Thirteen years since the first iPhone:
A systematic review on the effectiveness of language learning apps on smart devices

Ziwei “Jo” Huang 黄梓蔚

Department of Education, University of Oxford

Kellogg College

Supervisor: Dr. Hamish Chalmers

Dissertation submitted in part-fulfilment of the requirements for the degree of

Master of Science in Applied Linguistics and Second Language Acquisition

Trinity Term, 2020

Word count: 19,916

PLAIN LANGUAGE SUMMARY

Background

Since the release of the first iPhone in 2007, language learning apps such as Duolingo have been popular with language educators and learners. The online language learning market is valuated to worth as much as $21 billion by 2027. Some apps claim to provide the "best" and "fastest" ways of learning a foreign language.

Objectives

This review explores the extent and nature of research studies that investigated whether language learning apps are effective as claimed. Besides critically analysing and synthesising their results, the review also assesses the studies' trustworthiness.

Methods

The review uses a best-evidence, systematic review approach. I included original research studies that reported quantitative data related to language proficiency and were published in all languages. With a theory-informed search string, I identified 7,510 records from 10 electronic databases. Using pre-defined inclusion and exclusion criteria, I screened the records twice with the help of a second reviewer. I also evaluated the trustworthiness of each included study's results based on their study design and reporting practices.

Results

The search found 32 eligible studies, covering a dozen countries and nine languages, the most popular being the USA and English. Researchers were interested in two types of comparison: one compared the effectiveness of different apps or versions of the same app; the other compared language learning apps with other methods of language learning. Altogether, these studies surveyed more than 10 commercially available apps and a larger number of researcher-developed ones. Vocabulary was the most popular area of interest among the nine measured language skills and most studies used adults as participants.
As a whole, the results show that learners improved on their language proficiency tests after using the apps. Functions like gamification and more access to multimedia resources are good for language learning. Some studies found learning with an app as effective as with a face-to-face teacher. Only in two studies, participants received either lower or the same scores after learning language with apps. However, the collective quality of these

studies is low; only four studies were considered “strong” on their study design and reporting practices.

Implications

For language learners, apps can provide affordable and accessible educational resources. However, users should beware that all apps are not the same and personal usage significantly impacts learning outcomes. Different apps have different functions and users should choose the app according to their preferences. There is not enough evidence to show that learners of all ages and proficiency levels benefit from apps equally.

In the future, researchers should strive to produce more rigorously designed and reported studies. More attention should be directed to study non-adult learners, more advanced learners, and languages other than English.

ABSTRACT

Powered by advanced calculating power and novel technologies like artificial intelligence, smartphone apps offer unprecedent creative solutions in education. They have the potential to change the reality of language teaching and learning. This systematic review critically evaluates the empirical basis for the effectiveness of language learning apps.

A rigorous search on electronic databases and two screenings identified 32 eligible studies, mostly with adults and learners with lower levels of proficiency. The research comprises of two types of comparisons: one compared the effectiveness of different apps or versions of the same app; the other compared language learning apps with other methods of language learning. The research covered a dozen countries, nine languages, nine language skills, 10 commercially available apps and a larger number of researcher-developed ones. 25 of 32 studies found unequivocally positive effects of learning foreign language on apps regardless of functions. Such features as gamification and adaptive learning are conducive to improving language proficiency. Only two studies found participants to score either lower or equally after treatment. A risk of bias assessment reveals that the overall quality of research is low due to inadequate study design and reporting practices.

For language learners, apps can provide affordable and accessible educational resources. However, there is insufficient evidence to show that learners of all ages and proficiency levels benefit equally from apps. Future research needs more robust study designs and should direct attention to study non-adult learners, more advanced learners, and languages other than English.

Keywords: systematic review, language learning apps, m-learning, mobile language learning

Acknowledgements

Thanks to Guido van Poppel, who offers me a haven and deals with all my tantrums with the most unselfish support during this challenging time. Thanks to my parents who have always provided for me. I couldn't have done it without you. I love you.

Thank you, Hamish Chalmers, who inspired me with great ideas and gave the most needed advice and suggestions.

Thank you, Atsushi Kanayama for your invaluable support with this project. And thank you, Jiawei Shi, for the moral support in this journey.

Thank you to my teachers over the years, especially Dr. Greg Robbins (and Ms. Peggy Milne), Dr. Feng Xiao, and Dr. Carmen Fought. Your works motivated me to be a linguist and an educator.

At last, thank you and congratulations to my supportive, fellow ALSLA students. It has not been an easy year, but we did it!

Table of contents

List of figures

List of tables

Quick reference to abbreviations

App: application

ASR: automatic/automated speech recognition

CALL: computer-assisted language learning

D-learning: digital learning

DCT: dual-coding theory

E-learning: electronic learning

EPHPP tool: Effective Public Health Practice Project Quality Assessment Tool for Quantitative Studies

ESL: English as a Second Language

FL: foreign language

GPS: global positoning system

L1: first language

L2: second language

M-learning: mobile learning

MALL: mobile-assisted language learning

PDA: personal digital assistant

RCT: randomised-controlled trial

RoB: risk of bias

SLA: second language acquisition

CHAPTER 1. Introduction

1.1. Background to the study

In 2019, a company called Duolingo was listed by Forbes as the next billion-dollar start-ups among 24 other companies; it was the only company in the education sector on this list (Feldman & Carson, 2019). Endorsed by 300 million registered users and 28 million monthly active users[1], Duolingo teaches them languages through "smart," mobile devices (Lardinois, 2018). Though considered the most popular language learning app, Duolingo is not an isolated case of success in the market of online language learning. Multiple market reports valuated the online language learning market to grow at a compound annual growth rate (CAGR) as high as 18%, reaching $21 billion by 2027 (Meticulous Market Research Pvt. Ltd, 2020; Technavio, 2020).

Language learning apps are the successor of a long-standing area of study in second language acquisition (SLA): mobile-assisted language learning (MALL). Kukulska-Hume (2009) and Godwin-Jones (2011, 2017) review the affordances of the mobile technologies and highlight the potential of mobile learning in altering the current reality of language education. Powered with advanced calculating power and novel technologies like artificial intelligence, smartphone apps offer an unprecedented range of creative solutions. Whereas in MALL, mobile devices only played an assistive part, language learning apps now take the central teaching role. For less than £10 a month, learners can have full-fledged language courses at their disposal, "anytime, anywhere" (Kukulska-Hulme & Shield, 2008, p. 273).

1.2. Rationale and objectives of the study

Is the commercial success of Duolingo and its peers warranted in academic research? How effective are these apps in teaching people language? In their marketing materials, Duolingo claim to be the "best way to learn a language" and Memrise the "fastest." These claims are somewhat supported by empirical evidence: Vesselinov and Grego's (2012) study found that 34 hours of learning Spanish on Duolingo would equal to a semester of university-level course. Other studies have also found similar results (see 2.2. for a review on the literature). Despite these claims, there has yet been a systematic and comprehensive examination on the empirical evidence on this topic. Then, how much

[1] These numbers are likely much higher at the time of writing this report. However, no reliable source could be located for the latest statistics in August 2020.

research has been conducted to investigate the effectiveness of language learning apps, and can we trust their results?

The current study seeks to answer these questions with a systematic review approach. I purposefully select empirical studies from the existing literature using rigorously defined criteria. Then, I evaluate the studies' results and their trustworthiness. Results of this review will inform language learners, educators, and researchers with a state-of-the-art synthesis on the current evidence on the effectiveness of language learning apps.

1.3. Dissertation outline

In this report, I first review the existing literature on mobile language learning, which includes relevant concepts in applied linguistics and educational technology. I also identify the research questions following the literature review. Then, I outline my research methodology, followed by results and analysis. Finally, I discuss the results and provide conclusions and recommendations for future research.

CHAPTER 2. Literature review

2.1. M-learning, e-learning, d-learning, and MALL

Smartphones have become an indispensable technology for the 21st century humans. Features of high calculating capacity, screen size, and stable access to the Internet afford it capabilities to transform almost all sectors, including education. M-learning (mobile learning), e-learning (electronic learning) and d-learning (digital learning) are three commonly cited concepts in the literature on education and technology. In a systematic review, Basak et al. (2018) surveyed 126 articles to synthesise the definitions of these concepts. They concluded that while closely related, these three types of learning construct a superset-subset relationship. D-learning includes a wide range of learning experiences, facilitated by any use of digital technology. E-learning specifies learning activities accomplished through or supported by electronic devices and media, such as laptop/desktop computers, television and phones. M-learning specifically refers to e-learning activities with mobile electronic devices that function without "physical connections to cable networks" (Georgiev et al., 2004, p. 2) such as mobile phones, tablet computers, and PDAs. M-learning and e-learning are, thus, two types of solutions for d-learning, and m-learning is a subset of e-learning (Basak et al., 2018). This categorisation echoes that in Georgiev et al. (2004), who depict the relationship in this diagram:

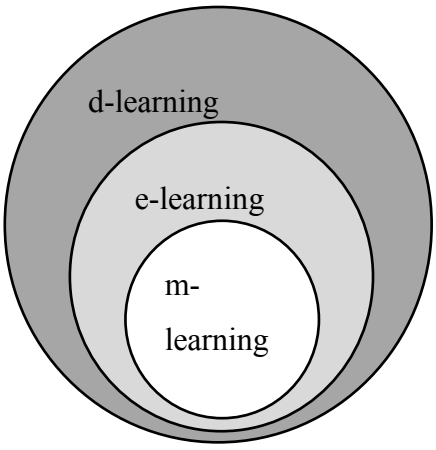

Figure 1 The place of m-learning as part of e-learning and d-learning.
Reproduced from Georgiev et al. (2004, p.1)

In m-learning, a widely researched area is the topic of mobile-assisted language learning (MALL). As its name suggests, MALL sits at the intersection of language learning and m-learning, denoting learners' use of mobile devices to enhance their language learning experience. The era of MALL could be divided into two phases: pre-2007 and post-2007. The iPhone, arguably the first smartphone and its accompanying operating system (iOS)

were released by Apple Inc. in 2007. It was then followed by Android by Google, the currently dominant smartphone operating system in the global market (O'Dea, 2020). Before smartphones, tech-savvy language educators utilised technologies like MP3, PDAs, and portable digital dictionaries for delivery of some contents (see 2.2. and Burston, 2014a, 2015; Kukulska-Hulme & Shield, 2008 for further discussion). The launch of smart devices in 2007, however, takes language learning into brand new horizons.

2.1.1. The concept of "smartness"

The concept of smartness in electronic devices is a recent development in the field of information technology. The search for artificial and machine intelligence led to the use of the adjective "smart" for non-human entities (Alter, 2020). Lamenting the vagueness of its definition, Alter (2020) conceptualised "smart" as follows:

> Purposefully designed entity X is smart to the extent to which it performs and controls functions that attempt to produce useful results through activities that apply automated capabilities and other physical, informational, technical, and intellectual resources for processing information, interpreting information, and/or learning from information that may or may not be specified by its designers (p. 384).

In the definition, X is a device, socio-technical system or automated system. This conceptualisation comprises of four dimensions: information processing, internal regulation, action in the world, and knowledge acquisition (see Alter, 2020 for a detailed explanation). On each dimension, X's capabilities sit on a gradient from "not smart at all" to extremely smart. An extremely smart device is capable of accomplishing tasks even when they are unprespecified by its designers.

Some functions and capabilities of an iPhone 11 running on the iOS operating system are examples of these smart dimensions. For instance, the iPhone actively listens for the audial input "Hey Siri (a virtual assistant)" in all contexts, which activates the phone's automatic speech recognition (ASR) function. When activated, Siri requests more input from the user to produce information. When the user asks questions like "what's the weather," Siri provides an answer using ASR (to recognise the question), Global Positioning System (GPS) (to locate the user), and Internet search (to look for the weather conditions at the given location). iPhone also stores information about the user's behaviours on the phone and generate shortcuts based on their habits, which is an example of knowledge acquisition.

In sum, the advanced calculating capabilities and functions of smart devices distinguish them from their predecessors, namely non-smart cell phones and PDAs. In reality, smartphones are the combination of a computer and a mobile phone, allowing users to accomplish a range of digital learning activities on the go (Godwin-Jones, 2017). In the current paper, smart devices refer to mobile phones and tablet computers that operate on a "smart" operating system, that is, Android, Windows, iOS (iPad OS) or Amazon.

2.1.2. The powers of the apps

An outstanding feature that marks the difference between smart devices and their predecessors is the availability of mobile applications, which are often shortened as *apps*. Apps are software packages that allow its users to access different functions on the device and provide users with various services and capabilities. As of 2020's first quarter, there were over 2.5 millions apps on the Android app store (Google Play), 1.8 million on Apple App Store, 0.67 million on Windows Store, and 0.49 million on Amazon App Store (Clement, 2020). Compared to traditional access to Internet services using Web browsers, apps afford more capabilities thanks to smart devices' high calculating power and built-in functions (for details on the technology see Godwin-Jones, 2011). App developers exploit these functions in creative ways to facilitate learning.

GPS functions, for example, allow for contextualised or situated learning. Situated learning refers to learners' interaction with real and physical surroundings to facilitate their acquisition of knowledge (Huang et al., 2016). This interaction serves to help learners integrate their knowledge in real-life situations. Huang et al. (2016) designed a mobile vocabulary learning tool that features GPS-supported situated learning on a mobile phone. The tool obtains the physical location of the user through GPS and provides them with matching learning materials. In Huang et al.'s (2016) teaching and learning experiment, students who learned with the mobile, situated learning tool were compared to those who studied the same materials in a traditional, face-to-face classroom. Students using the situated learning tool showed significantly more improvement on their vocabulary quiz score (Mean=11.25, SD=10.61, t=6.71, p<.000) than those who studied with a teacher, face-to-face (Mean=0.88, SD=11.15, t=0.5, p=0.62). The researchers claim that this difference results from the fact that the mobile language learning tool can deliver materials at the learner's request. This helps learners focus on understanding the meaning of new words in real-time without consulting other sources.

