiences. These technological advances have stimulated research efforts in various sectors, especially education, to harness the power of AR to transform the prevailing work.

In the ever-growing number of research studies exploring how AR applications could help realize specific educational goals, a plethora of design and evaluation methodologies has been employed. Understandably, many of these studies focus on their methodological approaches from the pedagogical perspective, such as applying the constructivist learning theories to develop AR-based learning materials (e.g. [13, 68, 97]) and employing the traditional pretest-posttest method to evaluate AR-induced learning effects (e.g. [27, 57, 96]). Despite the uptake of AR technology in education started only about two decades ago, several *systematic literature reviews* (SLRs) or survey[1] on **AR educational applications (AREAs**) have already been conducted,

[1] Some authors use the term 'survey' when their work actually follows the standard process of a systematic literature review (i.e. PRIMSA statement) whereas some authors seem to use the term 'survey' to signify that their work does not follow the SLR process. Other terms such as 'analytic review' and 'meta-review' are also used to indicate that the related review is non-SLR-compliant but still more comprehensive than a literature review typically performed as an integral part of a scientific publication.

albeit with varied quality. In a nutshell, an SLR aims to identify relevant research studies on a specific topic, analyze and synthesize constructs of interest systematically, thereby producing a broad as well as deep understanding of that topic and drawing implications for future research and practice [83].

The existing SLRs on AREAs address primarily their educational impacts rather than their usability and user experience (UX), which are critical qualities for determining the acceptance and adoption of AR as new teaching and learning tools. As usability and UX are the main concepts of this work, it is necessary to define them upfront here (cf. Section 2.2). Usability is a core notion in the field of Human-Computer Interaction (HCI) with a widely accepted definition documented in the standard ISO 9241-210: 2019, 3.13: "*The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*". Accordingly, an interactive system is usable when it can support its users to achieve their goals by completing related tasks with low or no error-rate, using optimal resources in terms of time and mental effort, and feeling satisfied with comfort. Otherwise, the design of the system is flawed with usability problems that undermine user acceptance. The notion of UX emerged when the HCI community had become aware of the limitations of the traditional usability paradigm. Moving beyond the non-utilitarian aspect of human-technology interactions, UX puts emphasis on user *affect* and *sensation*, and the meaningfulness of such interactions in everyday life [36]. In ISO 9241-210:2019, 3.15, UX is defined as "*user's perceptions and responses that result from the use and/or anticipated use of a system, product or service*." This broad definition seems to imply a subsumptive relation between usability and UX, though it is not explicitly stated in the standard. Aligning with the view of some but not all HCI professionals, we adopt the stance that usability is part of UX. Nonetheless, to accommodate the range of research studies with some addressing only the usability aspect of AREAs (e.g. user performance) and some covering the UX aspect as well (e.g. user emotion), we use both terms throughout this paper.

A handful of reviews studying usability/UX of AR-based applications in education and other domains are available. The survey conducted by Santos and five colleagues [79] covers the AREA research studies published in 2002-2012 with discovering usability issues being one of their review foci. The SLR carried out by Dey and colleagues [19] covers the related work published in 2005-2014 with education being one of the AR application domains. In their review of AREA publications in 2011-2015, Akçayır and Akçayır [1] briefly mentioned usability as a factor undermining the positive learning effect of AR. However, none of the three studies analyzed the usability issues systematically. Furthermore, two usability/UX-focused reviews on *non-education-specific* AR applications are available. The work of Swan and Gabbard [88] was probably the first endeavor of this kind. They argued for the need of user-based studies to advance the development and uptake of AR applications. Built upon the work of [88], Bai and Blackwell [6] conducted an analytic review to investigate usability/UX studies in the context of AR research in different domains other than education.

Overall, while the aforementioned reviews did provide some useful information on the usage of AREAs in general and their design and evaluation issues in particular, there remain questions to be answered: Which usability/UX core concepts are used to inform the design and evaluation of AREAs? How established usability/UX methodologies are employed to design and evaluate AREAs? Are any novel usability/UX methods and tools created to address AR features? What are the relations between the usability/UX and learning effect evaluated in the research studies on AREAs? Methodologically, these and other questions along this line of inquiry can viably be explored with an SLR.

To explore the aforementioned questions, we conducted an SLR on AREAs designed for learners in K-12 education (i.e. from kindergartens up to secondary schools). A key rationale for this inclusion criterion is that such end-users of AREAs are sensitive to usability/UX issues, which can undermine their acceptance of new educational technologies [69, 84]. This is particularly relevant as many of them are yet to develop skills to circumvent interaction issues arisen, which their older counterparts in tertiary education are more equipped to handle. Furthermore, we will examine whether and how perceived usability/UX and learning efficacy of AREAs vary with learner age.

Our SLR followed the well-recognized *Preferred Reporting Items for Systematic Reviews and Meta-analysis* (PRISMA) guidelines [67] and involved searches in four databases and existing SLRs. The process of identification, screening and filtering has resulted in a batch of 49 included papers (Section 3). In particular, when planning and implementing the process of SLR, we identified some limitations such as the lack of explicit guidelines for the quality assessment of the articles retrieved (Section 2.3) and the insufficient transparency of reporting intermediate results. We have introduced alternative approaches, namely,

employing publicly available citation count and h-index as *complementary* quality criteria, and using tree as and Venn diagrams (i.e. figures in Section 3) as *supplementary* reporting tools. Nonetheless, without the intention to claim that they are perfect solutions for the limitations found, we aimed to invite feedback from the wider research community on these attempts to improve the process of SLR.

Overall, the main research goal of our SLR is ***to gain data-driven insights into design and evaluation of augmented reality educational tools used by schools.*** This goal informs six research questions *(RQs):*

RQ1. Are there any discernible patterns of target groups, learning subjects and settings in deploying AREAs?
RQ2. What is the trend in hardware and software tools used for developing AREA over time?
RQ3. Which usability/UX frameworks, concepts, methods and tools have been used for the design and evaluation of the AREA?
RQ4. What usability/UX problems of AREAs have been identified and whether as well as how they have been addressed?
RQ5. What are the relations between usability/UX qualities and learning efficacy of AREAs?
RQ6. How are usability/UX qualities and learning effect of AREAs related to age groups?

Answers to these RQs are based on qualitative synthesis of the studies presented in the 49 papers included in our SLR, which is *not* a meta-analysis, as it does not rely on statistical approaches. A caveat is that our work does *not* aim to prescribe a set of usability/UX frameworks and methodologies that the research studies on AREAs should use. In contrast, we use an inductive approach to identify which concepts, models, methods and tools have been applied to gain an in-depth understanding about their potentials as well as limitations, thereby drawing implications for improvement. Overall, the contributions of our SLR are:

- ground the analysis deeper in the usability and UX core concepts;
- analyze the relation between the learning effect of AREAs and their usability/UX;
- examine how usability/UX issues of AREAs vary with age groups;
- reflect on the prevailing SLR process and introduce alternative approaches for improvement;
- draw relevant implications for future work on AREAs;

The rest of the paper is structured as follows: In Section 2, three strands of the related work are reviewed. In Section 3, a detailed description of the SLR methodology, which comprises three main stages – identification, screening/filtering, and synthesis, is presented. In Section 4, results are reported with Section 4.1 and 4.2 focusing on basic attributes pertinent to AREAs and Section 4.3 on usability and UX. In Section 5, the six research questions are answered with respect to the insights gained from the SLR outcomes, implications and limitations are discussed. The paper is concluded in Section 6.

2 RELATED WORK

Three strands of work are relevant to our realization of SLR: uses of AR in education (Section 2.1); core concepts of usability and UX (Section 2.2); how systematic review differs from scoping review (Section 2.3). While the relevance of providing the background for the first two strands is self-explanatory, reviewing the arguments for the unique characteristics of SLR, especially the notion of quality assessment, is pertinent as it informs our methodological decisions.

2.1 Augmented Reality in Education

While the coinage of the term "augmented reality" is generally credited to Tom Caudell and David Mizell in the early 1990s, precursors to today's AR technology were already created in the 1960s [9, 10]. In its 60-year history, many definitions of AR have been developed, and Azuma's [4] has been widely accepted [10]. Accordingly, AR is a technology with three core characteristics: (i) it combines real and virtual content; (ii) it is interactive in real-time; (iii) it is registered in 3D. The first characteristic differentiates AR from Virtual Reality (VR), which involves a total immersion of its user in simulated worlds, completely masking the real-world environment. Another term closely related to AR is Mixed Reality (MR) [63]. While AR and MR are increasingly used as synonyms, controversies and confusions prevail, which are attributable to the varied usage of these terms in the industrial and academic venues [87]. Nonetheless, according to the widely recognized taxonomy [63], MR refers to everything in the reality-virtuality continuum, including AR and VR.

In principle, AR-based applications are motivational as they allow learners to relate efficiently the contextualized information, be it visual, audio, haptics, taste/flavor or even smell [87], to the real objects upon which it is overlaid, resulting in a deeper understanding of the topic and thus better learning outcomes (e.g. [P074, P107, P220]). AR-based applications also enable learners to perceive and manipulate 3D representations of abstract concepts (e.g., complex molecular structures), which are usually harder to grasp when presented in 2D format (e.g. [P108, P205]). With enriched sensory experiences enabled by AR-based educational applications, which are mostly multimodal these days, interacting with them can elicit positive emotional responses such as fun and pleasure in learners, contributing to stronger learning effects (e.g. [Ex003, Ex004, W106]). Nevertheless, deploying such applications can have negative effects, including cognitive and sensory overload, frustrations arising from poor usability and other technical shortcomings (e.g. unstable GPS signals for location-based AREAs), considerable costs for equipment (e.g. head-mounted devices) and content development (Section 4.3).

To identify the status and trend of AREAs, a number of SLRs of various scale and scope have been documented, albeit some, strictly speaking, do not meet the requirements for being an SLR (Section 3). We identify nine eligible ones and summarize them in Table 1, characterizing each by the name of the first author, key research goal, number of papers reviewed, sources where literature searches are performed (total number of returns), range of publication years for the papers reviewed, and main findings.

Seven of the SLRs cover *all* educational levels from primary to tertiary whereas [79] covers the primary and secondary level and [26] is the only one that covers the pre-school level. Similarly, while most of the SLRs are non-domain-specific, [26] focuses on language and [44] on STEM. Note that we leave out some SLRs for the following reasons: a mix of VR and AR articles [23]; vague information about search strings or sources used [14, 90]. There exist other types of reviews such as a *meta-review* of the papers comparing student learning in AR and non-AR applications to identify factors influencing educational effectiveness [77] and a *scoping review* where no quality assessment of the papers identified from sources is indicated [85] (cf. Section 2.3).

Overall, there are three consistent findings on the educational effectiveness from these SLRs: the use of AR can result in learning gain to a moderate extent; increased motivation is the salient mediating variable contributing to the positive learning effect of AR; STEM is the most common domain for AREAs.

Three of these SLRs - [1, 19, 79] - investigated usability/UX issues, but the corresponding review results are more descriptive rather than analytic, lacking grounding in conceptual frameworks. [79] listed the usability evaluation methods employed in 24 publications (out of 87 with qualitative data) and concluded that questionnaire was the most commonly used method. [1] did not specify how many of the 68 papers reviewed had addressed usability/UX, and stated a general remark that usability issues could hamper student learning performance and experience. [19] extracted from each of the 42 papers the values for a set of attributes: data type (objective, subjective), displays used, dependent measures, study type (formal, field) and number of participants. Their main conclusion was that there was a variety of methods for evaluating usability of AR applications. In fact, the scope of education they defined was broad and might be debatable, for instance, AR-based rehabilitation training programs for patients (e.g. [2, 17]) could rather be classified as health applications.

The reviews of [6] and [88] are not listed in Table 1 as they are not education-focused. [88] selected 21 publications from four conference venues (i.e., ISMAR, ISWC, IEEE VR, Presence), covering the period of 1992-2004, whereas [6] analyzed 71 publications from one conference venue, ISMAR, covering the period of 2001-2010. One key contribution of [88] is their classification of three types of user-based studies: *perception* – involving low-level tasks to understand human perception and cognition in AR contexts; *performance* – studying user task performance within AR-based environments to understand the impact of AR on the task; *collaboration* - examining AR-mediated interaction and communication between users. This classification has been applied and expanded by [6] with the additional fourth type – user experience – analyzing subjective feelings. Further, [6] could be credited for their attempt to ground their work in UX, although a deeper analysis of the concept is warranted - something that we are going to address next.

Table 1: Systematic literature reviews on AR educational applications

Author*	Research Goal	Paper	Source	Year	Main Findings
Fan (2020) [26]	To provide an overview of the landscape of AR for early language acquisition, and practical and forward-looking knowledge of this area.	53	ACM, ERIC, PsycINFO, IEEE, WoS, ScienceDirect, Springer Link (*1247*).	2010-2019	The effectiveness of varied combinations of design and instructional strategies for AR applications for early language learning activities.
Garzon (2019) [30]	To analyze the impact of AR on students' learning gains and the influence of moderating variables.	46	WoS, Scopus, Google Scholar (*1342*)	2010-2018	The effect size on learning gains varies with the educational level (e.g. Bachelor: high) and with the field (e.g. Engineering: very large).
Pellas (2019) [73]	To lay the groundwork for improving school students' motivation and learning outcomes with AR-Game-based Learning (GBL)	21	JSTOR, ERIC Scopus, WoS, ScienceDirect, EBSCO, Wiley (not given)	2012-2017	STEM and marker-based AR are most common domain and type; Motivation and enrichment of learning experience were pillars of AR-GBL. Challenges for teachers to use the system and develop content.
Ibáñez (2018) [44]	To identify specific design features, instructional processes and outcome measures for AR apps for STEM learning.	28	ACM, ERIC, IEEE, WoS, Scopus, ScienceDirect, Springer (*1358*)	2010-2017	Most AR apps for STEM offered exploration or simulation. Few provided assistance for carrying out learning activities. Most evaluated conceptual understanding and affective outcomes.
Dey (2018) [19]	To provide a high-level view of how the landscape of user-based AR research has evolved.	42 on Edu-cation	Scopus (*1147*)	2005-2014	A wide range of AR usage and design; physical/cognitive training, gamified. A variety of methods for evaluating educational outcomes and usability.
Akcayir (2017) [1]	To fill the gap of the lacking literature review on the use of AR in educational settings.	68	WoS (*102*)	2011-2015	AR enhanced learning achievement. Usability and technical issues were noted challenges.
Diegmann (2015) [20]	To review benefits of different types of AR applications used in educational environments	25	IEEE, AIS, ProQuest, ACM, EBSCO, ScienceDirect (523)	2005-2013	14 benefits identified; improved learning curve and increased motivation were the most salient two.
Bacca (2014) [5]	To address the lack of review on analyzing factors for using AR in educational settings	32	5 SSCI and 4 SCI indexed journals (*not given*)	2003-2013	AR had reported advantages in various aspects, especially learning gain and motivation. Most studies on science and for Bachelor's level
Santos (2013) [79]	To measure the effect of AR educational content to show whether it is useful.	7 quan 87 qual	EdITLib, IATED, Wiley, InderScience, Sage, Springer ScienceDirect, TaylorFrancis (*503*)	2002-2012	A moderate effect size of student performance. AR supports real world annotation, contextual and vision-haptic visualization. The most common usability evaluation tool is questionnaire.