Automatic speech recognition is another commonly used function in language learning apps. ASR identifies the spoken words of a speaker and translate them into text readable by the computer and humans. Liakin et al. (2015) conducted an experiment to investigate whether the use of ASR on a smartphone can improve French learners' production and perception of the phoneme /y/. The results show that the use of ASR has a positive effect on production only (Liakin et al., 2015). See 5.1. for more discussion on the study.

Other smart functions that are commonly used by m-learning apps include the use of multimedia materials, gamification, and adaptive learning technology (Heil et al., 2016).

Multimedia learning. Even before smart phones, multimedia learning was promoted by researchers of language learning at large and in computer-assisted language learning (M.-P. Chen et al., 2014). It refers to the use of different types of stimuli, visual and audial, in the presentation of information. The manifestations of these stimuli include written text, pictures, animations, videos, sounds, and so on. The benefits of employing multimedia in learning are supported by an integrated cognitive theory of multimedia learning (Mayer, 2002), which comprises of elements from the dual-coding theory (J. M. Clark & Paivio, 1991), the cognitive load theory (Sweller, 1994), and the working memory model (Baddeley, 1992; Baddeley & Logie, 1999).

Mayer defines multimedia learning as learning activities that takes place by "building a mental representation from words and pictures" (Mayer, 2002, p. 85), which extends to include not only written text and static pictures, but also animated images (videos) and spoken speech.. Contending that multimedia learning is superior than learning with only one type of instructional message, Mayer (2002) and colleagues surveyed 11 studies to support this theory. They found that in all experiments, learners who had access to multimedia instructional messages had better learning results in non-verbal problem-solving skills than those who did not. While Mayer and Anderson (1991) replicated this positive effect of simultaneous multimedia messages in two other experiments, it was unclear whether verbal tasks could produce the same results.

For the purpose of the current project, evidence for the effectiveness of multimedia learning in SLA is needed. Al-Seghayer (2001) investigated the effects of multimedia modes on vocabulary acquisition using a within-subjects design where 30 ESL students studied vocabulary with annotations in video and text, in picture and text, and in text only. The results from recognition and production post-tests showed that

participants performed the best on vocabulary learned with videos and text (M=6.1, SD=0.759), compared to those with pictures and text (M=4.7, SD=0.952) and text only (M=4.03, SD=1.586), χ^2 (df= 2, N= 30) = 28.876, p<0.001 (Al-Seghayer, 2001, pp. 219–220). Pair-wise comparisons showed that video-text (z= 4.018, p< 0.001) and picture-text (z=2.038, p=0.042) stimuli were superior than text-only ones (Al-Seghayer, 2001, p. 220). Furthermore, the participants performed better on video-text words than on picture-text words (z= 4.018, p< 0.001) (Al-Seghayer, 2001, p. 220).

Results from Mayer (2002) and Al-Seghayer (2001) not only evidence the positive effects of multimedia materials on learning results but also relate to the dual-coding theory (DCT). The conceptual representations of knowledge are represented in two distinctive structures, verbal and non-verbal (or imaginary), according to DCT (J. M. Clark & Paivio, 1991). Although learners have access to both systems, they don't necessarily use both. Multimedia learning materials that present information in different modes and adapt to different learning goals are conducive to effective learning (Ganan et al., 2014; Mayer, 2002). Since multimedia instructions involve multiple sensory systems of the learner when they process the materials, the use of multimedia instructions may be successful assuming that these two systems have an additive effect for superior memorization (Al-Seghayer, 2001; J. M. Clark & Paivio, 1991; de Groot & van Hell, 2005; Mayer & Anderson, 1991).

Although learning foreign languages with multimedia could be effective, more multimedia stimuli do not directly and linearly correlate with better learning outcomes. The multimedia effect is mediated by limited cognitive load (Sweller, 1994) and working memory (Baddeley, 1992; Baddeley & Logie, 1999) (cf. Al-Seghayer, 2001; Chun & Plass, 1996; Plass et al., 2003). The cognitive load theory contends that the two primary mechanisms for learning, schema acquisition and automation, are moderated by the learner's cognitive effort, i.e. cognitive load (Sweller, 1994). When the learner is attempting to develop a complex intellectual skill, they use more cognitive effort. Similarly, working memory, the brain system that temporarily stores and manipulates information, can only handle a limited number of items at a time (Baddeley, 1992).

Different multimedia designs could promote or hinder learning depending on the amount of information they present and consequently the working memory and cognitive load they demand in processing this information (Mayer et al., 2014; Plass et al., 2003). For example, Mayer et al. (2014) conducted two experiments to evaluate how low-load

instructional aid (video + audio) and high-load aid (on-screen captions + video + audio) to narrated videos could impact second language learning. On-screen captions in the foreign language are considered high-load because the additional text might divide the viewer's attention from the video and require more efforts in viewing when they are not trained to process multiple visual stimuli (Ayres & Sweller, 2014). The results show that audio + video group scored significantly higher on a content comprehension test than the audio-only control group while reporting less, though not significantly less effort (Mayer et al., 2014, p. 656). In contrast, the captioned-video group and their audio-only comparison group did not score significantly differently(Mayer et al., 2014, p. 658). Given these results, the authors of this experiment (Mayer et al., 2014) assert that high cognitive load multimedia input do not facilitate SLA.

For developers of multimedia learning experiences, it is thus imperative to consider the designs of multimedia materials according to the learners' characteristics in the best attempt to reduce their cognitive load. R. C. Clark and Mayer (2016) provide guidelines for the design and use of multimedia learning. The discussion on such guidelines, nonetheless, is beyond the scope of this paper. In the current review, I assume that app developers are aware of these theories and that they use them to their advantage.

Gamification. Gamification or gamified learning refers to "the use of game design elements in non-game environments" (Deterding et al., 2011, p. 10). Although educators sometimes integrate gamified design in physical classrooms or offline courses (Gressick & Langston, 2017; Hanus & Fox, 2015; Zainuddin, 2018), the term is more commonly used in the digital world when non-gaming software possess certain video-game characteristics. In the current paper, gamification solely refers to its application in the digital learning context. In a gamified learning environment such as a gamified language learning app, game design elements are manifested on various levels. (Deterding et al., 2011) identify these levels with examples:

Table 1 Levels of game design elements level.

Level	Description	Example
Game interface design patterns	Common, successful interaction design components and design solutions for a known problem in a context, including prototypical implementations	Badge, leaderboard, level
Game design patterns and mechanics	Commonly reoccurring parts of the design of a game that concern gameplay	Time constraint, limited resources, turns

Game design principles and heuristics	Evaluative guidelines to approach a design problem or analyse a given design solution	Enduring play, clear goals, variety of game styles
Game models	Conceptual models of the components of games or game experience	MDA; challenge, fantasy, curiosity; game design atoms; CEGE
Game design methods	Game design-specific practices and processes	Playtesting, playcentric design, value conscious game design

Note. Adapted from Deterding et al. (2014, p. 12)

An increasing number of studies has been dedicated to investigating the impact of gamification on learning. In theory, gamification influences learning by effecting changes in learning-related behaviours and attitudes (Landers, 2014). For instance, the use of badges and leaderboards can influence learners' motivation by engaging them with rewards and in competitions. In a literature review on the impact of gamification across different sectors, Hamari et al. (2014) synthesised the results from 24 peer-reviewed articles. 15 of 22 included studies that reported quantitative data found positive and significant effects for at least some of the tests and mostly for motivational measures (Hamari et al., 2014, p. 3028). Nonetheless, further examination of the data and qualitative studies revealed that the gamification effects might be short-term and indirectly affecting the desired outcomes (Hamari et al., 2014). For instance, Landers and Landers (2014) conducted an experiment in which 86 university students in an online project course were randomly assigned to a gamified group with a leaderboard (n=42) and a control group (n=46) without one. Students were measured on the times in which they engaged with the task as well as their academic performance, which was blindly rated by two researchers (Landers & Landers, 2014). The results found that the leaderboard group students were 29.61 times more likely to interact with the task than their non-leaderboard peers (p<0.001); they also found that time-on-task was positively associated with four matrix of the academic performance (r ranged from 0.343 to 0.474, p<0.01) (Landers & Landers, 2014, pp. 778–779). However, the experimental condition did not correlate with the learning outcomes, suggesting that effects of gamification might be indirect and subtle. Orhan Göksün and Gürsoy (2019) obtained similar results in an experiment where gamified teaching tools led to positive but non-significant impact on pre-service teachers' learning outcomes.

Despite the somewhat mixed outcomes in the literature on gamification, many language learning apps incorporate game elements in their design (e.g. C.-M. Chen et al., 2019; Moreno & Gatewood, 2019; Usai et al., 2018). Among the more popular commercial products, Duolingo is heavily gamified in that users earn experience points

(XP) by completing learning activities and progress accordingly on leagues and leaderboards (Rachels & Rockinson-Szapkiw, 2016; Teske, 2017). Figure 2 comprises of two screenshots from my Duolingo account: by completing learning exercises or passing in-app exams, users gain certain amount of XP that allows them to claim achievement badges (left) which are publicly displayed on their profile and enters them into competitions with other users (right). In addition, Duolingo also limits resources and sets clear goals for progress in the curriculum. The discussion on whether gamification is effective for language learning is beyond the scope of this paper. However, some included papers in this review (C.-M. Chen, Liu, et al., 2019; Purgina et al., 2020; Rachels & Rockinson-Szapkiw, 2018) centre on the effects of gamification in their studies; therefore, like multimedia learning, gamification is considered a factor in the theory of change (see 2.3. and 3.10.) in this review.

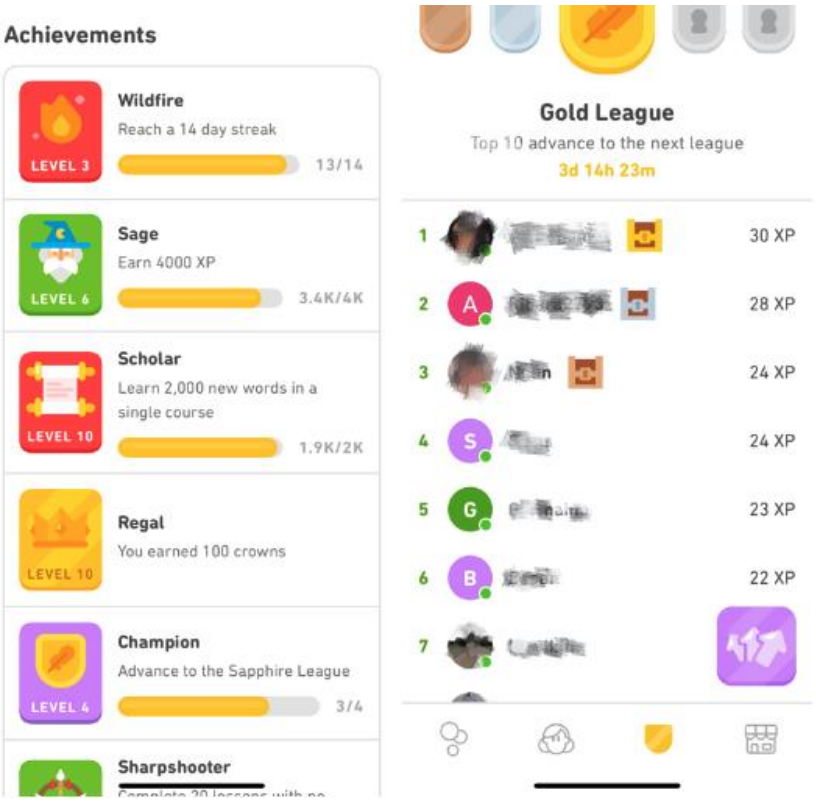

Figure 2 Illustration of gamification on Duolingo.

Note. Profiles are blurred to protect user privacy.

A crucial distinction should be noted between serious games and gamified learning (gamification). In short, serious games are full-fledged gaming contexts with a focus on education instead of entertainment and gamified learning happens in *non-gaming* context with some game elements, which were extracted based on the needs of the learning

environment (Deterding et al., 2011; Landers, 2014). In the current study, serious games are not a subject of interest.

Digital adaptive learning. In e-learning, EdSurge-Pearson (2020) defined adaptive learning as "education technologies that can respond to a student's interactions in real-time by automatically providing the student with individual support" (p. 17). Personalisation can be manifested in content, assessment, and sequence (EdSurge-Pearson, 2020). While conventional adaptive technologies try to combine the teacher's input with educational technology on the computer (Xie et al., 2019), many language learning apps now claim to provide adaptive and personalised lessons (e.g. Duolingo, Busuu, Mango Languages, Memrise). These adaptive technologies are often powered by statistical models and algorithms that consider characteristics like scores, preferences, and progress (cf. EdSurge-Pearson, 2000; Paramythis & Loidl-Reisinger, 2004).

In an early systematic review on the effectiveness of adaptive learning systems, Verdu and colleagues (2008) concluded that adaptive e-learning can significantly improve students' learning outcomes when used appropriately. Due to the rapid development in portable technologies, the review's conclusion that adaptive learning is hardly scalable is outdated; its results still offer a positive outlook for adaptive e-learning in the digital age.