Note: * for brevity's sake, only the first author is cited here

2.2 Usability and User Experience (UX)

Usability, as a core HCI concept, has a history of about 40 years. The first scientific paper with usability in its title was published in 1979 [8], although some argue that the related ideas already emerged in the early 20th century [81]. In its history, a range of usability frameworks, methods, instruments and metrics has been developed [34, 54]. To name a few examples: Frameworks such as User-centred Design [33] and Usability Engineering Lifecycle [70]; Methods such as Think-aloud [22] and Heuristic Evaluation [71]; Instruments like System Usability Scale (SUS, [12]), Computer System Usability Questionnaire (CSUQ) [54], Technology Acceptance Model (TAM, [18]); Metrics like task completion rate, time on task, and error [43].

User Experience (UX) emerged around the turn of millennium, expanding the focus on cognition to acknowledge the crucial role of emotion when interacting with computing technology. Nevertheless, despite various attempts in the last decade (e.g. [31, 53, 66]), there is still a lack of universal definition of UX. Similar to the discussion on defining mixed reality [87], some HCI researchers and practitioners may not see the necessity or utility of having *the* definition of UX. This overly broad ISO definition of UX (cf. Section 1) comes with the two notes (cf. three in the 2018 version). Note 1 refines the ambiguous phrase '*users' perceptions and responses*' by referring to "*users' emotions, beliefs, preferences, perceptions, comfort, behaviours, and accomplishments*". Note 2 refines it even further with a comprehensive list of constructs, covering almost all aspects of human psychology and drawing close to the widely cited UX definition of Hassenzahl and Tractinsky [38]. Furthermore, in the earlier and current version of the ISO standard, no clarification about the relation between usability and UX is given. In the 2010 version, Note 3 suggests a relation by stating that "*usability criteria can be used to assess aspects of user experience*". However, for some reason, this Note does not exist in the current 2019 version. Overall, as mentioned in Introduction, we interpret that usability is subsumed by UX, as depicted in Figure 1. However, as this is *not* a consensual interpretation in the HCI community, we opt for the expression of usability/UX.

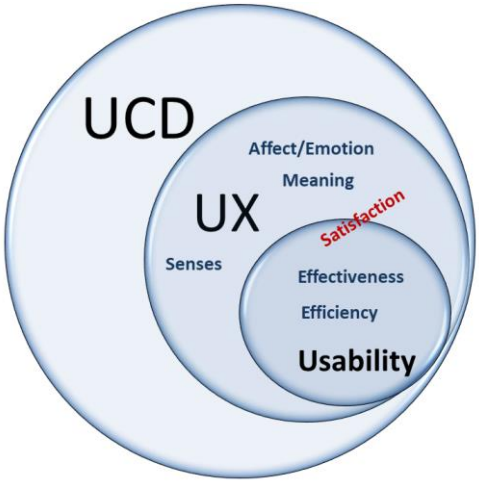

Figure 1. Relations between UCD, Usability and UX

Furthermore, a host of UX models, frameworks and methodologies, which are primarily derived from psychological research, is available. Frameworks such as Hassenzahl's pragmatic-hedonic model [35] and McCarthy and Wright's [62] sense-making experience (see [52] for an overview); Methods such as experience sampling [92] and psychophysiological measurement (e.g. [59]); Instruments like *AttrakDiff* [36, 38], Self-Assessment Manikin (SAM, [11]), UX Curve [49]; Metrics like positive and negative affective ratings. However, citing just these works does not do justice to the rich literature of usability and UX methods, which is documented in a number of projects and websites (e.g. usability.gov; allaboutux.org).

As acknowledged in the early as well as recent AR research work (e.g. [10, 28]), traditional HCI methods such as user needs analysis and tasks analysis are effective to determine *what* but not *how* content should be presented to users in AR. Usability/UX studies are essential for ensuring the quality of AR applications as they help identify design flaws in the early phases of development. In particular, user feedback allows insights into user expectations and preferences to improve the interaction design of AR applications.

Given the focus of our SLR, it is relevant to give a brief overview how usability/UX approaches have been applied in educational settings. However, we do not delve into this issue as it entails a separate publication. Generally speaking, usability/UX studies are conducted primarily for evaluating rather than designing educational tools [52]. Based on a recent analysis on evaluating the use of technology in education [50], one of the eight themes is usability/UX, involving constructs such as perceived usefulness, perceived ease of use, and affective responses (e.g. enjoyment, anxiety). Established scales such as TAM and SUS are instruments widely adopted in educational research [50, 93]. In contrast, applying UX design principles for educational technology seems nascent, based on the comprehensive review of Minichiello and colleagues [65]. Accordingly, the impact of UX approaches can be realised through designing educational experience (e.g. curricular innovation) and designing educational tools (e.g. interactive learning resources). Nonetheless, a

striking observation made by Minichiello et al [65] is that a general trend within the educational research literature is to ignore the methodological detail of UX design tool development. Hence, significant implications inferred are that the emerging UX education scholarship should be grounded deeper in the UCD work and that "*a base of methods-based knowledge*" (p.23) should be implemented to inform the development of more elegant and cost-effective UX-based approaches in education contexts. These insights resonate well with the findings of our SLR (Section 4.3) and associated implications (Section 5.3).

2.3 Systematic review vs. Scoping review

Systematic literature reviews (SLRs) are distinct from other types of literature review by their width as well as depth, and, a critical attribute, replicability [83]. The well-recognized PRISMA statement [67] is to ensure that the SLR process is replicable. Accordingly, the process is described in such a way that others can follow it and produce a comparable outcome.

The first and foremost stage of an SLR is to undertake a ***comprehensive search*** to identify relevant work on a specific topic. The search should be guided by an explicit statement of research goals and questions, which determine the selection of words and phrases to be used as search keys for identifying sources. It is recommended to perform searches with at least two databases [83]. The second stage is ***methodical screening and filtering*** of the search results based on a set of well-defined inclusion/exclusion criteria. On top of applying such criteria is ***quality assessment*** of the methodological rigor of individual studies. This makes an SLR distinct from a *scoping review* for which quality assessment is *not* considered as a priority or characteristic of its methodology [3, 15]. This view seems embraced by the research community, as observed by Pham and colleagues [75], who conducted a scoping review of 344 scoping reviews and found that quality assessment of individual studies was infrequently performed. While it is taken for granted that an SLR should account for quality, there is a lack of a standardized approach for such quality assessment. Several attempts to develop study quality tools or metrics were undertaken, for instance, a summary score over a set of attributes [46], but their limitations raise concerns and imply the need for improvement [95]. Sole reliance on either researcher-based subjective perceptions or machine-based objective parametric indicators is untenable [29, 91]. A combination of both may be more viable. We realised this integrated approach for our SLR (Section 3.2.4). As the last step of this stage, data extraction is performed to identify values of attributes pertaining to the research questions of the SLR, and output is usually tabled in a spreadsheet to facilitate subsequent work. Preliminary coding may be applied to some of the attributes.

The third stage is ***critical synthesis***, involving meticulous analysis and integration of the output from the second stage. Further data aggregation and coding are applied to determine whether and how the SLR research questions can be answered, including discernible new trends, discrepant findings across studies and plausible reasons, relations between theories and empirical evidence, significant gaps and limitations. Furthermore, there are two major types of synthesis: quantitative (meta-analysis) and qualitative. While SLRs are generally associated with meta-analysis, which entails the use of statistical methods to integrate quantitative information, reviews of qualitative information can also be replicable, rigorous and transparent in terms of methodology and be informative and insightful in terms of outcomes [83]. It is this type of qualitative review that our SLR aimed to produce.

3 METHOD

In the following we report the three stages of our SLR: Identification, Screening/Filtering, and Synthesis. While the two authors are the core contributors of all three stages, they have been supported in Stage 1 and Stage 2 by a research assistant and eight trained postgraduate students to realize the laborious process of the SLR. The first and second author have about twenty and ten years of research experience in HCI methodologies, respectively, and both are actively involved in exploring AREAs from the HCI perspective, as indicated by their partnership in an international project on AR interactive educational systems[2].

[2] https://www.areteproject.eu/

3.1 Stage 1: Identification

As stated in Introduction, the main research goal of our SLR is to gain insights into design and evaluation of AREAs used by schools. The details of the search are as follows:

Search string:
"Augmented Reality" AND ("Education" OR "Learning") AND "School" AND ("Design" OR "Evaluation")
Databases:

- Scopus: Document Search → Article title, Abstract, Keywords
- Web of Science (WoS) Core Collection: Basic search → Topic (article title, abstract, author keywords)
- ACM Digital Library Full-text Collection: Advanced Search → Title, Abstract, Author Keyword
- IEEE Xplore: Advanced Search → Document title, Abstract, Author Keyword

We constructed our search string with the following rationales. First, similar to the existing SLRs (e.g. [19], [30], [79]) we did not use the acronym "AR" as a search term based on the assumption that it should be spelled out in the title/abstract of an article. Otherwise, a huge number of false positive would be returned. Second, we did not use the term "mixed reality", which subsumes both AR and VR (Section 2.1); we assume that researchers use Augmented Reality as their primary term. Third, we used "school" as a broad contextual term to include school-related activities (in/outside classroom) and stakeholders (children, teachers, parents). Fourth, we did not want to restrict our searches by the term "usability" or "user experience". It is because some researchers could apply the related design and evaluation concepts, methods and tools without using the terms "usability" or "user experience" explicitly. Hence, we opted for a higher-level search substring ("Design" OR "Evaluation").

Concerning the choice of the databases, Scopus and WoS are two large bibliometric databases with a broad coverage of subject areas in sciences, social sciences, life and health sciences. ACM and IEEE cover subject areas in computing and engineering. Articles hosted in these databases are assumed to have been peer-reviewed. Nonetheless, this assumption is contentious (see Section 5.7 for the discussion on this issue). We did not choose Google Scholar due to various concerns, especially its limited advanced search function (i.e. searches in either article titles or whole texts), making the number of records returned overwhelmingly high of which a significant portion can be of very low scientific impact [60]. Furthermore, we did not include any unpublished studies based on the quality concern [30], although some SLR guidelines suggest otherwise [83].

The last set of searches was conducted on 1 July 2020. The initial searches resulted in altogether 714 records. Each record was assigned an identifier. As depicted in Figure 2, Scopus returned the highest number of records, followed by WoS. The total number of records having duplicate(s) in one or more than one of the other databases is 144. Only one instance of a duplicate record is placed in the source labelled "Overlap", but such a record can have two or more identifiers. None of the 144 records is a duplicate of all four databases. After consolidating the duplicates, 536 unique records remain. Interestingly, ACM returned only 18 records and 11 are duplicates of Scopus and 6 of WoS (see the top Venn diagram in Figure 5). It is worthy to point out that, to the best of our knowledge, many of the existing SLRs do not provide such details of overlaps across databases. However, we advocate that it is a good practice because it is relevant to know the distribution of resources to improve the integrity, transparency and reliability of an SLR. Furthermore, Cohen's kappa was computed to indicate the inter-coder reliability between the two authors at each exclusion level (Figure 5).

3.2 Stage 2: Screening and Filtering

The course of screening and filtering results in four progressively refined scopes. As depicted in Figure 2, the two outer circles contain the papers fitting the scope of design and evaluation to different extents, whereas the two inner circles narrow the scope of papers to usability and UX, with the innermost one (the lightest blue) meeting the strictest eligibility criteria for synthesis.

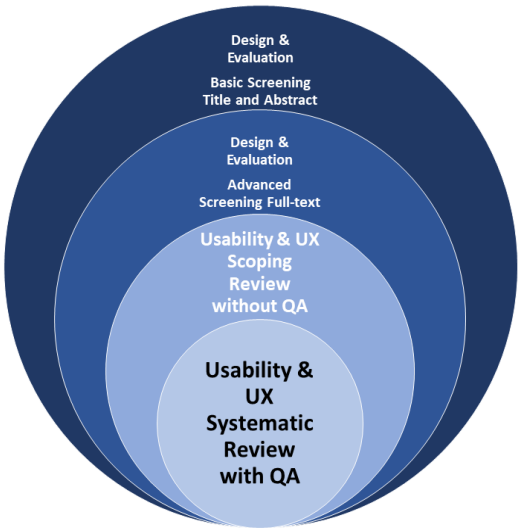

Figure 2. Four scopes of papers from the progress of screening and filtering.

3.2.1 Basic screening

The relevance of each of the 536 unique records was checked by the two authors and the research assistant trained for the task. Specifically, the title and abstract of each record was inspected to check for relevance by applying some of the inclusion/exclusion criteria (Table 2). Note that some of the criteria are applied (e.g. in3, in4, in5, ext4, ex5) when full papers are inspected (Section 3.2.2). This first screening filtered out 213 records for reasons such as the target groups were university students.

Note that the first two papers in this final batch – [P001] and [P007], authored by the same research group, had "virtual reality' in the title but "collaborative augmented reality system" in the abstract, and the work was actually about AR not VR. The mixed use of the terms can be attributed to the nascent stage of the AREA research in the early 2000s when the two papers were published.

Table 2: Inclusion/Exclusion criteria for basic screening

Inclusion Criteria	Exclusion Criteria
in1) The design and/or evaluation of the AR application is aimed to serve an educational goal(s); in2) Target group is from pre-school up to secondary schools (pre-university); in3) Access to full-text; in4) Essential information about the AR application and methodological approaches is provided; in5) Peer reviewed	ex1) Target group is from post-secondary institutions; ex2) Theoretical or review-focused (e.g. SLR); ex3) The term 'augmented reality' mentioned while actually virtual reality is used; ex4) Written in non-English; ex5) Insufficient information is provided about the AR application or methodological approaches

3.2.2 Advanced screening

The 323 papers retained after the basic screening were further screened in full text by the two authors and the research assistant with the use of the inclusion/exclusion criteria (Table 2). As a result, 98 papers were eliminated with the major reasons being inaccessible full text (in3), literature review only (ex1), VR instead of AR (ex3), and non-English (ex4). This left 225 papers for subsequent analysis. The distribution of the papers filtered in and out over the four databases and overlap is shown in Figure 3 (Yes vs. No). A *data extraction scheme* was developed by building upon the first author's previous SLR work [51] and some of the aforementioned SLRs, including [30, 73]). The scheme was used to pull out relevant information from individual papers and its attributes are listed in Table 3.

Eight postgraduate students were trained to carry out the data extraction task. First they were paired to apply the scheme to an initial set of five papers and discussed the results thoroughly with one of the two

authors as their trainer. The discussion helped identify their misunderstanding of the scheme. Each student was then given a batch of thirteen papers to analyze individually. The results were checked by one of the two authors, who fed back comments to the students. This exercise covered almost half of the 225 papers. The remaining data extraction was carried out by the two authors and the research assistant. Note that the criteria in4/ex5 (Table 2) were more applicable in this process of data extraction. Papers with low-quality information were marked for prudent consideration under quality assessment (Section 3.2.4).

Table 3: The data extraction scheme

High-level Attribute	Low-level Attribute	
Paper information	identifier, author, title, publication year, source	
Basic	domain, research goals/questions, theoretical framework	
Methodological approaches	*Context*	activity, setting, hardware, software
	Participant	target group, special condition of participants, participant age range, sample size
	Data	method, data collection instrument and data type, data analysis instrument and data type
Results	challenges, perceived quality by learner, perceived quality by educator, effectiveness for learner, effectiveness for educator	
Miscellaneous	Comments	

3.2.3 Usability and UX without Quality Assessment

For identifying papers addressing usability/UX of the AREAs, we examined the 'research goals/questions' (Table 3). The majority (128 out of 225) are pedagogical in nature *only*: design, develop and/or evaluate AREAs to validate if they can enhance specific knowledge and ability in learners *without* any usability/UX goals. Examples are: to foster problem-solving skills in children through an intelligent AR game [40]; to design and implement an AR-based inquiry courseware [56]. Of the remaining papers, 47 address both pedagogical and usability/UX goals and the other 50 the usability/UX goals only. An example of the former is to identify the benefits of AR technologies in science education and evaluate technological usability and learners' perception about the system [P167] whereas an example of the latter is to evaluate the usability of the remote collaboration with the enhanced AR system [48].

In accordance with the defining characteristics of scoping review (Section 2.3), the batch of 97 usability and UX papers is eligible for synthesis (i.e. the third inner circle of Figure 2). Nonetheless, to allow a synthesis to base on papers of a higher standard, the process of quality assessment is recommended. But how such a quality assessment should be implemented remains controversial (Section 2.3). While many of the existing SLRs seemed dependent on the subjective assessment of researchers, a few used bibliometrics (e.g., [5, 19]).

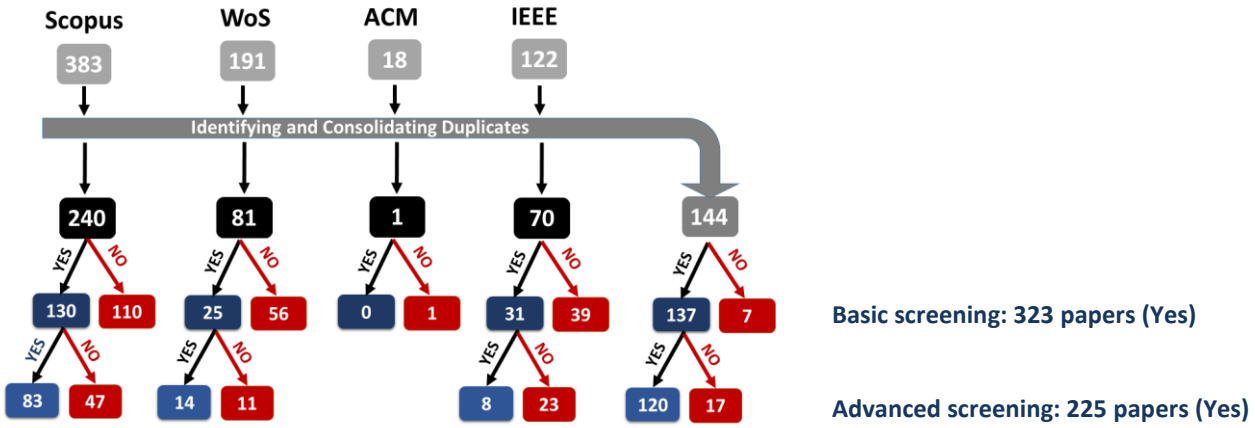

Figure 3. The results of basic and advanced screening stage.

3.2.4 Usability and UX with Quality Assessment (QA)

We employed two measures - Google Citation Index (GCI) and h-index provided by Scimago[3] Journal Rankings (SJR) - to support us to make informed decisions on including papers in the final batch for synthesis. Although the two metrics are increasingly used to indicate paper impact, they are known to have limitations such as accuracy of citation count and temporal instability of h-index [58, 72]. Nonetheless, the indices can still be useful for triangulating our assessment.

For h-index, [5] used Google Scholar Metrics h5-index for the category of "educational technology" to select top five journals as their SLR sources (Table 1). However, we did not use h-index-based rankings but we looked up h-indices for individual sources (i.e. journals, conference proceedings). The advantages of SJR over Google Scholar Metrics are that it supports free-text searches (cf. category-based lookup in Google) and covers conference proceedings (cf. journals only in Google). A caveat is that the SJR h-index information for conference proceedings is incomplete; for some it is presented as an overall value (not-year-specific) (e.g. *Conference on Human Factors in Computing Systems* has h-index of 177) and for others it is year-specific, albeit with missing years (e.g. *IEEE Virtual Reality Conference 2015* has h-index of 12). If h-index for a specific conference year is missing, the closest one available after that year is used.

For citation index, [19] used GCI to compute so-called Average Citation Count (ACC) of a paper by dividing total lifetime citation by lifetime (years), and they arbitrarily chose 2.0 as a threshold for filtering in (above)/out (below) a paper. In contrast, we applied year-based and type-based calibration (Table 4).

Table 4: Google citation indices (GCI) and Scimago h-indices for quality assessment

GCI	h index	Type	ID	Status	GCI	H index	Type	ID	Status
17	7	J	P123		5	U	B	P122	
53	38	J	P128	I	18	U	C	P125	x
41	84	J	P131	I	16	8	C	P129	x
38	44	J	P134	I	2	7	C	P130	
36	38	J	P136	I	7	8	C	P132	I
fU	U	J	P140		19	12	C	P135	I
187	164	J	WOS106	I	1	U	C	P137	
10	3	J	WOS116		3	2	C	P141	
38	44	J	WOS117	I	13	8	C	P142	x
15	12	J	WOS124		0	2	C	P158	
					10	64	C	WOS104	I
					3	9	C	WOS107	
					20	12	C	WOS123	I
					14	19	C	IEEE002	I

Note: B = Book chapter, J = Journal article, C = Conference article, U = Unknown index

Specifically, for each of the 225 papers retained after Advanced Screening (Section 3.2.2), their GCI and SJR h-index were found. Take the batch of 24 papers published in year 2016 as an example (cf. Figure 6). In Table 4, the GCIs and h-indices of 10 journal (J) and 13 conference (C) papers are listed[4]. The median GCI and median h-index are both 38 for the type J, and 10 and 8, respectively, for the type C. Median rather than mean is used as a threshold, because the ranges are large (GCI-J: 10-187; h-index-J, 3 -184; GCI-C: 0-24; h-index-C: 2-64). Those papers scoring equal or higher than the threshold are considered as quality-marked. We first applied this rule of thumb to the papers' GCI and changed the status of those passed the threshold as included (I). Then we checked h-index of those papers with GCI lower than the threshold. In case their h-index passed the threshold then they were included. This happened to [P136] whose GCI was 36 (lower than the median) and h-index was 38 (equal to the threshold). The assumption is that a quality paper published in a good venue might be overlooked by the research community and became under-cited. Conversely, one could argue that a quality paper published in a poor venue could also be under-cited, but presumably this happens less often. The Spearman rho's correlation coefficient between the two indices (N = 184, r = .415, p<.001) is positively significant. N is lower than 225, because h-index cannot be found for a number of conference proceedings; another justification for the prioritized use of GCI. Note that the status of some

[3] https://www.scimagojr.com/index.php

[4] We exclude a book chapter (B) from quality assessment, because the related peer review process is different from that of journals or conferences (NB: only three instances of book chapter among 225)

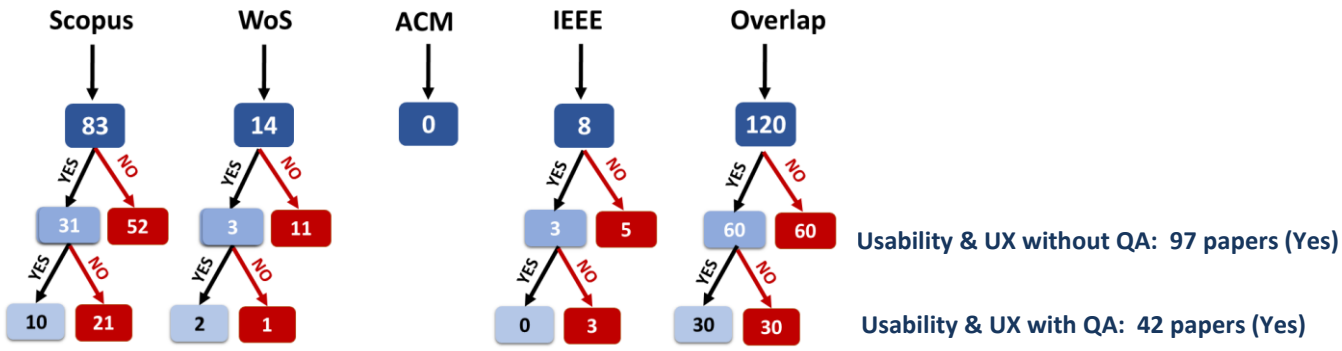

Figure 4. The results of filtering for usability and UX with/without quality assessment (QA)

conference papers are classed with "x" (i.e. P125, P129, P142). While they are above threshold, it was already noted in the Advanced Screening (Section 3.2.2) stage that they have some quality concerns such as lack of relevant details about empirical findings. The process of quality assessment concluded with 41 papers for synthesis (Figure 4).

As shown in Table 1, three of the existing SLRs on AREAs [1, 19, 79] studied the usability/UX aspect as a part of the review (Section 2.1). From the list of papers included in these SLRs, we identified papers that were duplicates of our batch and also unique ones, given different scopes and databases used, 7 of which meet our criteria. With these extra papers, we have altogether 49 (= 42 + 7) eligible for the SLR (Figure 5).

3.3 Stage 3: Qualitative Synthesis

In addition to the data extraction process as described in Section 3.2.2 (Table 3), the final batch of 49 papers were further analysed with the following coding scheme (Table 5), which consists of two major dimensions – Methods and Data, Results and Follow-up – and attributes. The scheme was developed by the two authors based on their expertise and experience in usability/UX work. The information coded was synthesized to identify patterns and insights (Section 4).

Table 5: The coding scheme for usability and UX articles included in the SLR

Usability/UX: Methods and Data ▪ *Usability/UX Frameworks*: HCI theoretical or methodological framework that supports or informs the usability/UX design and evaluation work in the paper; ▪ *Scope*: The usability/UX work covers design or evaluation or both aspects of AR app development; ▪ *Design Goals*: Usability/UX (e.g. specific user interface elements, specific emotions) aspects were designed for; ▪ *Evaluation purpose*: Formative (diagnostic to get improvement ideas), Summative (towards/after the end of the AR app development); ▪ *Research protocol:* ○ Established: well-recognized method is used; ○ Adapted: modified an established method; ○ Novel: new method; ○ Loose: a generic method is mentioned without any detail (e.g. interview without listing the questions); ○ Mixed: a combination of above ▪ *Informant*: Who provided the usability/UX data; ▪ *Data type*: Type of data collected to infer the usability/UX level; ▪ *Data collection instrument*: Provide details, including names, sources, and psychometric properties, if available; ▪ *Data analysis techniques*: Statistics, content analysis, and other alternatives; comment on the appropriateness with regard to data types.
Usability/UX: Results and Follow Up ▪ *Overall results*: High/low usability, Positive/negative UX, or both; ▪ *Detailed descriptions*: Usability/UX problems identified; Specific emotional experiences reported ▪ *Relation with Learning Effect:* Report both Usability/UX and Learning Effect? If yes, Independent; if no, Positive or Negative relation; ▪ *Mediating variables*: If the relation is positive/negative, the variable(s) contributing to it; ▪ *Responses*: Follow up actions to address the findings related to usability/UX, if any;