Accessible input and interaction. Input plays a crucial role in second language acquisition. Although Krashen's (1982) input hypothesis lacks precision in its conceptualisation (see White, 1987 for a detailed discussion) and such factors as interaction and output (Long, 1981, 1985; Pica, 1994) are also important, few applied linguists and language teachers would deny that the quantity and quality of input significantly impact language learning outcomes (Unsworth, 2016). One of the many benefits of mobile language learning is its ability to provide learners with access to foreign language input regardless of time and place and in various forms. On the tip of their fingers, users of mobile language apps can access text, audio, pictures and videos.

Furthermore, certain apps offer opportunities for interaction, another important factor for second language learning. Studies like those of Pica et al. (1987) and Mackey (1999) point out that interaction facilitates L2 development in many aspects (see Mackey, 2007 and Pica, 1996 for discussions on interaction). For their users, some apps promote interaction with chatbots (e.g. Mondly, Memrise) or with humans (e.g. Busuu), which encourages users to produce their L2 in both speech and text.

With the latest technological advancement, language learning apps seem to hold tremendous potential in providing language learners with accessible contents and powerful facilitative learning tools. So, is there any evidence for their effectiveness so far?

2.2. Literature on language learning apps

With the popular development of MALL, scholars have conducted numerous studies on its use and effectiveness as well as many reviews of studies. Tracing MALL research to the mid-1990 when PDAs were used in L1 development, Burston (2014a, 2015) reviewed 291 MALL studies and conducted a meta-analysis on their learning outcomes. It is unclear whether Burston (2014) adopted a systematic approach in the review process, but his results offer some insight into the current state of MALL research. The results demonstrate that few MALL studies were adequately rigorous in study design, and little quantifiable information on learning outcomes was reported. Of the 19 studies that met Burston's (2014) standards for inclusion, 15 reported unequivocal positive effects from MALL and four no significant differences. Nevertheless, the small numbers of samples and short duration of treatment in these experiments undermine the reliability and validity of their results. These results corroborated with those by Viberg and Grönlund (2012). He also states that MALL applications had been geared towards teaching English to adults in teacher-oriented settings; the pedagogical focus was primarily on vocabulary acquisition (Burston, 2014). According to Burston, the state of narrowly-focused MALL was explicated in the report's title: still on the fringes (2014). A major shortcoming of this review is that its ambition to cover a comprehensive range of mobile devices ignores the heterogeneity in functions of different mobile devices. The review includes studies from two decades and on devices ranging from MP3s to smartphones. Despite that these devices share the mobility feature, the diversity in their designs and purposes can result in inaccurate inferences about their comparability.

In another review on mobile language learning in authentic (i.e. not in a lab) environments, Shadiev et al. (2017) identifies latest trends and research foci in this field. The results show that smartphones are the most popular MALL medium and that students in college and primary schools comprised the most users. Like Burston (2014a), however, Shadiev et al. (2017) focuses on studies on a variety of MALL devices. Moreover, these reviews in mobile language learning have only focused on the assisting role of devices in

language teaching. In other words, mobile devices hitherto have been an accompaniment to traditional teaching methods in the classroom and less attention has been given to full-package language learning apps on smartphones.

Although reviews on empirical studies are lacking, Heil et al.'s literature review (2016) provides an overview of the current trends and future outlook on mobile language learning apps by surveying the top 50 most popular commercial apps. The findings show that currently, the majority of language learning apps use drill-like mechanisms to teach isolated vocabulary (Heil et al., 2016). Only a few of them intend and are equipped to provide learners with a complete learning experience similar to that with a human teacher. These results are reflected in Rosell-Aguilar's (2017) taxonomy of apps for language learning:

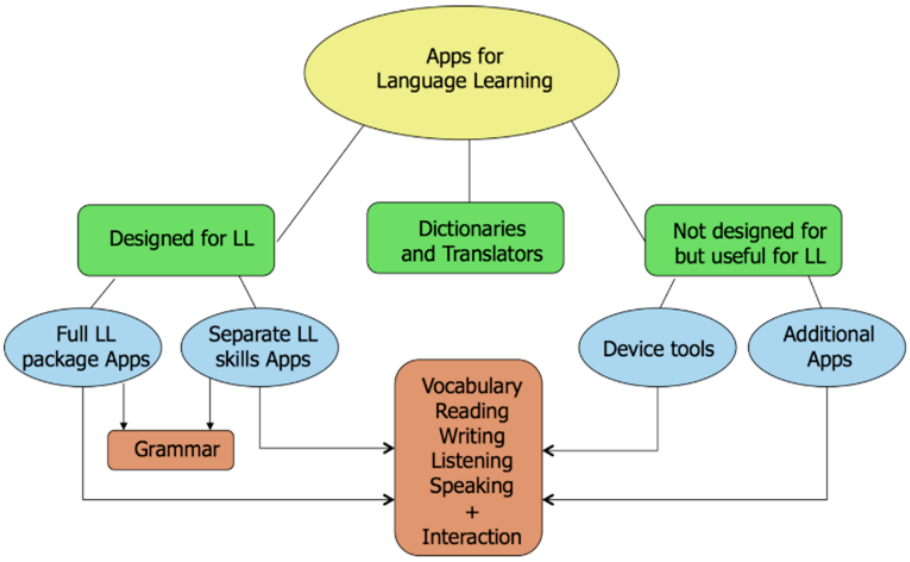

Figure 3 Taxonomy of apps for language learning.

Note. From Rosell-Aguilar (2017, p. 249).

In Rosell-Aguilar's (2017) taxonomy, three types of apps are included: those that are designed for language learning (LL), those not designed by useful for language learning, and supporting tools like dictionaries and translators. The current review is interested in the first category: apps designed for language learning. In this category are two subcategories: full LL-package apps and separate LL skills apps. The former refers to apps that are, in theory, capable of replacing human teachers and offer language learning solutions in all language skills; examples of these apps include Duolingo and Busuu (Rosell-Aguilar, 2017). Apps for separate language skills are those that only provide services in certain language skills, such as vocabulary drilling and listening training; examples of this category include 501 Spanish verbs and Conjuverb (Rosell-Aguilar,

2017). In this review, I include studies that use both types of language learning apps for analysis and I extract information about the measured language skills.

2.2.1. Gaps in the literature

Numerous studies have endeavoured to measure the effectiveness of language learning apps. Rachels and Rockinson-Szapkiw (2018), for instance, compares the learning outcomes of primary school students learning Spanish on Duolingo and with a teacher. Vesselinov and colleagues (e.g. 2016b, 2019b) conducted a line of studies on commercial apps with current users to evaluate their efficacy. Gangaiamaran and Pasupathi (2017) have attempted to review the empirical evidence of language learning apps by surveying selected studies. They classify the apps based on the age groups of target users (primary, secondary, and tertiary) and concluded that the listening skill is best acquired through these apps (Gangaiamaran & Pasupathi, 2017). There are several flaws in their review: the categorisation of apps has no basis in the app design, as most apps do not explicitly state the age of their target users. Moreover, the authors do not provide an explanation on the methodology with which they selected and analysed their studies and data, leaving doubts for the validity of this review. Although they provide a general conclusion on the current empirical findings, only narrative descriptions were included without any attempt in synthesis.

Thus, the current systematic review seeks to fulfil this gap in the literature by providing a scrutiny of the existing empirical evidence on the effectiveness of these apps.

2.3. A theory of change

A theory of change (Weiss, 1998, p. 98) refers to "the chain of causal assumption that link programme resources, activities, intermediate outcomes and ultimate goals." Popay et al. (2006) recommend developing a theory of change in a systematic review to inform research question development and decisions in developing inclusion criteria. Consulting previous literature and relevant studies on the topic, I theorise that mobile apps could be effective for language learning because it can provide accessible, multimedia input and personalised learning. Mobility allows users to access learning activities regardless of time and space. Features like gamification and multimedia materials might increase students' motivation in learning. Personalisation can offer guidance on progress and performance. All these results are conducive to effective language acquisition.

2.4. Research questions

The review sets out to answer the following questions:

RQ1. What is the extent of the literature on the effectiveness of language learning applications on smart devices?

RQ2. What is the evidence of their effects on outcomes related to language proficiency?

RQ2a. Is the effectiveness of these apps different for different language skills?

RQ2b. Is the effectiveness of these apps mediated by user characteristics like age?

CHAPTER 3. Methodology

To address the research questions, I chose to conduct a systematic review on the evidence of language learning apps' effectiveness and efficacy. Effectiveness and efficacy research are two types of evidence-based studies commonly used in evaluating the impact of a treatment (Institute of Education Sciences & Foundation, 2013). Efficacy studies implement the treatment under ideal and highly controlled conditions and answers the question "how well does the intervention work in a strictly controlled environment?" In contrast, effectiveness studies apply the treatment under circumstances closer to the reality. In other words, effectiveness studies answer the question "how well can the intervention work in the real world?" For the purpose of this review, I do not distinguish efficacy and effectiveness studies and use the term "effectiveness" generally to refer to both types of research.

A systematic review follows rigorous and systematic methods to assess the empirical evidence on a given topic and with pre-determined eligibility criteria (Boland et al., 2017). Although the tradition of systematic reviews originated in medicine and health to evaluate different medical treatments, social scientists have adapted the methodology to assess empirical studies in their disciplines. I followed Boland et al.'s (2017) guidebook on conducting systematic reviews because it was specifically written for graduate student projects. I also followed Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) to generate a protocol to guide this project (Moher et al., 2016; Shamseer et al., 2015) (see Appendix A for the protocol).

Meanwhile, to align closely with practices in education and applied linguistics, I consulted resources from Education Endowment Foundation (hereafter EEF, n.d.-a), an independent English organisation that has commissioned numerous evidence reviews and whose works are recognised in the education sector. In addition, I utilised resources by The Campbell Collaboration (n.d.), a non-profit organisation that specialises in producing systematic reviews in various disciplines to inform policymakers and scholars. I also consulted resources by Cochrane Collaboration, a recognised foundation in public health and that provides high-quality, validated resources for systematic reviews. In this chapter, I outline my methodological decisions and the review process.

3.1. Eligibility criteria

Acknowledging that studies with the highest standards of rigor, i.e. using double-blind randomised controlled trials (RCTs), could be scarce in educational research, I adopt the

best-evidence approach (Slavin, 1986) in my review. In a best-evidence synthesis, studies that meet minimal methodological standards are included and a systematic assessment of individual studies' quality and validity is conducted. Consequently, I chose to include primary studies published after 2007 that reported quantitative data. Articles were not excluded based on their language and publication status. See below for a complete list of eligibility criteria with rationale.

Table 2 Eligibility criteria

Item	Inclusion	Exclusion	Rationale
Complete reference	Include studies with a complete reference.	Exclude studies with an incomplete reference.	
Time	Include studies published after 2007.	Exclude studies published before 2007.	The first smartphone, the iPhone, was released in 2007. Therefore, it provides a suitable watershed moment for investigating the use of mobile apps on smartphones.
Publication status	Do not exclude studies based on their publication status.		I want to include the grey literature on this topic to minimise the impact of publication bias.
Language	Do not exclude studies based on the language in which they are written. However, if the main text of the study is in languages other than those I read (English and Chinese), I would make efforts to obtain the translation of the text or to ask a speaker of that language to perform data extraction.		
Duration	Include studies that are longitudinal, with any duration.	Exclude studies that are not longitudinal, for example, one-off cross-sectional studies.	Cross-sectional studies cannot show improvement.
Study design	Include primary studies that use an RCT, non-randomised comparison, or one group, pre-test/post-test design, with, at minimum, one experimental group. Include one cohort, post-test-only design. Include mixed methods studies that comprise of mentioned designs. Include studies that use intact groups as subjects.	Exclude observational studies, ethnographies, secondary data analyses, and studies that only collected qualitative data. Exclude single case studies where individuals are the subjects.	I want to investigate the direct relationship between language learning apps and language proficiency on the basis of substantive, measurable outcomes such as test scores. Primary studies with experimental groups are the more reliable design for making these kinds of causal inferences. I include one cohort, pre-test/post-test designs because they are prevalent in the scoping search. The risk of bias based on study design will be evaluated.
Participants	Include studies on typically developing language learners of all ages and	Exclude studies that exclusively recruit non-typically developing participants, for example	Since the feature of mobile language learning is "anytime, anywhere," it's reasonable to infer that they are widely

	background. Include studies that use intact groups as subjects, if they are reasonably assumed to comprise of typically developing students.	those with physical disabilities or Developmental Language Disorder.	accessible to all populations of language learners. However, non-typically developing populations might exhibit traits and factors not applicable to the wide populations. Therefore, I will not include studies that only recruit participants from this population.
Intervention	Include studies that use language learning apps on smart phones or tablets as the intervention for second/foreign language learning for the experimental group. These applications have self-contained content and curriculum which allow users to progress. Examples include Duolingo and Busuu. These apps do not have to be commercially available as long as they have the potential to be used by all populations.	Exclude studies that do not use an app on mobile devices as the intervention and those that *only* use the website versions of the apps. Exclude studies that use apps that require user-generated content and those that only serve as tools to enhance learning experience (e.g. Anki, Quizlet, dictionaries). Exclude studies that use mobile devices as platforms for teaching or as an addition to other methods of teaching. Exclude studies that study first language skills, such as literacy.	Some applications have a website version with different layouts and features from their mobile counterpart, which could impose additional confounding variability to the study results. However, studies that do not limit the platform on which they use the app can be included because it is to the participants' discretion and convenience which one to use, a core feature of mobile learning. Moreover, participants in studies that use mobile devices as platforms or additional teaching tools might benefit from interaction with instructors, which presents additional confounding variability.
Outcome	Include studies that report quantitative data of post-tests on the language skills they seek to measure.	Exclude studies that don't report any quantitative data and those that report survey results as quantitative data. Exclude studies that report only conclusions drawn from quantitative data.	I want to investigate the direct relationship between language learning apps and increased language proficiency on the basis of substantive, measurable outcomes such as test scores. Access to quantitative data allows for cross-study comparison.