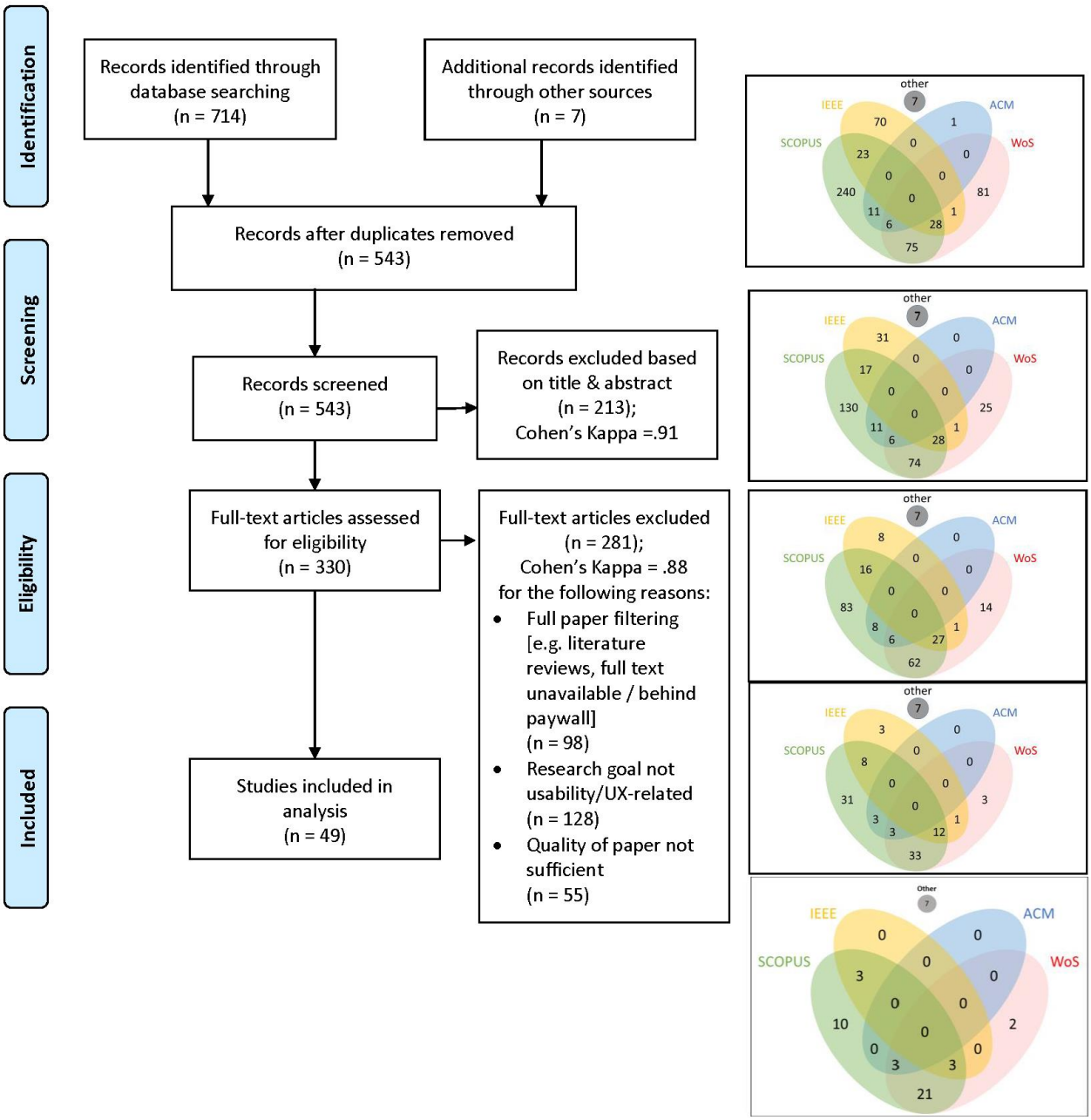

Figure 5: PRISMA Statement for the SLR

4 RESULTS

Given the focus of this SLR, we mainly present the synthesis results based on the 49 papers included in the final batch of usability/UX with quality assessment (i.e. the innermost circle of Figure 2). The above detailed descriptions of the three-stage process are to ensure the transparency and replicability of our SLR. In the following, we first present the results about the basic attributes of the papers, including year/source of publication and application domain (Section 4.1), followed by the results on the contextual attributes, including hardware and software, age groups, conditions and settings (Section 4.2). These two sections serve as a significant backdrop for understanding the results on usability/UX (Section 4.3).

4.1 Patterns of Basic Attributes

4.1.1 Papers by year

In searching the four databases (Section 3.1), the earliest publication year of the records returned is 2000. Figure 6 illustrates the changes over time in last twenty years. While the increase was gradual in the first decade (2000-2009), it was more rapid in the second decade (2010-2019) with a visible jump from 2017 to 2018. Nonetheless, the percentage of the papers (orange line vs. blue line) addressing usability/UX as (part of) their research goals was fluctuating over years with no discernible patterns, for instance, 71% in 2008 dropped to 17% in 2012, climbed to 68% in 2017 and down again to 23% in 2019.

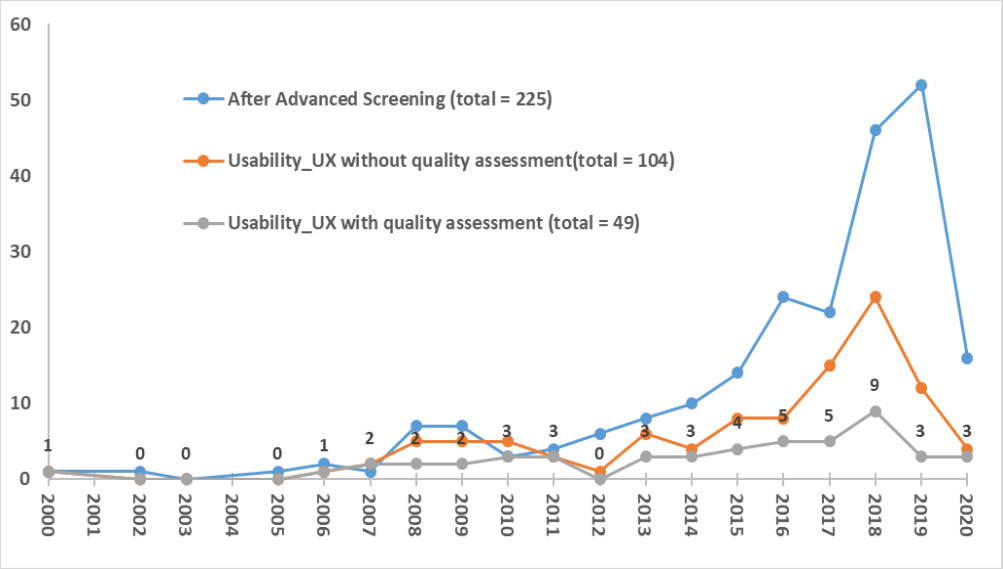

Figure 6. Distribution of papers over years for the three screening results

4.1.2 Papers by sources

The papers were published in three types of sources: journals, conferences and books. We categorised them by seven disciplines, which inevitably overlap to some extent (Table 6). Out of the 49 papers 32 are sourced from journals. Given our focus on usability and UX of AREAs, it is not surprising that Education Tech is the most frequent category, followed by HCI.

Table 6: Distribution of the SLR papers by sources

	Design	Education Tech	Engineering & Comp. Sci.	Entertainment & Games	HCI	Science	VR	Subtotal
Journal	1	*23*	1	1	5	1	0	*32*
Conference	0	4	3	2	5	2	1	*17*
Subtotal	*1*	*27*	*4*	*3*	*10*	*3*	*1*	**49**

4.1.3 Paper by application domain

The range of application domain of AREAs as described in the papers is broad (Table 7). We categorised them at the subject level and then clustered them to three major domains of which STEM, subsuming seven subjects, is the largest with 57% (28 out of 49 papers). The subject "Integrated Science" is referred to general science education for primary school level when the division of biology, chemistry and physics is not yet in place. Maths, mostly geometry, proved a popular subject, given the power of AR for 3D visualisation. Language learning is another popular subject where AR is typically used to visualise learning scenarios, enhancing the motivation. The subject "Common Knowledge" is referred to the integrated study at the primary/lower secondary level, exploring basic scientific, social and civic topics. The subject "Cognitive and social skills" covers topics like creativity, computational thinking, memory management, emotional intelligence and symbolic play.

Table 7: Distribution of the SLR papers by application domain

STEM							Humanities		General Knowledge & Skills			
Bio-logy	Chem-istry	Phy-sics	Integrated Science	ICT	Environ-mental Science	Maths	Language	History	Common knowledge	Cognitive & Social Skills	Physical Edu.	Art & Design
5	3	5	6	1	2	6	7	3	6	2	2	1

4.2 Patterns of Contextual Attributes

4.2.1 Pattern in Hardware

Different types of hardware were deployed in the AREAs as described in the papers reviewed. Typically more than one type were used in the individual studies, accounting for a total larger than 49. As indicated in Table 8, the trend corroborates the argument that the increasing use of mobile devices has contributed to the rise of AR applications.

Table 8: Distribution of the hardware used in the SLR papers

Mobile Device	Computer /Laptop	Webcam	HMD	Tracker	Large screen	Custom made
27	14	14	8	4	3	2

By 'Mobile devices', we refer to phones and tablets. For the category of 'Custom made', it refers to the technical setup where the researchers integrated different hardware components, such as displays, cameras, projectors, headsets and scanners, in specific ways to address their research questions (e.g. Spinnstube; [74]). Salient examples of the categories 'HMD' (head-mounted display), 'Tracker', and 'Large screen' are Hololens,

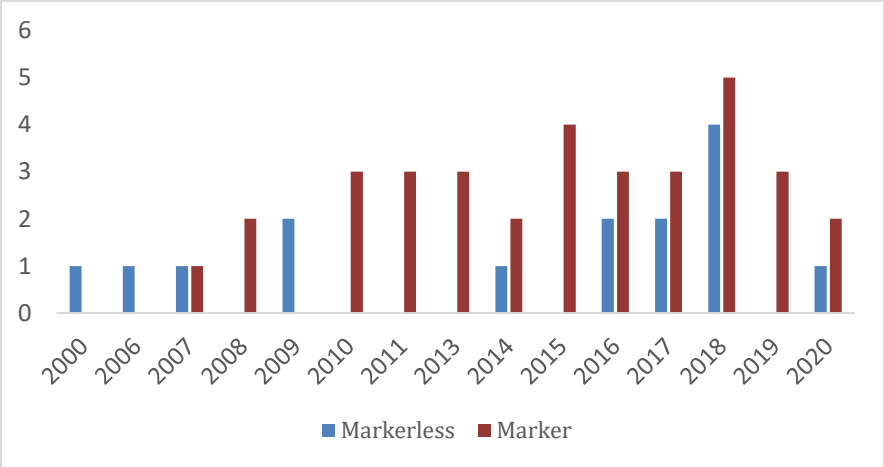

Figure 7. Distribution of marker-based and markerless AR educational applications

Kinect and smart TV, respectively. A discernible trend is the miniaturization of the hardware components to improve the portability and usability of AR applications. An intriguing observation is that the number of marker-based AR applications has been consistently higher than their marker-less counterparts (Figure 7). One plausible reason is that the reliance on GPS to support outdoor marker-less AR experience, but it is hard to ensure the stability and precision (high resolution) of GPS. Another marker-less setup is mid-air gesture-based interaction such as Kinect, but the need of equipment might hamper its adoption. In contrast, markers are easy and economical to produce, for example, with the support of a tool such as Vuforia, and everyday objects can be used as markers (e.g. [P179]), thereby fostering natural interaction and immersive experience.

4.2.2 Pattern in Software

Similar to our analysis of hardware used in the AREAs, we identified six categories (Table 9). Surprisingly, almost half of the papers did not provide any information on the software used to create their applications. Many of the studies deployed multiple software tools; among them, Vuforia and Unity are common and often used together. Examples of '3D modelling software' are Blender, Google SketchUp, and 3DS MAX. The

category 'Frameworks/Toolkits/Libraries' includes tools for low-level programming support, such as Android SDK, ARCore SDK, Open Inventor toolkit, OpenGL, NyArToolkit, Wikitude SDK, Java Media Framework (JMF), Google Maps API, OsgART library, Open Scene Graph, and Cubase SX. Examples of 'Existing AR software (customized)' mentioned are Studierstube, ARIS editor and app, Aurasma, TraceReaders AR platform, PapARt system API, TaleBlazer, Junaio, Lightning Studios & Sketch. For 'Asset editing software', examples are Windows Movie Maker, Adobe Photoshop, Audacity, clip studio paint, and Adobe Premiere. Overall, there seems no discernible trend in the software tools deployed.

Table 9. Distribution of the software tools used in the SLR papers

Vuforia	Unity	3D Modelling Software	Frameworks/ Toolkits/ Libraries	Existing AR Software (customized)	Asset editing software	Not specified
9	8	6	14	8	1	21

4.2.3 Pattern in Target Group Age

Our SLR focused on the research studies with K-12 as target groups (Section 1). We applied the *International Standard Classification for Education* (ISCED) 2011 scheme, which defines different levels without specifying associated age ranges. With reference to different educational systems, we identified the respective ranges of the ISCED levels, as shown in Table 10, which clearly indicates that the majority of AREAs reviewed were for Level 1 (Primary Education).

Table 10: Distribution of target age groups in the SLR papers

ISCED Education Age range	Level 0 Early Childhood 4-5	Level 1 Primary 6-12	Level 2 Lower Secondary 13-16	Level 3 Upper Secondary 17-19
Count	0	32	14	6

As three of the papers [P009, P205, Ex006] covered two age groups, the total count is 52 rather than 49. Furthermore, as depicted in Figure 8, the sample size of the empirical work tended to be moderate with 16 studies having 30 to 49 participants. There were a handful of studies involving more than 70 participants.

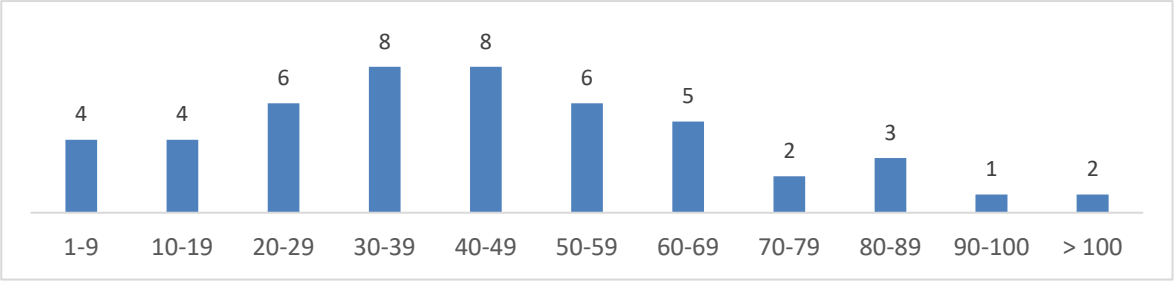

Figure 8: Distribution of sample sizes in the SLR papers

4.2.4 Pattern in Target Group Condition

The existing SLRs (Table 1) reported that very few research studies of AREAs focused on target groups with special needs. We corroborate this observation with the batch of papers we reviewed. Among the 49 papers, only one targeted students with physical disabilities to learn science [P019] and one on autism [P256]. In other words, only about 4% of these AREA research studies addressed students with special needs. This is the issue worthy to investigate which factors contribute to the low rate of application.