3.2. Information Sources

To minimise the impact of publication bias, which refers to the phenomenon that studies finding significant effects are favoured for peer-reviewed publication, I searched beyond peer-reviewed journals to include the grey literature. Grey literature sources include but are not limited to conference proceedings, commercial and/or government reports, trial register (if any), master's and doctoral theses.

Since the first smart phone was released in 2007, the literature search only included those published after this date. Because of this date restriction, it was reasonable to assume that all relevant literature was published electronically; therefore, I only conducted the search on electronic databases.

3.3. Search strategy

3.3.1. List of databases

I conducted the main search on the following databases, which cover fields of education, language and linguistics, psychology, computer science and general social sciences. I generated this list by selecting those that are relevant on the Search Oxford Libraries Online (SOLO) system under categories of education, linguistics and social sciences. I also consulted with an experienced librarian to remove redundant and irrelevant databases, such as those indexing only books or newspapers.

Table 3 Main search databases

Discipline/Category	Database(s)/Journals
Education	The Education Collection on SOLO (incluing ERIC), British Education Index, Child Development & Adolescent Studies, Education Abstracts, EThOS[2]
Linguistics	Linguistics Collection (including LLBA), *Journals*: CALICO, ReCALL, Language Learning and Technology[3]
Psychology	PsycINFO
General social sciences	ProQuest Social Science Premium Collection[4], ASSIA Applied Social Sciences Index and Abstracts, SCOPUS, Sociological Abstracts
Multidiscipline	Web of Science Core Collection, Google Scholar
Conference proceedings	Oxford University Research Archive, PapersFirst, ProceedingsFirst (OCLC)
Theses and dissertations	ProQuest Dissertations & Theses Global

I conducted a simple search at the advice by the librarian, i.e. not using the full search string, on SCOPUS, PsycINFO and Child Development & Adolescent Studies and Sociological Abstracts, to determine if they could be relevant databases for a full search. The simple search generated no relevant records and those databases were excluded from the mains search. I also used Google Scholar to look for additional citations. Furthermore, I conducted backward and forward citation searches. Where I identified a study that fit all

[2] EThOS is included in ProQuest Dissertations & Theses Global

[3] Originally planned for hand searching because of their high relevancy to the topic, but the journals were fully included in the ProQuest Social Science Premium Collection

[4] The ProQuest Social Science Premium Collection includes Linguistics Collection (including LLBA), ASSIA Applied Social Sciences Index and Abstracts, and Sociological Abstracts

criteria for inclusion, I searched in its reference list as well as looked for subsequently published research that have cited this article for eligible studies. In my scoping search, I found that many commercial language learning apps such as Duolingo and Busuu have commissioned their own efficacy studies and their reports are publicly available. Therefore, I also included these reports.

I set automatic updates on searched databases for newly published articles during the review process. Given the deadline of this project, my last update on the main search was on 15th May 2020.

3.4. Search terms

The tradition of systematic reviews is rooted in medical research, in which methodologies are more meticulously defined. Terminologies and medical tests are generally clearly defined and standardised. Ambiguity in terminologies results from updates in the industry and language differences. On the other hand, for an emerging topic in the social sciences, there could be a plethora of terms and methods for similar phenomena. Scholars that follow different academic conventions and schools of thoughts might employ different terms when referring to the same subject. For example, m-learning and mobile-assisted learning are similar in definition in that they both refer to learning on mobile devices. Learning and acquisition are also used interchangeably by some scholars in the field of SLA. Additionally, the focus of the research question might change the perspective through which the researcher defines their subjects. For instance, researcher who focuses on the gamification feature of a language learning app could define the app as a gamified app, whereas researchers exploring the mobility feature of the app define the app as a mobile app. Some scholars might also adhere to more traditionally defined terms such as MALL to acknowledge the legacy of existing literature.

Therefore, one of the main challenges in developing my search terms was to address the diversity of language use in this field while locating the most relevant studies.

In consultation with an experienced librarian in our department and through library skills training and preliminary research, I identified a number of concepts central to the research question of this study:

- Purpose: language learning
- Medium and features of the medium: mobile app(lication)s
- Outcome: effectiveness

After brainstorming search terms with the librarian, we tested their validity and efficiency using the SOLO Linguistics Collection. I modified the search terms according to the

search results so that they are neither too broad nor too narrow. Moreover, I also compared the search results against a candidate article (Rachels & Rockinson-Szapkiw, 2018) generated from the scoping search, which passed all inclusion criteria to validate the search terms. The final search terms are listed in Table 4 (note that boolean operators are included to identify their relationships):

Table 4 Concepts and search terms

Purpose		Medium/features	Outcome
language* OR vocabulary OR grammar OR syntax	learn* OR acquisition OR acquir*	mobile OR app OR apps OR smart?phone OR iPad OR iPhone OR Android OR iOS OR gamification OR gamified OR game?based OR m-learnig OR e-learning OR MALL OR mobile?assisted OR busuu OR duolingo OR memrise OR babbel	effect* OR impact OR efficacy OR efficien* OR develop*

3.4.1. Search string

I connected search terms with boolean operators to coin a search string for the electronic databases. Concepts were connected with AND and search terms within each concept were connected with OR. Boolean operators allow for combination of keywords to increase the relevance of search results[5]. To further increase accuracy of the search, I applied search limitations: keyword search was restricted to anything but full article (*noft*) if possible, and publication date was restricted to after 01/01/2007. The search string is as follows:

> noft(language* OR vocabulary OR grammar OR syntax) AND
> noft(learn* OR acquisition OR acquir*) AND
> noft(mobile OR app OR apps OR smart?phone OR iPad OR iPhone OR Android OR iOS OR gamification OR gamified OR game?based OR m-learnig OR e-learning OR MALL OR mobile?assisted OR busuu OR duolingo OR memrise OR babbel) AND
> noft(effect* OR impact OR efficacy OR efficien* OR develop*)

3.5. Study records

3.5.1. Data management

After performing the main search on the identified databases, I uploaded the generated citations to Rayyan, where I resolved duplicate and conducted screenings. Rayyan is an Internet based software program for systematic review screening and facilitates collaboration among multiple reviewers (Ouzzani et al., 2016). I used Rayyan for

[5] For a detailed guide on boolean, see Bodleian Libraries' (2020) guide on subject searching.

screenings and Mendeley, a reference manager, for full article review. Before the review process, I collaborated with my supervisor on Rayyan to ensure that I was familiar with the functions and procedure of screening citations on the software.

I used spreadsheets on Microsoft Excel to perform data extraction and Tableau Desktop Professional Edition (2020.2.4) to generate graphs for results.

At the same time, I kept track of my actions and records, including search strings, in a Microsoft Word research journal. All data management and analysis were completed on a MacBook Pro with a Mac OS Catalina 10.15 operating system.

3.6. Selection process

3.6.1. Title and abstract Screening

After removing the duplicates from main-search generated citations, I conducted the first round of screening: titles and abstracts generated were checked against the inclusion/exclusion criteria. They were included if they met all inclusion criteria and excluded if they matched any of the exclusion criteria. Records that couldn't be reliably excluded on the basis of the information in their titles and abstracts were included for full-text screening. I tried to obtain full reports for all titles that appeared to meet the inclusion criteria or where there was any uncertainty.

3.6.2. Full-text Screening

After the title and abstract screening, I tried to obtain the full articles of included studies from databases identified in Table 3. If full articles were not available online in those databases, I tried to contact the author for the full text.

Then, I assessed the full articles against the inclusion criteria. I sought additional information from study authors where necessary to resolve questions about eligibility. Again, the articles were labelled according to the inclusion/exclusion criteria.

3.6.3. Double screening by a second reviewer

In both screenings, I invited an independent second reviewer. He is trained in applied linguistics and education research and we discussed the details and requirements about the review before screenings. He screened 5% of the total items (357 items) in the title and abstract screening and 5% (8 articles) in the full-text screening. A higher percentage wasn't possible because of time constraint. We both recorded the reasons for exclusion. Reviewers were blind to the decisions of one another during this process.

I then compared my decisions to those of the second reviewer. Since the purpose of this dual-review is to test the validity and the clarity of the eligibility criteria, and the number

of dual-reviewed articles is low, I sought a high simple inter-reliability of 0.8-1.0 (McHugh, 2012). The inter-rater reliability was 0.97 for the title and abstract screening 0.97 and 0.75 for the full-text screening. The second reliability rate was lower than ideal

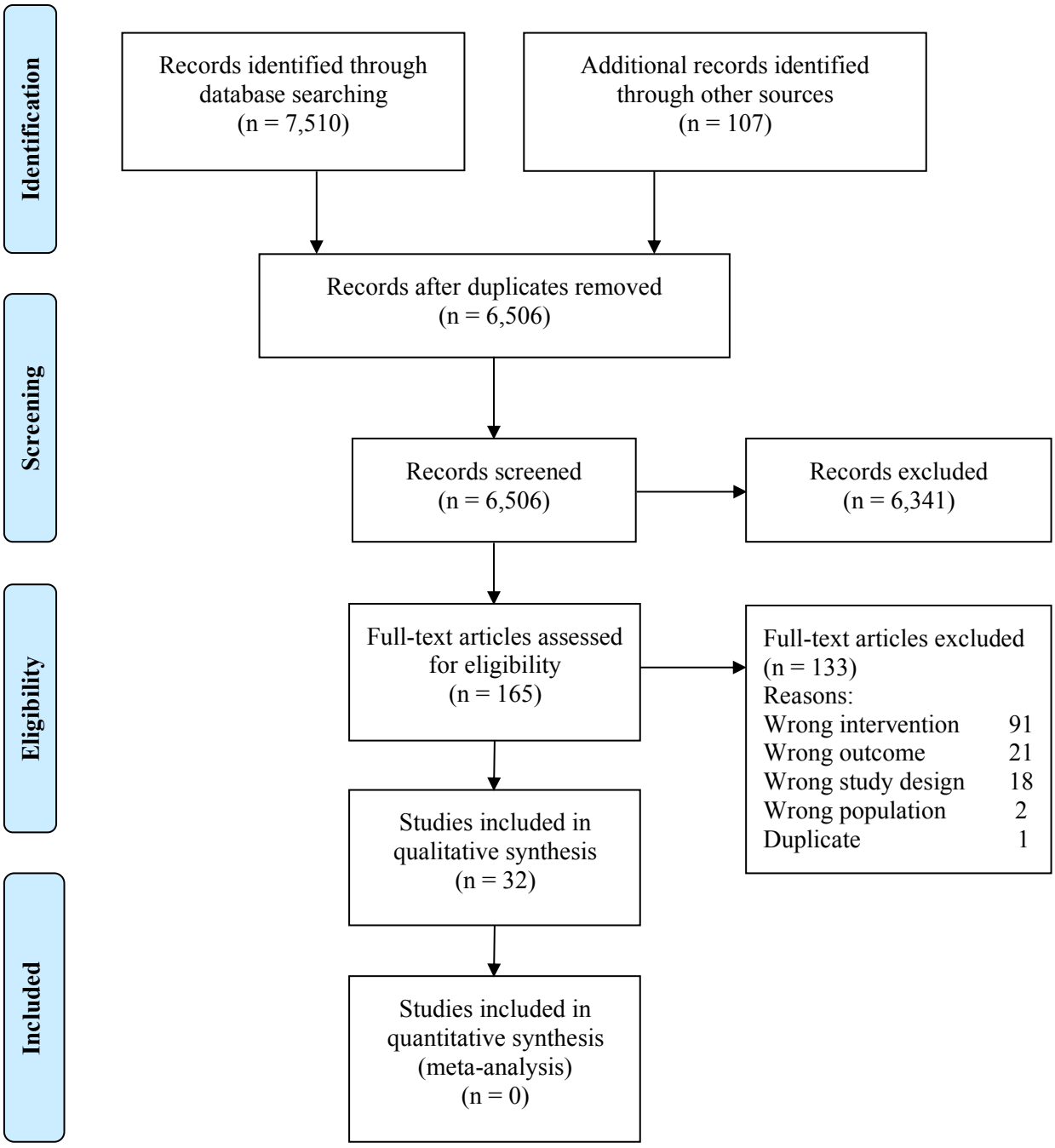

Figure 4 Flow diagram of the screening process.

Note. Template from Moher et al. (2009).

but within the acceptable range given that the number of screened articles was very small. I resolved all conflicts and rectify my eligibility criteria where more clarity was needed through discussions with the second reviewer.

Figure 4 shows the screening process with the numbers of studies included and excluded at each stage. As the diagram displays, 7,510 studies were identified through electronic databases and 107 through the scoping search and backward and forward citation searches. I resolved duplicates on Mendeley Desktop and then Rayyan, leaving 6506 titles and abstracts for the first screening. I excluded 6341 articles based on their titles and abstracts, leaving 165 for full-text screening. Subsequently, I excluded 133 articles based on their full texts. Among these articles, 91 had the wrong intervention, 21 wrong outcomes, 18 wrong study designs, and two wrong population. Two articles were two versions of the same study, because the earlier version was included as a doctoral dissertation in the database and the later a peer-reviewed journal article (Rachels & Rockinson-Szapkiw, 2016, 2018); I included the peer-reviewed version for review. In the end, 32 articles were included for qualitative, narrative synthesis. A meta-analysis was not possible because of the high heterogeneity in the study designs (including demographic, sample sizes and interventions) and outcome measurements of the studies.

3.7. Data extraction

Using a pre-determined data extraction form, I extracted data from studies onto an Excel spreadsheet. Extracted information includes study characteristics, participant characteristics, and the outcomes. See Appendix A for the detailed list of items.
It was not possible to have a second reviewer perform data extraction due to time constraint although it was planned in the protocol.

3.8. Outcomes and prioritisation

The primary outcomes are quantitative measures of language learning outcomes, for example, standardised test scores.

3.9. Risk of bias of individual studies

Bias refers to "a systematic error, or deviation from the truth," that can positively or negatively affect the results of a study and vary in magnitude (Boutron et al., 2019, sec. 7.1). In a systematic review, sources of bias include the actions of primary study authors, conflicts of interests, study designs, and actions of the reviewer. Thus, including an assessment of the risks of bias of individual studies in a systematic review is crucial because it informs the audience of the trustworthiness of the studies' findings. A review on the risk of bias assessment tools used in Campbell systematic reviews in the social sciences found that given the lack of discipline standard, the choice of quality assessment tools should depend on the topic and needs of the review (Li et al., 2017). Reviewers in

the social sciences widely adopt tools from health sciences, such as the Cochrane Risk of Bias Tool (Li et al., 2017). For this project I followed Chalmers (2019) in using the Effective Public Health Practice Project Quality Assessment Tool for Quantitative Studies (hereafter EPHPP tool; Thomas et al., 2004) because it is suitable for quantitative studies with different research designs. Moreover, Petticrew and Roberts (2006) recommend EPHPP among five other tools for systematic reviews in the social sciences. The EPHPP tool assesses the risk of bias of a study in the following eight components (2003):

A. Selection bias
   a. Are the individuals selected to participate in the study likely to be representative of the target population?
   b. What percentage of selected individuals agreed to participate?

B. Study design
   a. What is the study design?
   b. Was the study described as randomised?
   c. If Yes, was the method of randomization described?
   d. If Yes, was the method appropriate?

C. Confounders
   a. Were there important differences between groups prior to the intervention?
   b. If yes, what percentage of relevant confounders was controlled?

D. Blinding as part of a controlled trial
   a. Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?
   b. Were the study participants aware of the research question?

E. Data collection practices
   a. Were data collection tools shown to be valid?
   b. Were data collection tools shown to be reliable?

F. Withdrawals and dropouts
   a. Were withdrawals and drop-outs reported in terms of numbers and/or reasons per group?
   b. what is the percentage of participants completing the study?

G. Invention integrity
   a. What percentage of participants received the allocated intervention or exposure of interest?
   b. Was the consistency of the intervention measured?
   c. Is it likely that subjects received an unintended intervention (contamination or co-intervention) that may influence the results?

H. Analysis
   a. What is the unit of allocation?
   b. What is the unit of analysis?
   c. Are the statistical methods appropriate for the study design?
   d. Is the analysis performed by intervention allocation status (i.e. intention to treat) rather than the actual intervention received?

On all components except for G and H, the study first receives a local rating of "weak," "moderate," or "strong." Components G and H are not given ratings because while these constructs are important for understanding the results of included studies, potential deviations in these components do not constitute systematic errors caused by the researchers' design and implementation. See Thomas et al. (2004, p. 180) for a detailed explanation. Then, the study receives a global risk of bias rating of "strong," "moderate," or "weak." "Strong" studies have no weak ratings and at least four "strong" ratings; "moderate" studies have less than four "strong" ratings and/or one "weak" rating, and "weak" studies have two or more "weak" ratings (Thomas et al., 2004, p. 180). "Weak," "moderate," and "strong" refer to the strength of security of the study; the results of "weak" studies are less trustworthy than those of "moderate" and "strong" studies.

Like Chalmers (2019, p. 87), I made minor changes to the terminology and rating guidelines in my assessment. I used "non-randomised comparison" instead of "controlled clinical trial" in the Study Design component so that the term would be more applicable for educational research. For the Analysis component, I used "learner, class/group, school, local authority/school district" in lieu of "individual, practice/office, institution/organization, community." Under Blinding, if the tests for relevant outcomes did not required human graders (such as to grade pronunciation or fluency), the answers to D was defaulted to "not applicable." (See Appendix C and Appendix D For a detailed template and dictionary of this tool by the EPHPP).

Due to time constraint, it was not possible to have the second reviewer conduct this assessment although it was planned in the protocol.

3.10. Data synthesis

Following guidelines from Popay et al. (2006), I conducted a narrative synthesis with extracted data and use tools such as tables and graphs to explain the findings of included studies. According to Popay et al.'s (2006) guide, a narrative synthesis has four components:

- Developing a theory of how the intervention works, why and for whom.
- Developing a preliminary synthesis of findings of included studies.
- Exploring relationships in the data.
- Assessing the robustness of the synthesis.

The narrative synthesis roughly follows the above structure and serves as the main component of this review.

In the protocol, I specified that if included studies use sufficiently homogeneous study designs and reported measures for outcomes, I would conduct meta-analyses on the effect sizes of their outcomes. When effect sizes are not readily available in the included studies, I would convert outcomes to effect sizes using available information in the study. The included studies are, however, extremely heterogeneous in study characteristics. Thus, no meta-analyses are included in this review.

CHAPTER 4. Results

In this section, I present the characteristics of included studies with tables, graphs, and textual description. 32 studies are included in the analysis:

1. Bhide, A. (2017). Instructional methods for promoting the development of orthographic and phonological knowledge in second language learners of Indic languages [University of Pittsburgh]. In *ProQuest Dissertations and Theses* (Issue 10645799).
2. Çakmak, F., & Erçetin, G. (2018). Effects of gloss type on text recall and incidental vocabulary learning in mobile-assisted L2 listening. *ReCALL, 30*(1), 24–47.
3. Cavus, N., & Ibrahim, D. (2017). Learning English using children's stories in mobile devices. *British Journal of Educational Technology, 48*(2), 625–641.
4. Chang, W. C., Chen, C. M., & Yang, S. M. (2018). An English vocabulary learning app with self-regulated learning mechanism for promoting learning performance and motivation. *Proceedings - 2018 7th International Congress on Advanced Applied Informatics, IIAI-AAI 2018, 32*(3), 164–169.
5. Chen, C.-M., Liu, H., & Huang, H.-B. (2019). Effects of a mobile game-based English vocabulary learning app on learners' perceptions and learning performance: A case study of Taiwanese EFL learners. *ReCALL, 31*(2), 170–188.
6. de Jong, T., Specht, M., & Koper, R. (2010). A study of contextualised mobile information delivery for language learning. *Educational Technology and Society, 13*(3), 110–125.
7. García-Mencía, R., López-López, A., & Muñoz Meléndez, A. (2016). An Audio-Lexicon Spanish-Nahuatl: using technology to promote and disseminate a native Mexican language. In S. Papadima-Sophocleous, L. Bradley, & S. Thouësny (Eds.), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 155–159). Research-publishing.net.
8. Guaqueta, C. A., & Castro-Garces, A. Y. (2018). The use of language learning apps as a didactic tool for EFL vocabulary building. *English Language Teaching, 11*(2), 61.
9. Hao, Y., Lee, K. S., Chen, S. T., & Sim, S. C. (2019). An evaluative study of a mobile application for middle school students struggling with English vocabulary learning. *Computers in Human Behavior, 95*, 208–216.
10. Hsu, C. K., Hwang, G. J., & Chang, C. K. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. *Computers and Education, 63*, 327–336.
11. Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French/y/. *CALICO Journal, 32*(1), 1–25.
12. Liu, J. (2016). Designing intelligent language tutoring system for learning Chinese characters [Iowa State University]. In *ProQuest Dissertations and Theses* (Issue 10167832).

13. Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL, 31*(3), 293–311.
14. Martinelli, M. (2016). Effectiveness of online language learning software (Duolingo) on Italian pronunciation features: A case study [Oklahoma State University]. In *ProQuest Dissertations and Theses* (Issue 10191215).
15. Mohamad Ali, A. Z., Segaran, K., & Hoe, T. W. (2016). Effects of verbal components in 3D talking-head on pronunciation learning among non-native speakers. *Journal of Educational Technology & Society*, *18*(2), 380–389.
16. Moreno, I. D., & Gatewood, K. (2019). Gamification of foreign language mobile-learning: A quantitative study on Spanish-as-a-foreign language adult learners' satisfaction, perception, motivation, and knowledge [Keiser University]. In *ProQuest Dissertations and Theses* (Issue 13897893).
17. Nicholes, J. (2016). Measuring the impact of language-learning software on test performance of Chinese learners of English. TESL-EJ, 20(2), 1–20.
18. Ou-Yang, F. C., & Wu, W. C. V. (2017). Using mixed-modality vocabulary learning on mobile devices: Design and evaluation. *Journal of Educational Computing Research, 54*(8), 1043–1069.
19. Purgina, M., Mozgovoy, M., & Blake, J. (2020). WordBricks: Mobile technology and visual grammar formalism for gamification of natural language grammar acquisition. *Journal of Educational Computing Research, 58*(1), 126–159.
20. Rachels, J. R., & Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning, 31*(1), 72–89.
21. Ramírez, G., ElMiligi, H., Walton, P., & Billy, C. (2018). Revitalization of an indigenous language: Which is better the teacher or the app? *International Conference on E-Learning*, 2018-July, 344-350, XVII.
22. Terantino, J. (2016). Examining the Effects of Independent mall on vocabulary recall and listening comprehension: An exploratory case study of preschool children. *CALICO Journal, 33*(2), 260–277.
23. Thongsri, N., Shen, L., & Yukun, B. (2019). Does academic major matter in mobile assisted language learning? A quasi-experimental study. *International Journal of Information and Learning Technology, 36*(1), 21–37.
24. Vesselinov, R., & Grego, J. (2015). *Efficacy of a New Language App*.
25. Vesselinov, R., & Grego, J. (2019). *Pimsleur Efficacy Study* (Issue April).
26. Vesselinov, R., & Grego, J. (2019). *Mango Languages Efficacy Study*.
27. Vesselinov, R., & Grego, J. (2016). *The Babbel Efficacy Study* (Issue September).
28. Vesselinov, R., & Grego, J. (2012). *Duolingo Effectiveness Study*.
29. Vesselinov, R., & Grego, J. (2016). The Busuu Efficacy Study (Issue May).
30. Vesselinov, R., Grego, J., Sacco, S. J., & Tasseva-Kurktchieva, M. (2019). *The 2019 Rosetta Stone Efficacy Study*.
31. Wu, Q. (2015). Pulling mobile assisted language learning (MALL) into the mainstream: MALL in broad practice. *PLoS ONE, 10*(5).

32. Wu, Q. (2015). Designing a smartphone app to teach English (L2) vocabulary. *Computers and Education, 85*, 170–179.

4.1. Description of study characteristics

Table 5 and Table 6 lists basic information of these studies, including their publication status, intervention details (age group of participants, sample size, location of study, app's details, and the studied language[s]). Table 5 lists studies that use different apps or app versions as experimental and control interventions, while Table 6 shows those that compare effectiveness of apps against other methods of language learning or with repeated measures.

4.1.1. Risk of bias of individual studies

Both Tables 5 and 6 group studies by their risk of bias (RoB) ratings. Table 5 shows the 10 studies that compare apps and app versions and groups them by global risk of biases. Seven studies[1,2,4,5,6,12,17] have a "weak" global rating and three[10,15,16] have a "moderate" rating. No studies in this category received a "strong" rating. Table 6 lists the remaining 22 studies, which do not use different apps or app versions as interventions. In this category, 11 studies[3,7,8,9,14,18,19,22,23,31,32] have a "weak" rating, seven[13,21,24,25,26,28,29] have a "moderate" rating, and only four[11,20,27,30] have a "strong" rating. Overall, over half (18/32) of included studies received a "weak" rating, while 10 received a "moderate" rating and four a "strong" one.

4.1.2. Publication status

Figure 5 shows that over half (19) of the included studies are peer-reviewed articles, the number of which is followed by those of organisation report (7)[24,25,26,27,28,29,30], master's thesis (2)[12,14], doctoral thesis (2)[1,16], and conference proceeding (2)[7,21]. Organisation reports refer to those published by an organisation but not included in any formal conferences or accepted to any academic journals. In the current review, all studies that classify as organisation reports came from a research team at Queens College, City University of New York and the University of South Carolina, who established a line of studies to evaluate the effectiveness and efficacy of commercial language learning apps (studies 24-30, see Table 5).

4.1.3. Age group and sample size

Figure 6 shows the age groups of included participants in each study as well as the sum of samples in each age group. As displayed, the majority (71.88%, 23) of studies had adult participants, which comprise the largest number of participants across studies (2,191).

18.75% of the studies recruited children younger than 12 years old as participants, totalling 756 children across six studies[1,2,4,20,21,22]. Finally, three studies[8,9,10] (9.38%) sampled 138 teenagers aging 12 to 18. It is noted that this categorisation lacks reflection of detailed segmentation of adult populations. Some studies (e.g. Çakmak & Erçetin, 2018; C.-M. Chen et al., 2019; Wu, 2015a, 2015b) specifically recruited university-age students while others (e.g. de Jong et al., 2010; Loewen et al., 2019; Vesselinov & Grego, 2019) had adult participants from a broader age range. Further segmentation of adult age groups would not be appropriate, and the three age groups is the minimally meaningful grouping for the current review.

The smallest sample size in the studies is five[14] and the largest is 467[20]. The average sample size across 32 studies is 44.94 (SD=98.23).