4.2.5 Pattern in Settings

We categorized the settings where the AREAs were deployed into four major groups: in classroom (n = 33), outdoors (n = 13), museum (n=5) and at home (n=2). A handful of studies involved more than one setting (e.g. in classroom & museum, [P019]; in classroom, museum, and at home, [P128]), accounting for the total

of 53. Most of the studies took place in classroom where the control of the learning activities and infrastructure (e.g. mobile devices, the internet connectivity) tended to be more manageable than outside classroom. Indeed, the studies taken place outdoors, including playgrounds within a school premise and field trips, faced different challenges such as low GPS accuracy [P018], poor visibility [P167], and bad detection of nature objects used as markers [P179].

In the two studies involving parents at home some intriguing results were reported. Specifically, in [P128], the participants were asked to exercise their self-directed learning skills to complete a worksheet on the topic of weather at home using the AREA referred to as *Manipulative AR* with their family. The authors commented that "[*t]he MAR system was not easy to use at home*" (p.218) based on the participants' rating on an item in a post-intervention survey, but they did not provide any underlying reason. In [P030], based on the observational data, the authors remarked "*instances of enhanced learning opportunities*" using the given AREA were more prevalent at school than at home (Note: no learning effect was measured), although the same interaction challenge due to the lighting and camera angles occurred in both settings. They also commented that more human interactivity between the participating children and their teachers as well as peers was observed at school than between the children and their parents as well as siblings at home. They argued that generally teachers are trained professionals to provide learning opportunities whereas parents are not and thus lack confidence to do so, despite having interest.

4.3 Usability and UX

In Table 5 (Section 3.3), we list the attributes for analyzing the usability and UX work described in the papers. In the following we present the main findings.

4.3.1 Usability and UX Frameworks

A framework is the structure that can hold or support a theory or methodology of a research study. Here we refer to the HCI theoretical and methodological frameworks underpinning the usability/UX work reported in the papers. Interestingly, only 16 out of 49 papers referenced such a framework. The User/Human-centred Design (UCD/HCD) approach is most frequently mentioned, albeit still low with only six papers (Table 11); the authors referenced the related ISO standards (9241-11:1998/9241-210:2010/9126-2:2003) and Empathic Design, which is built upon HCD and related to the notion of UX [61]. Only in one paper was Participatory Design (PD) described explicitly as a co-design process [94]. These frameworks put human needs, preferences and feelings at the core of designing interactive technologies. Despite their relevance, it is puzzling to see their low usage. Applying the principles of Usability Engineering [71] such as consistency and simplicity was mentioned in two papers. Two papers stated the use of Technology Acceptance Model (TAM) [18], which originated from the field of Information Science and shares constructs with usability and UX such as perceived ease of use and perceived enjoyment [42].

While integrating Cognitive Load Theory [89] into designing user interface and evaluating usability has been realized for decades (e.g. [41]), making an explicit reference to the theory is not common, as indicated by the observation here. On the other hand, with the emphasis on the experiential aspect of human-technology interaction, psychological frameworks such as Emotion Theory [76] and Flow Theory [16] have attracted more research interest. Furthermore, the notion of Tangible User Interface (TUI) is to give physical form to digital information, enabling human users to utilize their natural ability to grasp and manipulate the form to facilitate learning through digital technology [45]. This framework underpinned the emergence of Tangible AR [Ex002]. A recent interaction paradigm called World-as-Support (WaS) has been deployed in two papers by the same research group [P232]. Accordingly, it utilizes the power of context-aware recognition to project dynamically retrieved digital content onto real-life objects. In this way, the world serves as a support for meaning-making to enable positive user experience (cf. [62]).

Admittedly, it is hard to decide whether usability/UX frameworks have been implicitly applied, given that some constructs and instruments used in the studies can be seen as based on some related frameworks. For instance, in [P201], the construct 'total immersion' was measured in terms of flow and presence. However, instead of discussing any pertinent theoretical frameworks, the authors referenced the questionnaires for measuring flow and presence. On the contrary, some studies referred explicitly to a particular framework, which, however, lost its trace in the empirical work and analysis. This reflects the phenomenon called "*fading traceability*" [52]. It denotes the situation where authors cite the framework in the front sections of their

paper, claiming that their work is based on it, but when coming to the description of the actual design or evaluation work the framework is not referenced or applied.

Table 11: Usability/UX Frameworks referenced in the SLR papers

Framework	Paper
User/Human-centred Design (UCD/HCD)	P001, P011, P019, P167, P256, P376
Flow Theory	P074, P303, P330
Technology Acceptance Model (TAM)	P048, Ex005
Usability Engineering	P001, P131
World-as-Support Interaction Paradigm	P231, P232
Tangible User Interface interaction	Ex002
Participatory Design	P223
Cognitive Load Theory	P303
Emotion Theory	W106

Nonetheless, among 33 papers that did not refer to any usability/UX frameworks, 12 did not mention any framework at all whereas the other 21 referred to a range of pedagogical frameworks underpinning the studies (e.g. Inquiry-based Learning). It is out of the scope of this paper to elaborate on such frameworks.

4.3.2 Scope, Goals and Methods

For 38 of the 49 studies, the usability/UX work was for evaluation only whereas the other 11 studies aimed to address both design and evaluation goals. Examples of the design goals are “attaining optimal emotional experience or flow state” [P074], “design for humor” [P131], “adapted design principles of mobile phone for good gaming experience” [P033], “coupling between physical space and mediated experience” [P201], "make the learning process more interesting and enjoyable" [Ex005]. Only one study [P223] explicitly described the use of Layered Elaboration [94] as a participatory design method to elicit user requirements. The core concepts such as ease of use, satisfaction, efficiency, fun, flow, and engagement were reported to underline the design and evaluation of the AREAs. Nonetheless, some of the concepts such as satisfaction were not explicitly defined or operationalized. Furthermore, 36 of the usability/UX evaluation studies were summative, 8 were formative, and 5 were both.

The variety of usability/UX methods employed in the 49 studies was small with questionnaire being the predominant one used in 35 studies, followed by interviews (n = 19), observation (n = 13), and focus group (n = 3). This pattern corroborates the findings of the previous SLRs (e.g. [79]). Furthermore, four studies reported analyzing interaction behaviors by using video recordings of participants when they were implementing the task scenarios with the AREA. Slightly more than half of the studies (n = 27) employed more than one method (e.g. combining questionnaire, interview and observation in [P030]) whereas fifteen, three, and one studies relied only on questionnaire, observation and interview, respectively. Surprisingly, only a few studies attempted non-typical methods: two studies [P231, P232] asked children participants to draw their interaction experiences with the AREAs and be interviewed to explain the drawings; in one study [P019] researchers deployed objective physiological measures (heart rate, eye strain) and subjective questionnaires (i.e. Comfort Rating Scale, the Borg Rating of Perceived Exertion Scale) to yield quantitative data for their formative as well as summative evaluation. [P019] was one of the 13 studies that collected only quantitative data whereas 14 studies collected only qualitative data and 22 studies mixed data.

As implied by the types of methods applied, the range of data collection instruments was small. Out of 49 studies, 24 used homegrown questionnaires, which were either created from scratch by the authors (e.g., [P128, P134]) or taken from a combined set of existing questionnaires. Only eight of these homegrown questionnaires were reliability tested with Cronbach’s alpha, which, however, is not a measure of validity. Furthermore, 13 studies employed standardized usability/UX questionnaires, such as SUS, NASA-TLX and User Engagement Scale (UES), and only four reported Cronbach’s alpha. Other methods like interview, observation and focus group were conducted in a loose manner without using standardized questions or templates.

4.3.3 Perceived Usability/UX

Participants, including learners and educators, in 31 out of the 49 papers, were reported to have positive perceptions of the usability and UX of the AREAs concerned. Comments on high usability, such as easy to use, easy to scale the AR model, easy to navigate, low cognitive load, and high level of satisfaction, were documented (e.g. [P167, P001, Ex007]). Positive emotional responses, such as fun, engaging, and playful, were often reported (e.g. [P107, P109]). On the contrary, two and sixteen studies had negative ([P128, P220]) and mixed (e.g. [P030, P179]) usability/UX perceptions. When breaking the results by learner age (Table 12) (NB: as mentioned in Section 4.2.3, three papers covered two age groups), the youngest age group (69%) tend to perceive the AREAs more positively than their older counterparts (64%/50%), though the differences are non-significant ($\chi^2_{(2)} = 1.044$, $p>.05$) (NB: the categories 'negative' and 'mixed' were combined as 'non-positive' for statistical analysis) .

Table 12: Overall Usability/UX perceptions by learner age

		Level 1: 6-12 years old	**Level 2: 13-16 years old**	**Level 3: 17-19 years old**	Total
Overall Usability/UX Perception	**Positive**	22 (69%)	9 (64%)	3 (50%)	34
	Negative	1 (3%)	1 (7%)	0 (0%)	2
	Mixed	9 (28%)	4 (29%)	3 (50%)	16
	Total	32	14	6	*52*

4.3.4 Usability Problems (UPs)

Usability problems (UPs) are indicators of design flaws and causes for poor usability/UX (Section 2.2). Whereas 25 of the 49 papers provided information on UPs, the other half did not have such information. It can be explained by the fact that 31 studies had the overall usability/UX perception as positive, and nevertheless eight of them also reported UPs. Furthermore, as we aimed to know if there were any age-related patterns, we identified the target age groups of the respective papers for individual UPs (Table 13).

There are altogether 12 categories of UPs, UP1 to UP12. The most frequent category of UP1 was AR-specific, namely, the design and usage of markers (Section 4.2.1). All age groups experienced UP1. Two other types commonly experienced by all age groups are UP4 (i.e. screen size) and UP7 (i.e. handling dual objects); these UPs could be associated with cognitive load. One type, UP6, was solely experienced by the youngest age group (6-12 years old) who found it difficult to manipulate, control and position 3D virtual objects. In contrast, two categories of UPs that were only experienced by the oldest age group (17-19 years old) were UP9 and UP11. This could be explained by the observation that the related studies were conducted in 2000, 2006 and 2009; the HMD was too bulky and expensive to be evaluated with younger users. Similarly, the issue of slow rendering was reported in the papers published in early 2000s when the work on AREAs started to take off. These UPs have been ameliorated with improved algorithms and more powerful computer processors. The UP12 on video sound quality could arguably be relevant, because the feature might contribute to the holistic user experience with the AREA concerned.

In analyzing whether and how these UPs were addressed within the respective studies, only in three studies [P109, P256, P376] did the authors report that the related UPs were handled with success. Specifically, in [P109], the changes included installing UI buttons on both sides of a tablet to facilitate controlling the AREA; providing a tutorial on 3D depth, amplifying perceptual cues (e.g., adding shadowing), and rendering visuals simpler. In case of [P256], simplifying the AR game mechanics and adding meaningful images to indicate the start position of the game resulted in an improved understanding of interaction design. For [P376], one marker and a menu supporting switches between solids to be visually augmented were deployed to replace multiple markers, and a pinch-to-zoom feature was also added.

Furthermore, in eleven studies, the authors presented some planned improvement actions as future work, albeit with different degrees of concreteness. Among them, four suggested adding a tutorial could resolve some UPs; one was more specific: "a short tutorial for introducing the device by a virtual friendly pet" [P033] whereas one simply wrote "a short tutorial". Some proposed generic actions such as "robuster tracking" [Ex005], "focus on simplicity" [P128] and "adaptive information density" [P320] whereas some had UP-focused actions. For instance, in [P134], the authors proposed using a road map instead of a satellite map to address the issue of poor map tile quality. In [P179], to address the problem of marker recognition the

authors suggested using computer vision and machine learning to identify nature objects rather than transforming an object in nature into a marker. The remaining studies acknowledged the presence of UPs without specifying any remedial actions.

Table 13: Category of usability problems identified in the SLR papers and distribution by learner age

Category of Usability Problems (UP)	**Papers**		
Age group	*6-12 years*	*13-16 years*	*17-19 years*
UP1 Marker-related: usage, detection, control, occlusion, transfer across contexts, objects in nature	P030, P108, P179, P223, P376	P320	P074
UP2 Perceptual quality of 3D virtual object: realism, visibility (outdoor), aesthetic design	P030, P134, P220,	P167, P175, Ex001	
UP3 Precision: misplacement of virtual objects (avatar), GPS	P009, P030, P018, P131,		P001, P007
UP4 Small screen size of mobile devices	P220	P013, P320	P001, P007
UP5 Software stability: crashes and rebooting	P018, P072, P131, P175,	Ex001	
UP6 Virtual object manipulation and control: gestural and hand recognition	P030, P108, P109, P256, P335, P376		
UP7 Dual handling of physical device and virtual object	P220, Ex005	P167	P074
UP8 Understandability: AR mechanism, User interface element	P030, P134, P109, P256,	P128	
UP9 HMD: weight, motion sickness			P001, P007, P019
UP10 Infrastructure setup: camera position and image projection on real-life objects;	P030, P256, Ex005		
UP11 Slow rendering			P001, P007
UP12 Sound quality of video	P220		

4.3.5 Relations between Usability/UX and Learning Effect

By learning effect, we refer to measures taken to assess the extent to which specific knowledge, skill or ability is changed as a result of learning with the AREA concerned. Out of 49 papers 25 reported that the experimental group performed as good as or better than the control group on specific topics, such as writing [P182], animal classification [P011], electromagnetism [P074], pedestrian navigation [P175], and collaborative skills [P330]. No negative learning effect was reported. The common measurement method of learning effect was pre-post knowledge/skill-specific tests. Another method was systematic observation. The other 24 papers did not report on learning effect, because either the main focus was on developing the application right from the interaction design perspective or the pedagogical outcome was informally assessed with anecdotal comments from some participants (e.g. [Ex003]) or researchers (e.g. [P030]). When breaking down the results by learner age (NB: three papers covering two age groups), as shown in Table 14, it seems that the two younger age groups benefitted more from the AREAs than their older counterparts, but this claim may be untenable due to the number of 'no measure'.