4.1.4. Location of study

Figure 7 shows the locations of included studies, where the deeper the colour and the bigger the size of the square, the more frequently studies were conducted at that location. Five studies do not specify the location of study. Seven studies[12,13,20,22,25,27,29] were at least partially conducted in the USA (two[27,29] are included in Multiple Locations in Figure 7), five in Taiwan[4,5,9,10,18], four in Mainland China[17,23,31,32], two in Canada[11,21], and one each in Germany[27], the UK[29] (both included in Multiple Locations, see Table 5 for details), Colombia[8], Japan[19], Turkey[2], Cyprus[3], and India[1]. Moreover, four studies[16,24,28,30] were conducted exclusively online. Five studies[6,7,14,15,26] does not have information on their location of study.

4.1.5. Studied language(s)

The studied languages are shown in Figure 8. Nine languages were subjects of interests for the included studies, in which English is, unsurprisingly, the most popular (14 studies[2,3,4,5,8,9,10,15,17,18,19,23,31,32]), followed by Spanish[16,20,22,24,25,26,27,28,29,30] (10), Hindi[1,6] (2), and one each in Italian[14], Chinese Mandarin[12], French[11], Turkish[13], Secwepemctsín[21] and Nahuatl[7].

4.1.6. Type of study design

As Table 5 and Table 6 demonstrate, the study designs of included studies are highly heterogeneous. Overall, 11 studies[2,4,6,11,16,17,19,20,21,31,32] were RCTs, of which five[2,4,6,16,17] compare apps or app versions. The second most popular design is cohort (one group pre-test/post-test) design, also used by 11 studies[7,8,13,14,24,25,26,27,28,29,30] and all of which do not compare apps or app versions. Eight studies[1,3,5,10,12,15,18,23] chose to conduct non-

randomised comparison, five[1,5,10,12,15] of which compared apps or app versions. At last, one[9] study used other methods: Hao et al. (2019) studied a cohort's performance after using one app with a post-test only. The categorisation of study designs depends on the actual implementation of the study according to the author's description, not what the author has specified in the methods section. Details for each study's study design will be summarised in 4.2.

4.1.7. Duration of intervention

The duration of intervention for included studies ranges from less than a day (labelled as one day) to six months. Six studies[2,6,9,12,15,21] conducted and finished their experiments within a day and one within a week[7]. Three studies[4,14,18] had a two-week intervention and one[23] for three weeks. Two studies'[11,16] interventions were five weeks. The most popular duration of intervention is two months, or eight weeks, as represented in eight studies[24,25,26,27,28,29,30,32]. Three studies[13,20,31] conducted their treatment over three months (labelled as 12 weeks) and one[17] over 15 weeks. The longest duration is six months, or 24 weeks, which is seen in two studies[8,22]. Finally, one study[19] does not specify its duration of intervention.

4.1.8. Measured language skills

Figure 9 summarised the target language skills in studies. Nine types of language skills are of interest to researchers. Many authors employed several tests to measure a number of language skills in their study. One study[3] measured (content) comprehension skill. However, the author does not specify whether it refers to reading or listening comprehension (Cavus & Ibrahim, 2017). Two studies[1,12] investigated learners' orthographical knowledge such as character recognition and construction. Four studies[9,13,17,20] surveyed writing skills and five[1,3,11,14,15] on phonology-related knowledge and skills such as pronunciation, phoneme perception and production. Six studies[9,13,17,25,26,29] measured speaking skills and seven measured listening[2,3,7,9,13,17,22] skills. Ten studies[9,10,13,17,24,26,27,28,29,30] measured learners' reading performance while 11 measured lexicogrammar knowledge such as sentence construction and grammatical judgement[13,16,17,19,20,24,26,27,28,29,30]. The most popular measured language skill is vocabulary, which is measured by 21 studies[1,2,3,4,5,6,7,8,18,20,21,22,23,24,26,27,28,29,30,31,32].

4.1.9. Type of apps and app details

In Figure 10, I aggregate studies based on the type of treatment apps and the operating systems on which those apps run. 17 of 32 included studies used researcher-developed

apps as their intervention. Although five studies[4,5,10,15,24] do not specify the app's operating system, it is clear that researchers favour Android for app development, as seen in ten studies[1,2,3,7,9,18,19,21,31,32]. The preference for Android might result from the ease of developing and publishing apps on the platform as well as the prevalence of Android-equipped smartphones (O'Dea, 2020; Wu, 2015a). Commercially available apps, however, tend to be available on more operating systems: ten[8,13,14,23,25,26,27,28,29,30] out of 15 studies on commercially available apps used apps that run on both iOS and Android. Furthermore, Figure 11 shows the names of commercially available app names. The size of bubbles in Figure 10 represents the number of studies that used the app as its intervention. In total, 15 studies used nine commercially available apps as their intervention. Among these apps, Duolingo is the most popular as it is represented in five studies[8,13,14,20,28]. Two studies[17,22] used multiple apps as their interventions. Nicholes (2016)[17] used Rosetta Stone, Tell Me More and Memrise whereas Terantino (2016)[22] used Spanish Smash, LinguPinguin, Busuu Kids, Bilingual Child Bubbles, Bilingual Child. These apps are not included in the count.

Table 5 Summary of studies that compare apps or app versions (10)

Reference	Publication status	Age group	Final sample size	Location	Studied language	Type of study design	Duration intervention	App type and intervention details	Measured language skills
Studies that had a "weak" rating (7)									
1. Bhide (2017)	Doctoral thesis	Children (<12)	103[6]	India	Hindi	Non-randomised comparison	4 weeks	Researcher developed: An unnamed mobile game with two versions, narrow spacing and wide spacing **Control:** unseen, business-as-usual	Vocabulary, Orthography, Phonology
2. Çakmak & Erçetin (2018)	Peer-reviewed journal article	Adults (18+)	88	Turkey	English	RCT	1 day	Researcher developed: (1) Control group with no gloss access, (2) textual-only gloss group, (3) pictorial-only gloss group, (4) textual-plus-pictorial gloss group **Control:** none	Vocabulary, Listening
4. C.-M. Chen, Chen, et al. (2019)	Peer-reviewed journal article	Children (<12)	46	Taiwan	English	RCT	2 weeks	Researcher developed: EVLAPP-SRLM (English vocabulary learning app with a self-regulated learning mechanism): VOC 4 FUN **Control:** EVLASPP-NSRLM, no self-regulation	Vocabulary
5. C.-M. Chen, Liu, et al. (2019)	Peer-reviewed journal article	Adults (18+)	20	Taiwan	English	Non-randomised comparison	4 weeks	Researcher developed: MEVLA-GF: PHONE Words app (gamified) **Control:** MEVLA-NGF (non-gamified)	Vocabulary
6. de Jong et al. (2010)	Peer-reviewed journal article	Adults (18+)	35	NOT SPECIFIED	Hindi	RCT	1 day	Researcher developed: A mobile phrase book for learning Hindi with six conditions. **Control:** none	Vocabulary

[6] Details about final sample size by group were not reported in the article, but the author provided this information upon contact.

12. Liu (2016)	Master's thesis	Adults (18+)	86	USA	Chinese	Non-randomised comparison	1 day	Researcher developed: An unnamed app with two versions (A & B) **Control:** none	Orthography
17. Nicholes, J. (2016)	Peer-reviewed journal article	Adults (18+)	59	China	English	RCT	15 weeks	Commercially available: Rosetta Stone, Tell Me More, Memrise **Control:** none	Reading, Writing, Listening, Speaking, Lexicogrammar
Studies that had a "moderate" rating (3)									
10. Hsu et al. (2013)	Peer-reviewed journal article	Teenager (12-18)	108	Taiwan	English	Non-randomised comparison	4 weeks	Researcher developed: Experiment 1: Personalized recommendation-based mobile learning system in individual annotation mode. Experiment 2: Personalized recommendation-based mobile learning system in shared annotation mode **Control:** mobile learning system standard articles and individual annotation mechanism	Reading
15. Mohamad Ali et al. (2016)	Peer-reviewed journal article	Adults (18+)	60	NOT SPECIFIED	English	Non-randomised comparison	1 day	Researcher developed: Group 1: 3D talking-head with spoken text and on-screen text, Group 2: 3D talking-head with spoken text alone, **Control (group 3):** spoken text with on-screen text	Phonology
16. Moreno & Gatewood (2019)	Doctoral thesis	Adults (18+)	200	Online	Spanish	RCT	5 weeks	Researcher developed: A gamified mobile app prototype **Control:** a non-gamified mobile app prototype	Lexicogrammar (Spanish conjugation)

Table 6 Summary of studies that do not compare apps or app versions (22)

Reference	Publication status	Age group	Final sample size	Location	Studied language	Type of study design	Duration of intervention	App type and intervention details	Measured language skills
Studies that had a "weak" rating (11)									
3. Cavus & Ibrahim (2017)	Peer-reviewed journal article	Children (<12)	37	Cyprus	English	Non-randomised comparison	4 weeks	Researcher developed: Near East University Children's Story Teller (NEU-CST) **Control:** print story book	Vocabulary, Comprehension, Phonology, Listening
7. García-Mencía et al. (2016)	Conference proceeding	Adults (18+)	11	NOT SPECIFIED	Nahuatl	Cohort (one group, pre-test/post-test)	1 week	Researcher developed: The ALEN app. **Control:** none	Vocabulary, Listening
8. Guaqueta & Castro-Garces (2018)	Peer-reviewed journal article	Teenager (12-18)	20	Colombia	English	Cohort (one group, pre-test/post-test)	24 weeks	Commercially available: Duolingo **Control:** none	Vocabulary
9. Hao et al. (2019)	Peer-reviewed journal article	Teenager (12-18)	10	Taiwan	English	Other: specify in comments	1 day	Researcher developed: Detective ABC **Control:** none	Reading, Writing, Listening, Speaking
14. Martinelli (2016)	Master's thesis	Adults (18+)	5	NOT SPECIFIED	Italian	Cohort (one group, pre-test/post-test)	2 weeks	Commercially available: Duolingo **Control:** none	Phonology
18. Ou-Yang & Wu (2017)	Peer-reviewed journal article	Adults (18+)	108	Taiwan	English	Non-randomised comparison	2 weeks	Researcher developed: MyEVA Mobile **Control:** none	Vocabulary
19. Purgina et al. (2020)	Peer-reviewed journal article	Adults (18+)	21	Japan	English	RCT	Not specified	Researcher developed: WordBricks **Control:** traditional face-to-face with a print book	Lexicogrammar

22. Terantino (2016)	Peer-reviewed journal article	Children (<12)	7	USA	Spanish	Cohort (one group, pre-test/post-test)	24 weeks	Commercially Available: Spanish Smash, LinguPinguin, Busuu Kids, Bilingual Child Bubbles, Bilingual Child **Control:** none	Vocabulary, listening
23. Thongsri et al. (2019)	Peer-reviewed journal article	Adults (18+)	200	China	English	Non-randomised comparison	3 weeks	Commercially available: BW Vocabulary **Control:** none	Vocabulary
31. Wu (2015b)	Peer-reviewed journal article	Adults (18+)	199	China	English	RCT	12 weeks	Researcher developed: Word Learning-CET 4 Word **Control:** self-study print copy of the same materials	Vocabulary
32. Wu (2015a)	Peer-reviewed journal article	Adults (18+)	70	China	English	RCT	8 weeks	Researcher developed: Basic4Android **Control:** self-study print copy of the same materials	Vocabulary
Studies that had a "moderate" rating (7)									
13. Loewen et al. (2019)	Peer-reviewed journal article	Adults (18+)	9	USA	Turkish	Cohort (one group, pre-test/post-test)	12 weeks	Commercially available: Duolingo **Control: none**	Listening, Speaking, Writing, Reading, Lexicogrammar
21. Ramírez et al. (2018)	Conference proceeding	Children (<12)	96	Canada	Secwepemctsín	RCT	1 day	Researcher developed: Dress me Up app: Kindergarten to Grade 3, Matching UP: Grades 4 to Grade 7 **Control:** traditional face-to-face with the same materials	Vocabulary
24. Vesselinov & Grego (2015)	Organisation report	Adults (18+)	101	Online	Spanish	Cohort (one group, pre-test/post-test)	8 weeks	Commercially available: A new language app, not detailed **Control: none**	Lexicogrammar, Reading, Vocabulary

25. Vesselinov & Grego (2019b)	Organisation report	Adults (18+)	82	USA	Spanish	Cohort (one group, pre-test/post-test)	8 weeks	Commercially available: Pimsleur **Control: none**	Speaking
26. Vesselinov & Grego (2019a)	Organisation report	Adults (18+)	95	NOT SPECIFIED	Spanish	Cohort (one group, pre-test/post-test)	8 weeks	Commercially available: Mango languages **Control: none**	Vocabulary, Reading, Lexicogrammar, Speaking
28. Vesselinov & Grego (2012)	Organisation report	Adults (18+)	88	Online	Spanish	Cohort (one group, pre-test/post-test)	8 weeks	Commercially available: Duolingo **Control: none**	Vocabulary, Reading, Lexicogrammar
29. Vesselinov & Grego (2016b)	Organisation report	Adults (18+)	144	UK; USA	Spanish	Cohort (one group, pre-test/post-test)	8 weeks	Commercially available: Busuu **Control: none**	Vocabulary, Reading, Lexicogrammar, Speaking
Studies that had a "strong" rating (4)									
11. Liakin et al. (2015)	Peer-reviewed journal article	Adults (18+)	42	Canada	French	RCT	5 weeks	Commercially available: Nuance Dragon Dictation (ASR) **Control**: 1. face-to-face with the same materials (Non-ASR) 2. individual conversations, no focus (control)	Phonology
20. Rachels & Rockinson-Szapkiw (2018)	Peer-reviewed journal article	Children (<12)	467	USA	Spanish	RCT	12 weeks	Commercially available: Duolingo **Control:** traditional face-to-face with the same materials	Vocabulary, Writing (translation), lexicogrammar
27. Vesselinov & Grego (2016a)	Organisation report	Adults (18+)	325	Germany; USA; Online	Spanish	Cohort (one group, pre-test/post-test)	8 weeks	Commercially available: Babbel **Control: none**	Vocabulary, Reading, Lexicogrammar