Table 14: Learning Effect by Learner Age

	Level 1: 6-12 yrs.	**Level 2: 13-16 yrs.**	**Level 3: 17-19 yrs.**	Total
No Measure	16 (50%)	7 (50%)	4 (67%)	26
Positive Effect	16 (50%)	7 (50%)	2 (33%)	26
Total	32	14	6	

Furthermore, 11 papers did not relate the results of learning effect and usability/UX, quantitatively or qualitatively. In other words, whether learners gained knowledge, skill or ability from an AREA seems independent of their perceptions and responses from interacting with it. However, it could be that the

authors just did not discuss the relation explicitly. Eight and six studies described the relations between the usability/UX findings and learning effect qualitatively with positive and negative characterization, respectively. Only one study [P098] computed the correlation statistically. Specifically, low insignificant correlations were found (Note: the exact *r* values were not reported) between the usability score measured by SUS and the post-test score (negatively correlated) and between the task load score measured by NASA-TLX and the post-test score (positively correlated), both cases were statistically controlled by the pre-test score. It would be intriguing to understand these rather counterintuitive findings, but the authors did not analyze them. The mediating variables mentioned for the positive relations were novelty of the tool, motivation, flow, presence, and instant feedback, whereas those for the negative ones were task difficulty, lack of engagement, and difficulty in marker manipulation.

5 DISCUSSION

In this section, we revisit the six research questions posed in Introduction (Section 1) by referring to our analysis and synthesis results presented above (Section 4) and to the comparable findings of the existing SLRs (Section 2.1), where appropriate. Implications are drawn for the future work on AREAs, especially from the usability and UX perspective.

5.1 RQ1

Are there any discernible patterns of target groups, learning subjects and settings in deploying AREAs?

Answer: We identified some discernible patterns. Our findings indicate that the major target group of AREAs was primary education for children aged 6 to 12 years old (Table 10). Note, however, that only 4% of the studies targeted learners with special needs. The learning subjects were predominantly STEM with mathematics being the most popular domain (Table 7). The AREAs were mostly deployed within the school premises (classrooms, labs, computer rooms). A quarter of the studies took place outdoors, which, however, faced different challenges such as low GPS accuracy and poor visibility caused by glaring sunlight. Museums are deemed as rich informal learning spaces to complement classroom-based teaching, and are inviting setting for deploying AREAs to enhance learning experience, but the number of studies was rather low. Similarly, there were two only studies with the 'at home' setting [P030] and [P128].

Implication: The trends of target groups, learning subjects and settings suggest that there are significant gaps to be bridged. First, it is necessary to provide parents with enough training to support their children to deploy AREAs at home, given the proven benefits of such educational technologies. This is particularly salient in the wake of the current pandemic when homeschooling has become essential. Irrespective of the recurrence of such a crisis, which hopefully will never happen again, children's self-directed learning in formal as well as informal settings should be fostered with scaffolding to be provided by informed teachers, parents and carers. In addition, the skewed focus on primary schools should be balanced by developing more relevant AREAs to support their counterparts in secondary schools. The relative low number of AREAs for the latter can be related to the increasing complexity of curricular content with educational level. Clearly, developing AR-based content entails knowledge and skillsets different from those for traditional learning materials. Usable and useful authoring tools that can facilitate teachers and parents to co-create contents with children can be viable options to address the observed gaps. Furthermore, the use of AREAs for learners with special needs (e.g. Down Syndrome) should further be explored. Although the number of studies (only 2) was small, they all suggested the potential of AREA in this regard, especially the game-based approach (cf. [P256]).

5.2 RQ2

What is the trend in hardware and software tools used for developing AREA over time?

Answer: As pointed out in the previous SLRs, the popularity of mobile devices contributed to the growth of AR applications in education. Results of our SLR corroborate this observation (Figure 6). Nonetheless, while flexible, the small screen of mobile devices (phones) caused usability problems (Table 13) whereas devices with large screens such as tablets caused fatigue as they were heavy. The number of studies using HMD remained relatively small; the high price of such equipment (e.g. Holoens) can be an inhibitive factor. Noteworthy is the continuing trend of using marker-based as compared with marker-less applications (cf. [73]). One plausible reason is the low production cost of marker generation. Another advantage of markers

is that as a built-in interaction device they enable users to manipulate the virtual scene by manipulating the markers, for example to see a model from different angles by just rotating the marker instead of moving around it. For instance, [P320] presented an AREA on chemistry: to see the reaction of two atoms, the learner just brought the respective markers close enough together; without the markers the developers would need to design and implement virtual controls. Nonetheless, from the usability/UX perspective, marker-based interactions still need improvement. As listed in Table 13, different usability problems (UPs) were related to the use of markers; the frequently occurring UP was on detection, especially for applications using nature objects as markers. With regard to the trend in software tools, the use of Unity and Vuforia has increased in recent years, which can be explained with the benefits they bring, for example making the development process easier and allowing the output to be deployed on Windows, Android, and iOS devices.

Implication: Despite the advantages of mobile devices for AREAs, the potential of HMDs can be explored. Nonetheless, their affordability is a significant barrier. Clearly, high-quality tablets and phones can probably lead to good usability and positive user experience, but they are also more expensive. This can especially become an issue in school settings where several devices, not just one, need to be acquired to allow individuals or small groups of students to experience an AREA. This budgetary concern may be eased by some joint private-public partnership. Furthermore, the design and development of markers entail further research efforts to address the UPs identified, especially the issue of lighting and dealing with low-quality camera. While the ongoing technological advances can mitigate some of the UPs identified, the critical resolution is to ground the design and evaluation of AREAs in sound theoretical and methodological frameworks – the issue to be discussed subsequently for RQ3 and RQ6.

5.3 RQ3

Which usability and UX frameworks, methods and tools have been used for the design and evaluation of the AREA?

Answer: Only one third of the 49 papers referenced usability/UX frameworks to ground their work in the design/evaluation of the AREA. The range of frameworks was rather narrow with most of them being established in HCI such as UCD or in psychology such as Flow Theory (Table 13). The relatively newer and arguably AR-specific framework was the World-as-Support interaction paradigm described in [P232]. As a substantial portion of the 49 papers is in the realm of Education Technology, one may argue that the prominence of usability/UX frameworks is overshadowed by psycho-pedagogical theories (e.g. Situated Learning; Cognitive Theory of Multimedia Learning), though one may also argue that their distinction can be blurred. With regard to the methods used, questionnaire remained the most commonly used one (35/49 = ~71%) (cf. [79]), and, surprisingly, was even the only source of empirical data in 15 studies. This single-method approach is contradictory to the recommendation for multimethod approaches to triangulate different data sources in usability/UX work (e.g. [53]). No innovative or AR-specific usability/UX methods could be identified in the papers reviewed.

Implication: The low reference of usability/UX frameworks exposes the gap between theoretical frameworks and practical designs, which is a recurrent topic in the field of HCI [52]. A basic approach to bridging the gap is through HCI education, disseminating theories as well as demonstrating their uses. Furthermore, the observation of low usage of HCI frameworks in the work on educational technology seems to imply that the communication and collaboration between two research communities – HCI and TEL (Technology-enhanced Learning) – has been sporadic. In fact, only in 2019 was the subcommittee '*Learning, Education and Families*' introduced to the CHI conference. While it is an encouraging move, more research events can be organized to bring people in HCI and TEL together. Furthermore, the lack of AR-specific usability/UX frameworks or methods may hinder the advance of the particular area. Several attempts along this line have been undertaken, including usability principles for AR in smartphone environment (e.g. [47]), AR design heuristics [24], and validated questionnaire for handheld AR [80]. Nevertheless, the adoption of the proposed approaches seems modest. More research effort is called forth.

5.4 RQ4

What usability and UX problems of AREAs have been identified and whether as well as how they have been addressed?

Answer: For the batch of the 49 papers reviewed, user-based evaluations of the AREAs relied on questionnaire, which might not collect data on the qualitative descriptions of usability problems (UPs). In fact, only half of the studies reported UPs with different degree of elaboration. The most frequently identified UPs were marker-related, which were also AR-specific. Nevertheless, in acknowledging the existence of the UPs identified, only a small number of studies proposed concrete improvement actions, implemented them and confirmed their effectiveness. The others mentioned some generic suggestions as future actions.

Implication: To improve the design of interactive technology, including AREAs, gaining insights from usability problems is critical. Nonetheless, it is necessary to document UPs systematically to enable researchers and practitioners to analyze the UPs and derive change recommendations. The lack of such data in many of the SLR papers implies that explicit guidelines should be given, encouraging authors to document as well as share UPs with the community in an open forum for analyzing UPs and proposing fixes. Overall, the AREA researchers should be enabled to conduct comprehensive usability evaluations of prototypes. This practice can be promoted through the collaboration between the two research communities – HCI and TEL - as discussed in RQ3 (Section 5.3).

5.5 RQ5

What are the relations between usability/UX quality and learning effect of AREAs?

Answer: By usability/UX quality, we refer to the overall quality perception rather than prototypical usability metrics or specific emotional responses (Section 4.3.3). By learning effect, we refer to the measurable changes in targeted knowledge/skill as a consequence of using the AREA of interest. Nonetheless, it is challenging to characterize the relations based on the papers reviewed, because 24 of the 49 studies did not report or measure the learning effect at all, and 11 of those reporting the learning effect, however, did not attempt to relate it to the usability/UX findings. Only one of the 49 studies, [P098], formalized the relation with statistical analysis (Section 4.3.5). Several studies described the relations qualitatively with positive as well as negative characterization.

Implication: It is surprising to note the relatively low number of studies attempting to measure the learning effect in empirical research on educational technology. It can be a methodological artefact of the SLR process as we included papers focusing on usability/UX. But it can also be attributed to the fact that the research on AREAs is emerging; many studies were still at the exploratory phase, aiming to substantiate the design of the application. In other words, the prototypes were not mature enough to be fully functional to produce desired learning outcomes. Nevertheless, we deem it recommendable to encourage authors to assess systematically the learning effects, usability/UX qualities, and their relations. This allows a meaningful synthesis of research findings to enhance the growth of this emerging area.

5.6 RQ6

How are usability/UX quality and learning effect of AREAs related to age groups?

Answer: Learners in the younger age group (6-12 years old) tended to have more positive usability/UX perceptions of the AREAs than their older counterparts. The younger learners also tended to experience more the UPs concerning the manipulation and control of virtual objects and the understanding of user interface elements. Furthermore, younger learners demonstrated more positive learning effect. However, none of these observations are statistically significant. Nonetheless, the age-related trends can be attributed to the changes in psychological developments in children [78], especially visual perception [82] and sensorimotor skills [32]. Specifically, figure-ground perception improves around 3 to 5 years old and stabilizes at about 7. Form constancy shows notable improvement at 6 years old and continue to grow till 9. Spatial relationships improves from birth till about 10 years old. For neurotypical children, visual perception abilities become mature at 18 years of age [82]. Hence, young children may have problems distinguishing moving objects from a background or AR affordances [P108]. Similarly, gross and fine motor skills evolves from birth till 4 and 7 years old, respectively [32]. Hence, in designing AR apps for kindergartners (4-5 years old), it is imperative to ensure that control and manipulation of virtual objects can tolerate a large range of interaction precision and speed.

Requiring children to exercise their underdeveloped skills/abilities to interact with a technological tool can make them feel frustrated, and is a common cause for UPs, as shown in our analysis (Table 15). If the goal of the tool is to foster such underdeveloped skills, development-theory-driven designs can enhance the effectiveness of the tool and gain insights into interaction behaviours so as to identify improvement suggestions. As reported in [78], motor manipulation and spatial cognition are pivotal for children's experience with 3D objects in AR designs. They also argued that attention, memory and logic are highly relevant, but they did not look into other critical aspects such as visual skills and emotions.

Implication: A sound theoretical understanding of children's psychological development, specifically how their physical, cognitive, emotional and social skills evolve with age, is imperative for designers of children's technology (including AREAs) to make informed decisions (e.g. [7, 98]. However, the space of AR designs for children is yet to be fully explored [78]. Apart from the above mentioned psychological factors, a host of demographical factors should be taken into account for designing age-appropriate AR tools. The big challenge is to build an integrated model orchestrating these multifaceted factors to guide AR designs, especially the quality empirical data remain scant and scattered - an observation corroborated by our SLR. Overall, there is a clear need for more development-theory-informed empirical studies on AREAs.

5.7 Reflection on the SLR Process

The process of planning and implementing this SLR (and presumably all other SRLs) has been laborious and resource-demanding, as it has been iterative rather than linear. Among others, one factor complicating the process is the varying quality of the papers in individual databases. Although Scopus and Web of Science (WoS) are widely regarded as high-quality bibliometric data source for academic research, the recent trend that predatory open access journals are indexed in reputable databases, including Scopus, WoS, PubMed and MEDLINE, is worrying (e.g., [21]) as it poses real threats to research integrity. In fact, such worry can be amplified by the statement issued by WoS[5] that "*While most of the journals in Web of Science Core Collection are peer reviewed, Clarivate Analytics does not keep a log or list of which journals do have peer review status. The burden of proof regarding peer review lies with the journal editor/publisher*." Unfortunately, this honesty and trust-based system is at stake when even editorial board members can be faked [86]. In our filtering process we identified a set of at least 20 very low-quality papers from an open access journal, which is indexed in Scopus, and discarded all of them.

This alerted us the strong need for quality assessment of papers retrieved from the databases. Consequently, we resorted to two community-based metrics: the citation index of Google Scholar, which is well-recognized for its coverage [33], and h-index of Scimago (cf. Section 3.2.4), although we are aware of their imperfections (e.g. [58]). Furthermore, we attempted to further increase the transparency of the process by extending the PRISMA flowchart (Figure 5). Specifically, we used Venn Diagrams to illustrate the overlaps of the four databases we queried, which can increase the traceability of the sources. Apparently, a higher number of databases makes the visualization task more challenging.

5.8 Limitations

Despite our dedicated efforts in searching the four databases, we cannot claim the exhaustiveness of the papers retrieved. We could have included Google Scholar, which is known to have the largest coverage but no quality control [35], the sheer number of results returned is prohibitive. It would entail a huge amount of workload to filter hundreds, if not, thousands of results. The lack of useful advanced search features in Google Scholar makes its usage for this SLR undesirable. We have not looked into unpublished literature such as academic theses or any other "grey literature" (cf. [83]) based on the criterion of peer-review. However, as discussed in Section 5.7, there seems no guarantee on peer review of individual work even it is retrieved from a reputed source.

While we aimed to identify an objective approach for quality assessment – a distinctive factor, using Google Scholar citation index and Scimago h-index has its limitations. It is particularly tricky for locating h-index for conference proceedings of particular years in Scimago. In addition, we used the year-specific median as a cutoff point for deciding on inclusion whilst also checking against our own judgement on the

[5] https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Explanation-of-peer-reviewed-journals?language=en_US

overall quality of individual papers. The choice of using median seems debatable, although our informal benchmarking showed that this could be valid. Nonetheless, we are inviting the community to discuss these open issues.