30. Vesselinov et al. (2019)	Organisation report	Adults (18+)	143	Online	Spanish	Cohort (one group, pre-test/post-test)	8 weeks	Commercially available: Rosetta Stone **Control: none**	Lexicogrammar, Reading, Vocabulary

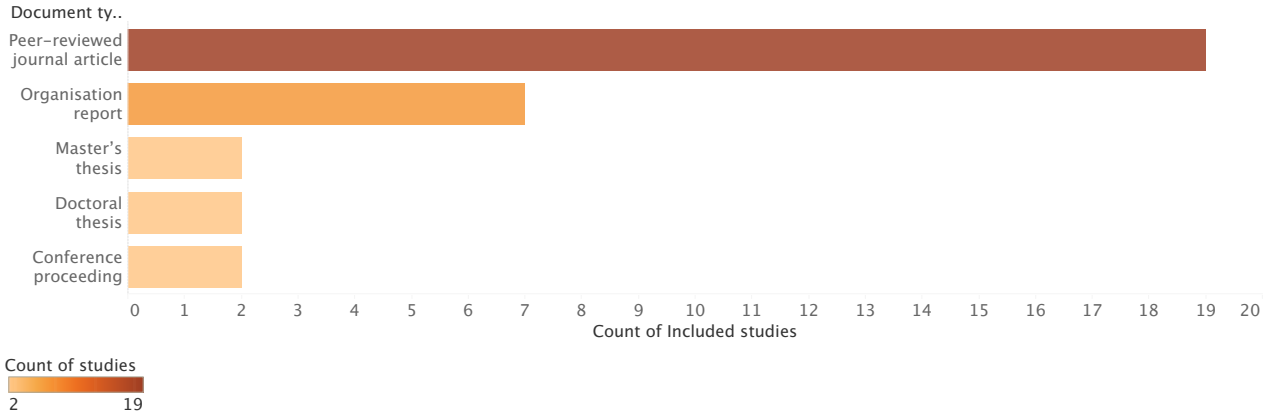

Figure 5 Publication status

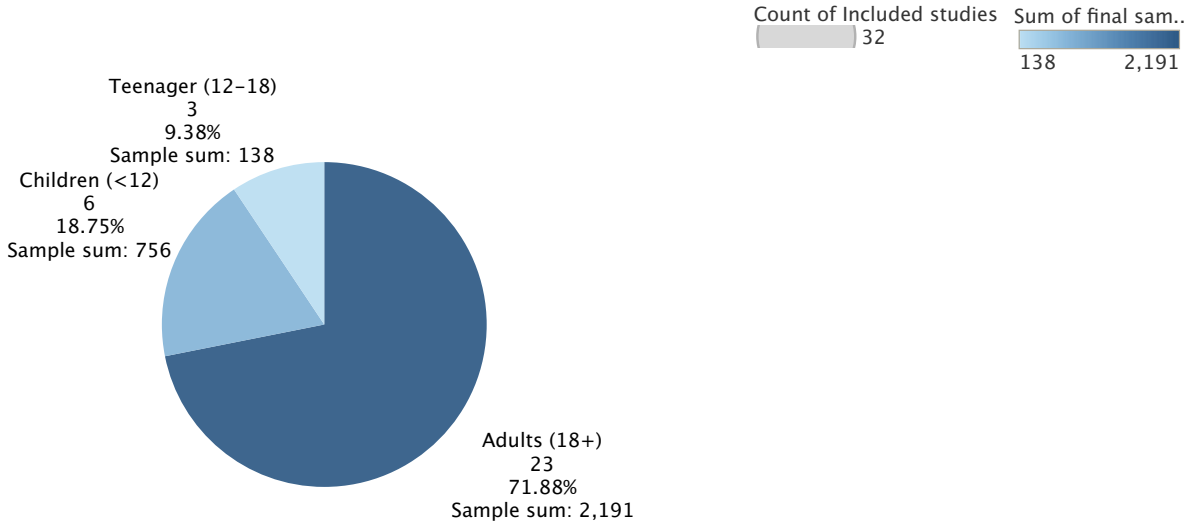

Figure 6 Age groups of participants

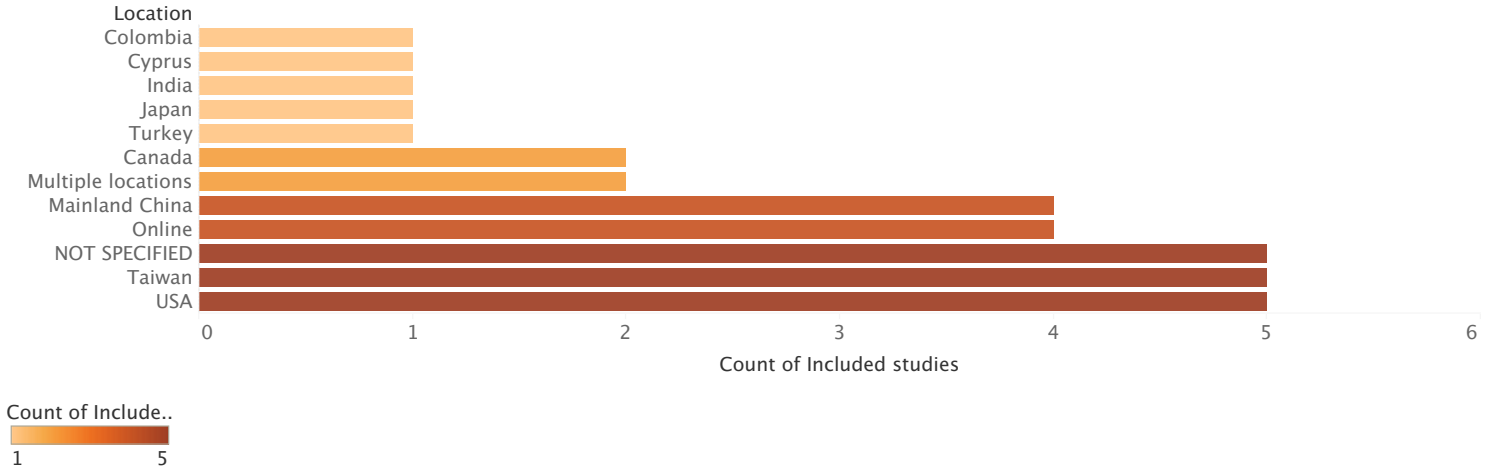

Figure 7 Location of study

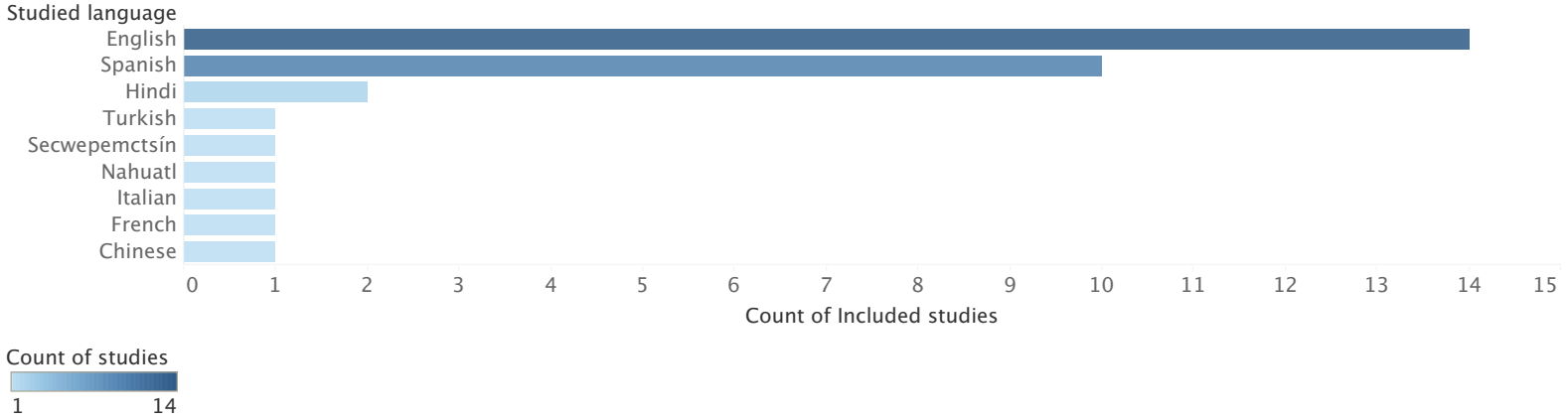

Figure 8 Studied language

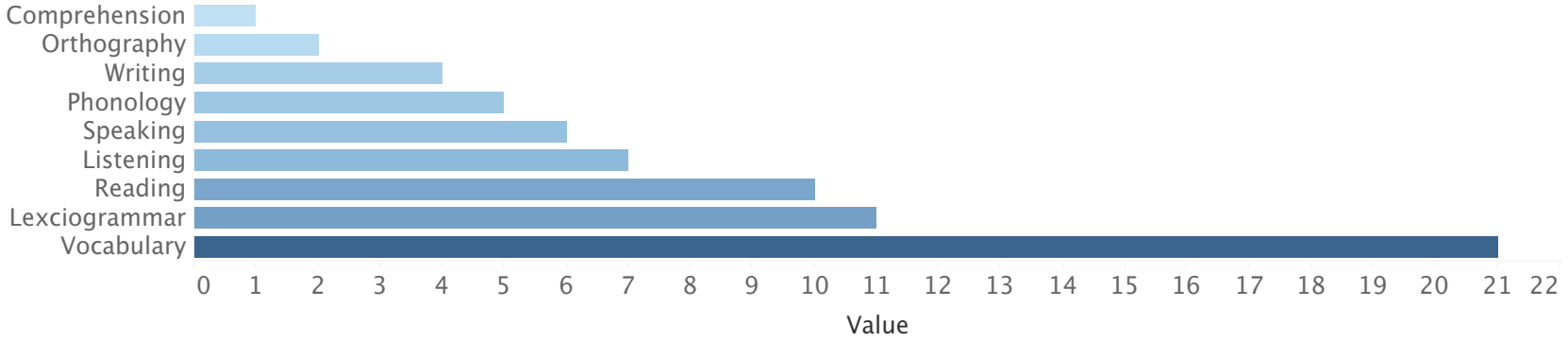

Figure 9 Measured language skills by the number of studies

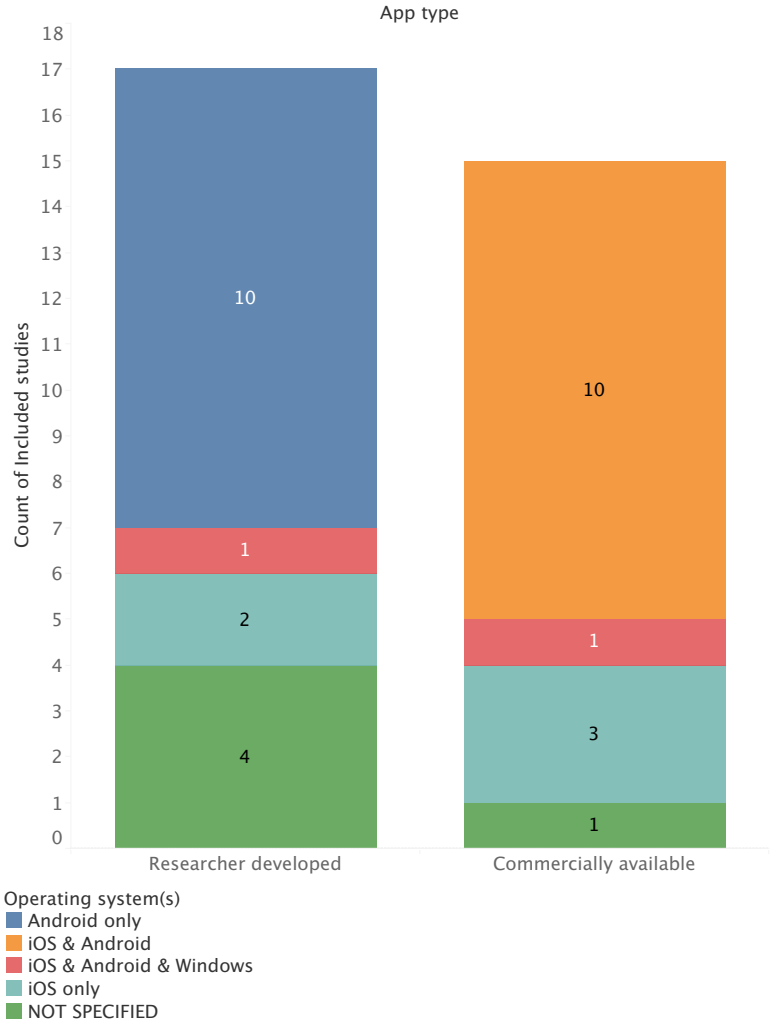

Figure 10 App type and operating system(s)