5.9 Future Research Agenda

In this section we summarise the insights gained from the SLR results in the form of a research agenda for future work on this research area – *Designing and evaluating AR tools for K-12 education from the human-computer interaction perspective*. Each of the agenda items involves three core stakeholders, namely, children, teachers and parents. Other stakeholders who can play a critical role in realising the research agenda are HCI specialists, software developers, curricular designers, and policymakers.

- To identify parents' and teachers' requirements for AREA authoring tools that will enable them to co-create AR contents with children;
- To understand challenges and resolutions for expanding the scope of AREAs beyond the mainstream education to special education and beyond predominant STEM contents to everyday skills;
- To study the potential and risks for young children to use HMD-based AREAs from the psychological, physical and ethical perspective;
- To build an integrated framework upon the usability/UX design principles and child development theories to inform the design of age-appropriate interaction mechanisms for AREAs;
- To develop innovative usability/UX evaluation methods to address the unique features of AREAs, enhancing their impacts on educational efficacy as well as educational experience.

6 CONCLUSION

The recent growth of research interest and effort in Augmented Reality, especially in the education sector, has inspired as well as motivated us to conduct a review of the published literature systematically. While clearly we are not the first (or the last) research group taking up this challenge, we aimed to bring specific contributions to this burgeoning area. In contrast to the existing SLRs, we endeavored to investigate the issues pertaining to usability and UX of AREAs by grounding the analysis in the relevant core concepts. It is surprising to observe that the number of studies referencing the established usability/UX frameworks or proposing new ones was limited. In fact, no novel AR-specific usability/UX methods were identified in the papers included in our SLR. Furthermore, the observation that types of usability problem are associated with learner ages calls forth stronger development-theory-driven design and evaluation of AREAs to ensure the compatibility between children's capabilities and system's features. Similarly, the number of studies attempting to measure the learning effect, qualitatively or quantitatively, was also small. It may imply that the collaboration and communication between the field of HCI and TEL needs to be strengthened, especially the training in the key theoretical and methodological approaches of both fields.

Furthermore, the research community at large is witnessing the serious threat from predatory journals and fake scholarships to research integrity; this can have significant negative impact on SLR in particular. We see the strong need to identify a robust approach to quality assessment. While we have attempted to apply two objective metrics for quality assessment, more research effort needs to be invested in this line of inquiry.

In addition, two issues worthy of more research attention and effort are the limited number of studies involving parents at home and a very few studies targeting children with special needs. Both issues can be related to the affordability of the equipment, including high-quality mobile devices or head-mounted displays.

Overall, we are witnessing the trend of utilizing Augmented Reality as a versatile educational tool. From the usability and UX perspective, there is still much room for improvement to make this promising tool to maximize its potential to benefit children to learn happily in as well as outside schools.

ACKNOWLEDGEMENTS

The publication has been supported by European Union's Horizon 2020 research and innovation program under grant agreement No 856533, project ARETE. We would like to express our thanks to anonymous reviewers of the earlier drafts of this paper; their constructive comments have helped improve its quality.

REFERENCES

Part I: All 49 papers included in the SLR sorted by ID number:

Ex001 Di Serio, Á., Ibáñez, M. B., & Kloos, C. D. (2013). Impact of an augmented reality system on students' motivation for a visual art course. *Computers & Education, 68*, 586-596

Ex002 Sin, A. K., & Zaman, H. B. (2010). Live Solar System (LSS): Evaluation of an Augmented Reality book-based educational tool. *2010 International Symposium on Information Technology (Vol. 1, pp.1-6).*

Ex003 Juan, C. M., Llop, E., Abad, F., & Lluch, J. (2010). Learning words using augmented reality. *10th IEEE International Conference on Advanced Learning Technologies,* pp. 422-426.

Ex004 Juan, C. M., Toffetti, G., Abad, F., & Cano, J. (2010). Tangible cubes used as the user interface in an augmented reality game for edutainment. *10th IEEE International Conference on Advanced Learning Technologies,* pp. 599-603.

Ex005 Liu, W., Cheok, A. D., Mei-Ling, C. L., & Theng, Y. L. (2007). Mixed reality classroom: learning from entertainment. *Proceedings of the 2nd International Conference on Digital Interactive Media in Entertainment and Arts*, pp.65-72.

Ex006 Sayed, N. E., Zayed, H. H., & Sharawy, M. I. (2011). ARSC: Augmented reality student card an augmented reality solution for the education field. *Computers & Education, 56*(4), 1045-1061.

Ex007 Chiang, T. H., Yang, S. J., & Hwang, G. J. (2014). An augmented reality-based mobile learning system to improve students' learning achievements and motivations in natural science inquiry activities. *Journal of Educational Technology & Society, 17*(4), 352-365.

P001 Kaufmann H., Schmalstieg D., Wagner M. (2000). Construct3D: A virtual reality application for mathematics and geometry education. *Education and Information Technologies*, *5*, 263–276

P007 Kaufmann H., Schmalstieg D. (2006). Designing immersive virtual reality for geometry education. *IEEE Virtual Reality Conference* (VR 2006) (pp. 51-58)

P009 Squire K.D., Jan M. (2007). Mad city mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology, 16*(1), 5-29

P011 Freitas R., Campos P. (2008). SMART: A system of augmented reality for teaching 2nd grade students. *People and Computers XXII Culture, Creativity, Interaction 22*, 27-30.

P013 Liu T.-Y., Chu Y.-L. (2008). Handheld augmented reality supported immersive ubiquitous learning system. *2008 IEEE International Conference on Systems, Man and Cybernetics* (pp. 2454-2458).

P018 Dunleavy M., Dede C., Mitchell R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of Science Education and Technology*, *18*(1), 7-22.

P019 Arvanitis T.N., Petrou A., Knight J.F., Savas S., Sotiriou S., Gargalakos M., Gialouri E. (2009). Human factors and qualitative pedagogical evaluation of a mobile augmented reality system for science education used by learners with physical disabilities. *Personal and Ubiquitous Computing, 13*(3), 243-250.

P030 Luckin R., & Fraser D.S. (20111). Limitless or pointless? An evaluation of augmented reality technology in the school and home. International Journal of Technology Enhanced Learning, 3(5), 510-524.

P033 Juan M.C., Furió D., Seguí I., Rando N., Cano J. (2011). Lessons learnt from an experience with an augmented reality iPhone learning game. *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology* (pp. 1-8).

P048 Wojciechowski R., Cellary W. (2013). Evaluation of learners' attitude toward learning in ARIES augmented reality environments. *Computers & Education, 68*, 570-585.

P055 Salvador-Herranz G., Pérez-López D., Ortega M., Soto E., Alcañiz M., Contero M.(2013). Manipulating virtual objects with your hands: A case study on applying desktop Augmented Reality at the primary school. *46th Hawaii International Conference on System Sciences* (pp. 31-39).

P072 Cai S., Wang X., Chiang F.-K. (2014). A case study of Augmented Reality simulation system application in a chemistry course. *Computers in Human Behavior, 37*, 31-40.

P074 Ibáñez M.B., Di Serio Á., Villarán D., Delgado Kloos C. (2014). Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education, 71*, 1-13.

P098 Lin H.-C.K., Chen M.-C., Chang C.-K. (2015). Assessing the effectiveness of learning solid geometry by using an augmented reality-assisted learning system. *Interactive Learning Environments, 23*(6), 799-810

P107 Lu S.-J., Liu Y.-C.(2015). Integrating augmented reality technology to enhance children's learning in marine education. *Environmental Education Research, 21*(4), 525-541.

P108 Fleck S., Hachet M., Christian Bastien J.M. (2015). Marker-based Augmented Reality: Instructional-design to improve children interactions with astronomical concepts. *Proceedings of the 14th International Conference on Interaction Design and Children* (pp. 21-28).

P109 Radu I., Doherty E., DiQuollo K., McCarthy B., Tiu M. (2015). Cyberchase Shape Quest: Pushing geometry education boundaries with augmented reality. *Proceedings of the 14th International Conference on Interaction Design and Children* (pp. 430-433).

P128 Hsiao H.-S., Chang C.-S., Lin C.-Y., Wang Y.-Z. (2016). Weather observers: a manipulative augmented reality system for weather simulations at home, in the classroom, and at a museum. *Interactive Learning Environments, 24*(1), 205-223.

P131 Laine T.H., Nygren E., Dirin A., Suk H.-J. (2016). Science Spots AR: a platform for science learning games with augmented reality. *Educational Technology Research and Development, 64*(3), 507-531.

P134 Laine T.H., Suk H.J. (2016). Designing mobile augmented reality exergames. *Games and Culture, 11(*5), 548-580.

P167 Ahmed S., Nasir B., Khan J.A., Ali S., Umer M. (2017). MAPILS: Mobile augmented reality plant inquiry learning system. *IEEE Global Engineering Education Conference (EDUCON)* (pp. 1443-1449).

P174 Joo-Nagata J., Martinez Abad F., García-Bermejo Giner J., García-Peñalvo F.J. (2017). Augmented reality and pedestrian navigation through its implementation in m-learning and e-learning: Evaluation of an educational program in Chile. *Computers & Education, 111*, 1-17

P175 Cai S., Chiang F.-K., Sun Y., Lin C., Lee J.J. (2017). Applications of augmented reality-based natural interactive learning in magnetic field instruction. *Interactive Learning Environments, 25*(6), 778-791.

P179 Alakärppä I., Jaakkola E., Väyrynen J., Häkkilä J.(2017). Using nature elements in mobile AR for education with children. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1-13)

P182 Wang Y.-H.(2017). Exploring the effectiveness of integrating augmented reality-based materials to support writing activities. *Computers & Education, 113*, 162-176

P201 Georgiou Y., Kyza E.A.(2018). Investigating the coupling of narrative and locality in augmented reality educational activities: Effects on students' immersion and learning gains. *13th International Conference of the Learning Sciences* (ICLS) 2018, Volume 1. London, UK: International Society of the Learning Sciences.

P205 Layona R., Yulianto B., Tunardi Y.(2018). Web based Augmented Reality for Human Body Anatomy Learning. *Procedia Computer Science, 135*, 457-464.

P214 Leonard S.N., Fitzgerald R.N. (2018). Holographic learning: A mixed reality trial of Microsoft Hololens in an Australian secondary school. *Research in Learning Technology, 26*.

P220 Efstathiou I., Kyza E.A., Georgiou Y.(2018). An inquiry-based augmented reality mobile learning approach to fostering primary school students' historical reasoning in non-formal settings. *Interactive Learning Environments, 26*(1), 22-41.

P222 Wu P.-H., Hwang G.-J., Yang M.-L., Chen C.-H.(2018). Impacts of integrating the repertory grid into an augmented reality-based learning design on students' learning achievements, cognitive load and degree of satisfaction. *Interactive Learning Environments, 26*(2), 221-234.

P223 Alhumaidan H., Lo K.P.Y., Selby A.(2018). Co-designing with children a collaborative augmented reality book based on a primary school textbook. *International Journal of Child-Computer Interaction, 15*, 24-36.

P231 Schaper M.-M., Santos M., Malinverni L., Zerbini Berro J., Pares N.(2018). Learning about the past through situatedness, embodied exploration and digital augmentation of cultural heritage sites. *International Journal of Human-Computer Studies, 114*, 36-50.

P232 Malinverni L., Valero C., Schaper M.-M., Pares N.(2018). A conceptual framework to compare two paradigms of augmented and mixed reality experiences. *Proceedings of the 17th ACM Conference on Interaction Design and Children* (pp. 7-18).

P256 Takahashi I., Oki M., Bourreau B., Kitahara I., Suzuki K.(2018). An empathic design approach to an augmented gymnasium in a special needs school setting. *International Journal of Design, 12*(3).

P303 Hsu T.-C.(2019). Effects of gender and different augmented reality learning systems on English vocabulary learning of elementary school students. *Universal Access in the Information Society, 18*(2), 315-325.

P314 Hsu H.-P., Wenting Z., Hughes J.E.(2019). Developing Elementary Students' Digital Literacy Through Augmented Reality Creation: Insights From a Longitudinal Analysis of Questionnaires, Interviews, and Projects. *Journal of Educational Computing Research, 57*(6), 1400-1435.

P320 Ewais A., Troyer O.D.(2019). A Usability and Acceptance Evaluation of the Use of Augmented Reality for Learning Atoms and Molecules Reaction by Primary School Female Students in Palestine. *Journal of Educational Computing Research, 57*(7), 1643-1670.

P330 López-Faican, L., Jaen, J.(2020). EmoFindAR: Evaluation of a mobile multiplayer augmented reality game for primary school children. *Computers & Education, 149*, 103814

P335 del Río Guerra, M.S., Martínez, A.E.G., Martin-Gutierrez, J., López-Chao, V.(2020). The limited effect of graphic elements in video and augmented reality on children's listening comprehension. *Applied Sciences*, *10*(2), 527

P376 Rossano, V., Lanzilotti, R., Cazzolla, A., & Roselli, T. (2020). Augmented Reality to Support Geometry Learning. *IEEE Access, 8,* 107772-107780.

W104 Gopalan, V., Zulkifli, A. N., & Bakar, J. A. A. (2016, August). A study of students' motivation using the augmented reality science textbook. In *AIP Conference Proceedings* (Vol. 1761, No. 1, p. 020040).

W106 Huang, T. C., Chen, C. C., & Chou, Y. W. (2016). Animating eco-education: To see, feel, and discover in an augmented reality-based experiential learning environment. *Computers & Education, 96*, 72-82.

Part 2: In-text bibliography

---

[1] Akçayır, M., & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review, 20*, 1-11. [h-index: 57; GDI: 600]

[2] Anderson, F., & Bischof, W. F. (2014). Augmented reality improves myoelectric prosthesis training. International *Journal on Disability and Human Development, 13*(3), 349-354.

[3] Arksey H, and O'Malley L. (2005). Scoping studies: towards a methodological framework. *International Journal Social Research Methodology,* 8, 19-31.

[4] Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators & Virtual Environments, 6*(4), 355-385.

[5] Bacca, J., Baldiris, S., Fabregat, R., Graf, S. & Kinshuk (2014). Augmented reality trends in education: a systematic review of research and applications. *Educational Technology & Society, 17*(4), 133–149.

[6] Bai, Z., & Blackwell, A. F. (2012). Analytic review of usability evaluation in ISMAR. *Interacting with Computers*, 24(6), 450-460.

[7] Bekker, T., & Antle, A. N. (2011). Developmentally situated design (DSD) making theoretical knowledge accessible to designers of children's technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2531-2540).

[8] Bennett, J. L. (1979). The commercial impact of usability in interactive systems. *Man-computer Communication, Infotech State-of-the-Art*, 2, 1-17.

[9] Berryman, D. R. (2012). Augmented reality: a review. *Medical Reference Services Quarterly, 31*(2), 212-218.

[10] Billinghurst M, Clark A, Lee G (2015). A Survey of Augmented Reality. *Foundations and Trends in Human-Computer Interaction. 8*(2-3). 73-272.

[11] Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49-59.

[12] Brooke, J. (1986). *System usability scale (SUS): a quick-and-dirty method of system evaluation user information*. Reading, UK: Digital Equipment Co Ltd, 43.

[13] Chang R.-C., Chung L.-Y., Huang Y.-M. (2016). Developing an interactive augmented reality system as a complement to plant education and comparing its effectiveness with video learning. *Interactive Learning Environments, 24*(6), 1245-1264.

[14] Cheng, K. H., & Tsai, C. C. (2013). Affordances of augmented reality in science learning: Suggestions for future research. *Journal of Science Education and Technology, 22*(4), 449-462.

[15] Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., ... & Moher, D. (2014). Scoping reviews: time for clarity in definition, methods, and reporting. *Journal of Clinical Epidemiology, 67*(12), 1291-1294.

[16] Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper and Row.

[17] Da Gama, A., Chaves, T., Figueiredo, L., & Teichrieb, V. (2012, May). Guidance and movement correction based on therapeutics movements for motor rehabilitation support systems. In *2012 14th Symposium on Virtual and Augmented Reality* (pp. 191-200). IEEE.

[18] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319-340.

[19] Dey, A., Billinghurst, M., Lindeman, R. W., & Swan, J. (2018). A systematic review of 10 years of augmented reality usability studies: 2005 to 2014. *Frontiers in Robotics and AI,* 5, 37.

[20] Diegmann, P., Schmidt-Kraepelin, M., Eynden, S., & Basten, D. (2015). Benefits of augmented reality in educational environments-a systematic literature review. *Benefits, 3*(6), 1542-1556.

[21] Duc, N. M., Hiep, D. V., Thong, P. M., Zunic, L., Zildzic, M., Donev, D., ... & Masic, I. (2020). Predatory open access journals are indexed in reputable databases: a revisiting issue or an unsolved problem. *Medical Archives, 74*(4), 318.

[22] Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing.* Intellect books.

[23] Durrani, U., & Pita, Z. (2018, December). Integration of Virtual Reality and Augmented Reality: Are They Worth the Effort in Education?. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 322-327). IEEE.

[24] Endsley, T. C., Sprehn, K. A., Brill, R. M., Ryan, K. J., Vincent, E. C., & Martin, J. M. (2017, September). Augmented reality design heuristics: Designing for dynamic interactions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 2100-2104). Sage CA: Los Angeles, CA: SAGE Publications.

[25] Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. the MIT Press.

[26] Fan, M., Antle, A. N., & Warren, J. L. (2020). Augmented reality for early language learning: A systematic review of augmented reality application design, instructional strategies, and evaluation outcomes. *Journal of Educational Computing Research, 58*(6), 1059-1100.

[27] Fokides E., Mastrokoukou A. (2018). Results from a study for teaching human body systems to primary school students using tablets. *Contemporary Educational Technology, 9*(2), 154-170.

[28] Gabbard, J., Swan II, J. E., Hix, D., Lanzagorta, M. O., Livingston, M., Brown, D. B., & Julier, S. J. (2002). Usability engineering: domain analysis activities for augmented-reality systems. In *Stereoscopic displays and virtual reality systems IX* (Vol. 4660, pp. 445-457). International Society for Optics and Photonics.

[29] Garside R. (2014). Should we appraise the quality of qualitative research reports for systematic reviews, and if so, how? *Innovation 27*:67–79

[30] Garzon, J., & Acevedo, J. (2019). Meta-analysis of the impact of Augmented Reality on students' learning gains. *Educational Research Review, 27*, 244-260.

[31] Gómez-López, P., Simarro, F. M., & Bonal, M. T. L. (2019, June). Analysing the UX scope through its definitions. In *Proceedings of the XX International Conference on Human Computer Interaction* (pp. 1-4).

[32] Gonzalez SL, Alvarez V and Nelson EL (2019) Do Gross and Fine Motor Skills Differentially Contribute to Language Outcomes? A Systematic Review. *Frontiers in Psychology. 10*:2670.

[33] Gould, J. D., & Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM, 28*(3), 300-311.

[34] Grudin, J. (2017). From tool to partner: The evolution of human-computer interaction. *Synthesis Lectures on Human-centered Interaction, 10*(1), i-183.

[35] Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the literature. *Journal of informetrics, 11*(3), 823-834.

[36] Hassenzahl, M. (2005). The thing and I: understanding the relationship between user and product. In *Funology* (pp. 31-42). Kluwer Academic Publishers.

[37] Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction, 25*(3), 235-260.

[38] Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & Information Technology, 25*(2), 91-97.

[39] Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In Mensch & computer 2003 (pp. 187-196). Vieweg+ Teubner Verlag.

[40] Hodhod, R., Fleenor, H., Nabi, S. (2014). Adaptive augmented reality serious game to foster problem solving skills. In: *Workshop Proceedings of the 10th International Conference on Intelligent Environments* (pp.273-284). IOS Press.

[41] Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior, 26*(6), 1278-1288.

[42] Hornbæk, K., & Hertzum, M. (2017). Technology acceptance and user experience: A review of the experiential component in HCI. *ACM Transactions on Computer-Human Interaction* (TOCHI), 24(5), 1-30.

[43] Hornbæk, K., & Law, E. L. C. (2007, April). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 617-626).

[44] Ibáñez, M. B., & Delgado-Kloos, C. (2018). Augmented reality for STEM learning: A systematic review. *Computers & Education, 123*, 109-123.

[45] Ishii, H. (2008). The tangible user interface and its evolution. *Communications of the ACM, 51*(6), 32-36.

[46] Juni P, Witschi, A., Bloch R, Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association, *282*:1054–60

[47] Ko, S. M., Chang, W. S., & Ji, Y. G. (2013). Usability principles for augmented reality applications in a smartphone environment. *International Journal of Human-computer Interaction, 29*(8), 501-515.

[48] Krauß, M., Riege, K., Winter, M., & Pemberton, L. (2009, September). Remote hands-on experience: Distributed collaboration with augmented reality. In Proceedings of European Conference on Technology Enhanced Learning (pp. 226-239). Springer.

[49] Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., & Sinnelä, A. (2011). UX Curve: A method for evaluating long-term user experience. *Interacting with Computers, 23*(5), 473-483.

[50] Lai, J. W., & Bower, M. (2019). How is the use of technology in education evaluated? A systematic review. Computers & Education, 133, 27-42.

[51] Law, E. L. C., Brühlmann, F., & Mekler, E. D. (2018, October). Systematic review and validation of the game experience questionnaire (geq)-implications for citation and reporting practice. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (pp. 257-270).

[52] Law, E. L. C., Hassenzahl, M., Karapanos, E., Obrist, M., & Roto, V. (2015). Tracing links between UX frameworks and design practices: dual carriageway. In *Proceedings of HCI Korea*, 188-195.

[53] Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009, April). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 719-728).

[54] Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction, 34(*12), 1148-1156.

[55] Lorusso M.L., Giorgetti M., Travellini S., Greci L., Zangiacomi A., Mondellini M., Sacco M., Reni G. (2016). Giok: An alien stimulates pragmatic and social skills in pre-school children. *Proceedings of the 4th Workshop on ICTs for Improving Patients Rehabilitation Research Techniques* (pp. 89-92).

[56] Liu Q., Xu S., Yu S., Yang Y., Wu L., Ba S. (2019). Design and implementation of an ar-based inquiry courseware - Magnetic field. *International Symposium on Educational Technology* (ISET) (pp. 134-138). IEEE.

[57] Lu S.-J., Liu Y.-C., Chen P.-J., Hsieh M.-R. (2020). Evaluation of AR embedded physical puzzle game on students' learning achievement and motivation on elementary natural science. *Interactive Learning Environments, 28*(4), 451-463.

[58] Mañana-Rodríguez, J. (2015). A critical review of SCImago journal & country rank. *Research evaluation, 24*(4), 343-354.

[59] Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology, 25*(2), 141-158.

[60] Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics, 12*(4), 1160-1177.

[61] Mattelmäki, T., Vaajakallio, K., & Koskinen, I.. (2014). What happened to empathic design? *Design issues 30.1*: 67-77.

[62] McCarthy, J., & Wright, P. (2004). *Technology as experience*. MIT Press.

[63] Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems, 77*(12), 1321-1329.

[64] Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995, December). Augmented reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and Telepresence Technologies* (Vol. 2351, pp. 282-292). International Society for Optics and Photonics.

[65] Minichiello, A., Hood, J. R., & Harkness, D. S. (2018). Bringing user experience design to bear on STEM education: A narrative literature review. Journal for STEM Education Research, 1(1), 7-33.

[66] Mirnig, A. G., Meschtscherjakov, A., Wurhofer, D., Meneweger, T., & Tscheligi, M. (2015, April). A formal analysis of the ISO 9241-210 definition of user experience. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 437-450).

[67] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med, 6*(7), e1000097.

[68] Moreno-Guerrero, A. J., Alonso García, S., Ramos Navas-Parejo, M., Campos-Soto, M. N., & Gómez García, G. (2020). Augmented Reality as a Resource for Improving Learning in the Physical Education Classroom. *International Journal of Environmental Research and Public Health, 17*(10), 3637.

[69] Munsinger, B., & Quarles, J. (2019, November). Augmented Reality for Children in a Confirmation Task: Time, Fatigue, and Usability. In *25th ACM Symposium on Virtual Reality Software and Technology* (pp. 1-5).

[70] Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.

[71] Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 249-256).

[72] Pajić, D. (2015). On the stability of citation-based journal rankings. *Journal of Informetrics, 9*(4), 990-1006.

[73] Pellas, N., Fotaris, P., Kazanidis, I., & Wells, D. (2019). Augmenting the learning experience in primary and secondary school education: A systematic review of recent trends in augmented reality game-based learning. *Virtual Reality, 23*(4), 329-346.

[74] Pemberton L., Winter M. (2009). Collaborative Augmented Reality in schools. *Computer Supported Collaborative Learning Practices: CSCL2009 Community Events Proceedings* (pp. 109-111). Rhodes, Greece: International Society of the Learning Sciences.

[75] Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Research synthesis methods, 5*(4), 371-385.

[76] Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*. American Psychological Association.

[77] Radu, I. (2014). Augmented reality in education: a meta-review and cross-media analysis. *Personal and Ubiquitous Computing, 18*(6), 1533-1543.

[78] Radu, I., & MacIntyre, B. (2012). Using children's developmental psychology to guide augmented-reality design and usability. In: *IEEE international symposium on mixed and augmented reality* (ISMAR) (pp. 227-236). IEEE.

[79] Santos, M. E. C., Chen, A., Taketomi, T., Yamamoto, G., Miyazaki, J., & Kato, H. (2013). Augmented reality learning experiences: Survey of prototype design and evaluation. *IEEE Transactions on Learning Technologies,*7(1), 38-56.

[80] Santos, M. E. C., Polvi, J., Taketomi, T., Yamamoto, G., Sandor, C., & Kato, H. (2015). Toward standard usability questionnaires for handheld augmented reality. *IEEE computer graphics and applications, 35*(5), 66-75.

[81] Sauro, F. (2013). *A brief history of usability*. https://measuringu.com/usability-history/

[82] Schneck, C. M. (2010). Visual perception. *Occupational Therapy for Children* (6th Ed.) (pp.363-403). Mosby Inc.

[83] Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology, 70*, 747-770.

[84] Sim, G., MacFarlane, S., & Read, J. (2006). All work and no play: Measuring fun, usability, and learning in software for children. *Computers & Education, 46*(3), 235-248.

[85] Sirakaya, M., & Alsancak Sirakaya, D. (2018). Trends in Educational Augmented Reality Studies: A Systematic Review. *Malaysian Online Journal of Educational Technology, 6*(2), 60-74.

[86] Sorokowski, P., Kulczycki, E., Sorokowska, A., & Pisanski, K. (2017). Predatory journals recruit fake editor. *Nature News, 543*(7646), 481.

[87] Speicher, M., Hall, B. D., & Nebeling, M. (2019, May). What is mixed reality? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

[88] Swan, J. E., & Gabbard, J. L. (2005, July). Survey of user-based experimentation in augmented reality. In *Proceedings of 1st International Conference on Virtual Reality* (Vol. 22, pp. 1-9).

[89] Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312.

[90] Tekedere, H., & Göke, H. (2016). Examining the effectiveness of augmented reality applications in education: A meta-analysis. *International Journal of Environmental and Science Education*, 11(16), 9469-9481.

[91] Valentine, J.C., & Cooper, H. (2005). Can we measure the quality of causal research in education? In Phye, G.D. Robinson, D.H., & Levin, J. (Eds.), *Experimental Methods for Educational Interventions: Prospects, Pitfalls and Perspectives* (pp. 85–112). San Diego, CA: Elsevier

[92] Van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR), 50*(6), 1-40.

[93] Vlachogianni, P., & Tselios, N. (2021). Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. Journal of Research on Technology in Education, 1-18.

[94] Walsh, G., Druin, A., Guha, M.L., Foss, E., Golub, E., Hatley, L., Bonsignore, E., and Franckel, S. (2010) Layered elaboration: a new technique for co-design with children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1237-1240).

[95] Whiting, P., Wolff, R., Mallett, S., Simera, I., & Savović, J. (2017). A proposed framework for developing quality assessment tools. *Systematic Reviews, 6*(1), 204.

[96] Winarni, E.W., Purwandari, E.P. (2019). The effectiveness of turtle mobile learning application for scientific literacy in elementary school. *Journal of Education and e-Learning Research, 6(*4), 156-161.

[97] Wojciechowski R., Cellary W. (2013). Evaluation of learners' attitude toward learning in ARIES augmented reality environments. *Computers & Education, 68*, 570-585.

[98] Wyeth, P., & Purchase, H. C. (2003, July). Using developmental theories to inform the design of technology for children. In *Proceedings of the 2003 Conference on Interaction Design and Children* (IDC) (pp. 93-100).