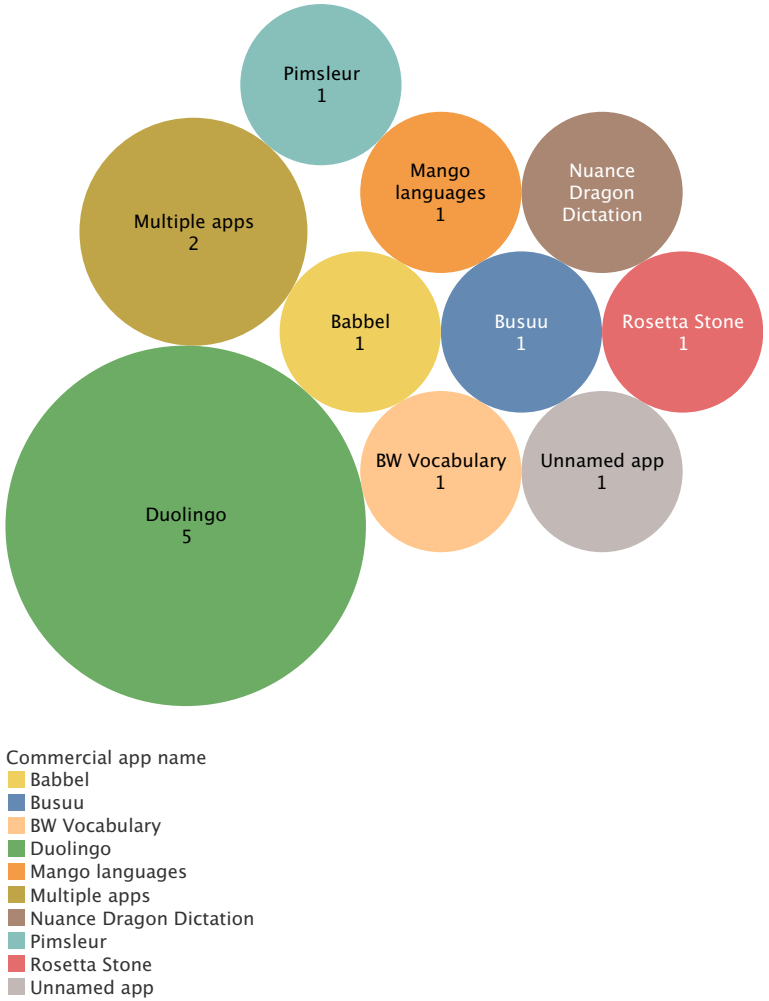

Figure 11 Commercial app names and count

4.2. Study findings

It is good practice to start the analysis of results in a narrative review with a summary of studies (Popay et al., 2006). Due to the limitation of space and the rather large number of included studies, I only provide a brief summary on the study design, results and conclusions of each study. Additionally, I present tables and graphs that illustrate the relationships between these factors. Table 7 and Table 8 contain the general direction of the effects, effect sizes and the raw test scores extracted from studies. I extracted effect sizes explicitly stated by the authors or where enough data were available for calculation. Figure 12 and Figure 13 show relationships between direction of effects, sample size, study design, age group of participants and risk of bias for the two categories of studies. The shape represents the age group of the participants and the colour represents the type of study design. Each geometric shape represents one study and the larger it is, the more participants the study recruited; the sample size is also noted in the graph.

The general direction of effects refers to the direction of conclusions by study authors. Studies are labelled "positive" (apps improve language proficiency) and "negative" (apps do not improve language proficiency or is not as effective as the comparator). The label "towards positive/negative" means that the study shows some effects towards the direction, but the evidence is not robust. Studies are labelled "mixed" if the authors conclude that there is not enough information to determine the effectiveness of tested apps or they have found differential effects in different tests. Some studies are labelled with "difference" if the author only found differences between treatment groups but do not provide conclusions on the effectiveness of apps; studies are labelled "no difference" if the results show no difference in treatments.

4.2.1. Studies that compare apps or app versions

Studied with a "weak" global RoB rating

1. Bhide (2017) studied the effects of a self-developed app for Hindi akshara with narrow-spacing and wide-spacing material distribution with a four-week non-randomised comparison design with 103 children. The experiment had two experimental groups and one business-as-usual control group. Results show that all participants improved from pre-test to post-tests, but the experimental groups particularly performed better on akshara recognition.

2. Çakmak & Erçetin (2018) studied the effects of a self-developed app for reading with three types of gloss in articles with a one-session RCT with 88 adults. Three

experimental groups participated, along with a control group with no gloss access to articles. Results show that gloss condition participants only had significantly higher scores in vocabulary tests. All groups showed no difference in text recall.

4. C.-M. Chen, Chen, et al. (2019) studied the effects of self-regulation functions in a self-developed app for English vocabulary learning in a two-week RCT with an experimental group (with self-regulation) and a control group (no self-regulation), totalling 46 children. Results show that both groups improved significantly from pre-test to post-test, but the experimental group showed more gains.

5. C.-M. Chen, Liu, et al. (2019) studied the effects of gamification in a self-developed app for English vocabulary learning in a four-week non-randomised comparison with an experimental group (gamified) and a control group (non-gamified), totalling 20 adults. The results show that both groups improved significantly from pre-test to post-test and delayed-post-test, but the experimental group showed more gains in both post-tests.

6. de Jong et al. (2010) studied the effects of selection method (semacode-, number-, list-, location-based) and filter (room/object) in a self-developed app for Hindi vocabulary learning in a one-session RCT with seven treatment groups of seven adults each. Results show that all participants improved from pre-test to post-test and both selection method and context filter affect knowledge gains.

12. Liu (2016) studied the effect of a self-developed app for Chinese orthography acquisition with version A (with metaphor and feedback) and version B (no metaphor nor feedback) in a one-session non-randomised comparison with an experimental group (A) and a control group (B), totalling 86 adults. Results showed that both groups gained Chinese orthographic knowledge after intervention but group A outperformed group B.

17. Nicholes, J. (2016) studied the effects of Rosetta Stone (RS), Tell Me More (TMM) and Memrise (MEM) on English language skills in a 15-week RCT with three treatment groups, each using one app and totalling 59 adults. The mixed showed that only MEM group improved on the total score from pre-test to post-test but RS and TMM groups improved on writing scores too. The RS group scored significantly lower on grammar post-test than on the pre-test.

Studied with a "moderate" global RoB rating

10. Hsu et al. (2013) studied the effects of a self-developed app for English reading with personalised recommendation and two annotation modes in a four-week intervention with 108 teenagers. Two experimental groups (individual annotation vs. shared annotation) were compared with a control group with no personalised recommendation. Results showed that all groups improved after intervention, but the experimental groups significantly outperformed the control group.

15. Mohamad Ali et al. (2016) studied the effects of a self-developed app for English pronunciation with a 3D talking-head and two multimedia modes (Group 1: spoken text + on-screen text vs. Group 2: spoken text only) in a one-session non-randomised comparison with 60 adults. Two treatment groups were compared against a control group with no 3D talking-head app (total n=20). Results show that while all groups showed improvement in pronunciation, group 1 outperformed both group 2 and the control group and the latter two showed no difference.

16. Moreno & Gatewood (2019) studied the effects of gamification in a self-developed app for Spanish conjugation learning in a five-week RCT with an experimental group and a control group (non-gamified), totalling 200 adults. Results show that while both groups gained knowledge on Spanish conjugation, there is no significant difference between their test scores.

4.2.2. Studied that do not compare apps

Studied that have a "weak" RoB rating

3. Cavus & Ibrahim (2017) studied the effects of a self-developed app for different English language skills in a four-week non-randomised comparison with an experimental group and a control group who used a print story book, totalling 37 children. Results show that both groups saw improvement from pre-test to post-test, but the experimental group had significantly higher gains on all four tests.

7. García-Mencía et al. (2016) studied the effects of a self-developed app for Nahuatl vocabulary and listening in a one-week intervention with a cohort (one group, pre-test/post-test) design, totalling 11 adults. Results show that most users improved their knowledge of Nahuatl and the differences in means were significantly different.

8. Guaqueta & Castro-Garces (2018) studied the effects of Duolingo on English vocabulary learning in a 24-week intervention with a cohort (one group, pre-test/post-

test) design, totalling 20 teenagers. Results show that all participants significantly improved from pre-test to post-test.

9. Hao et al. (2019) studied the effects of a self-developed app for English language skills in a one-session, post-test only intervention with a presumably low-achieving cohort of 10 teenagers. Results show that participants showed high achievement on vocabulary knowledge tests after the intervention.

14. Martinelli (2016) studied the effects of Duolingo for Italian pronunciation in a two-week intervention with a cohort (one group, pre-test/post-test) design with five adults. Mixed show that only one participant showed significant improvement in rated comprehensibility and accentedness.

18. Ou-Yang & Wu (2017) studied the effects of a self-developed app on English vocabulary learning in a two-week non-randomised comparison with students with different academic majors (English vs. non-English) as treatment groups, totalling 108 adults. Results show that regardless of major, participants improved significantly on their vocabulary knowledge.

19. Purgina et al. (2020) studied the effects of a self-developed app for English grammar in a RCT (unspecified length of treatment) with 21 adults. An experimental group who used the app was compared against a control group who had traditional face-to-face teaching with a print book covering the same materials. Results show that both groups saw improvement from pre-tests to post-test, but the experimental group had larger gains.

22. Terantino (2016) studied the effects of five commercial language learning apps on Spanish vocabulary and listening with a cohort of seven children (one group, pre-test/post-test) in a 24-week intervention. Results showed that participants improved significantly on both vocabulary and listening.

23. Thongsri et al. (2019) studied the effects of BW Vocabulary on English vocabulary learning in a three-week non-randomised comparison with students with different academic majors (STEM vs. non-STEM) as treatment groups, totalling 200 adults. Results show that the app had a significant positive effect on vocabulary learning for both groups, especially for STEM students.

31. Wu (2015b) studied the effect of a self-developed app for English vocabulary in a 12-week RCT with 199 adults. An experimental group was compared against a control group who received the same materials in print but no specified instructions.

Results show that the experimental group had a higher gain on the post-test scores than the control group although there is no indication of whether the difference is statistically significant.

32. Wu (2015a) studied the effect of a self-developed app for English vocabulary in an eight-week RCT with 70 adults. An experimental group was compared against a control group who received the same materials in print but no specified instructions. Data show that the experimental group had higher gains in the test scores than the control group, but whether the difference is statistically significant is unreported.

Studied that have a "moderate" RoB rating

13. Loewen et al. (2019) studied the effect of Duolingo on Turkish language acquisition with a cohort (one group, pre-test/post-test) design in 12 weeks with 9 adults. The results show that the participants knew more Turkish than they did before the treatment (beginner), but the test scores would not be satisfactory in a traditional college language course of the same length.

21. Ramírez et al. (2018) studied the effect of a self-developed app for Secwepemctsín in a one-day RCT with 96 children. Experimental groups of different grades were compared with respective control groups who received traditional face-to-face teaching covering the same material. Result show that only the control groups improved from pre-test to post-test and the gains are significant.

24. Vesselinov & Grego (2015) studied the effect of an unnamed language learning app on Spanish learning with a cohort (one group, pre-test/post-test) design in eight weeks with 101 adults. Results show that participants significantly improved their scores on the vocabulary, reading, and grammar post-tests.

25. Vesselinov & Grego (2019b) studied the effect of Pimsleur app on Spanish learning with a cohort (one group, pre-test/post-test) design in eight weeks with 82 adults. Results show that participants significantly improved their scores on speaking post-test.

26. Vesselinov & Grego (2019a) studied the effect of Mango languages app on Spanish learning with a cohort (one group, pre-test/post-test) design in eight weeks with 95 adults. Results show that participants significantly improved their scores on the vocabulary, reading, grammar, and speaking post-tests.

28. Vesselinov & Grego (2012) studied the effect of Duolingo on Spanish learning with a cohort (one group, pre-test/post-test) design in eight weeks with 88 adults.

Results show that participants significantly improved their scores on the vocabulary, reading, and grammar post-tests.

29. Vesselinov & Grego (2016b) studied the effect of Busuu on Spanish learning with a cohort (one group, pre-test/post-test) design in eight weeks with 144 adults. Results show that participants significantly improved their scores on the vocabulary, reading, grammar, and speaking post-tests.

Studied that have a "strong" RoB rating

11. Liakin et al. (2015) studied the effect of Nuance Dragon Dictation on French /y/ production and perception with a five-week RCT with 42 adults. The experimental group (ASR) using the app was compared with a non-ASR group who practiced the same materials with a teacher and a control group that only practiced in conversations without a teaching focus. Results show that the ASR group significantly improved on production but not on perception while the other groups showed no improvement on neither.

20. Rachels & Rockinson-Szapkiw (2018) studied the effects of Duolingo on Spanish vocabulary, writing, and grammar with a 12-week RCT with 467 children. The experimental group was compared against and a control group who used the same materials in a face-to-face classroom. Results show that while both groups improved from pre-test to post-test, there is no difference in their performance.

27. Vesselinov & Grego (2016a) studied the effect of Babbel on Spanish learning with a cohort (one group, pre-test/post-test) design in eight weeks with 325 adults. Results show that participants significantly improved their scores on the vocabulary, reading, and grammar post-tests.

30. Vesselinov et al. (2019) studied the effect of Rosetta Stone on Spanish learning with a cohort (one group, pre-test/post-test) design in eight weeks with 143 adults. Results show that participants significantly improved their scores on the vocabulary, reading, and grammar post-tests.

Table 7 Results: Studies that compare apps or app versions

Reference	Direction of effects	Outcome scores (experimental) Mean (SD) unless otherwise noted	Outcome scores (control) Mean (SD)	Effect sizes
Studies that had a "weak" rating (8)				
1. Bhide (2017)	Positive	Akshara recognition (in %, in the order of simple, CV, Complex) Narrow Pre-test 94.44 (9.78) 46.88 (21.77) 36.25 28.93) Post-test 98.96 (3.29) 57.29 (20.71) 53.13 (29.89) Wide Pre-test 94.29 (15.33) 55.71 (22.49) 34.29 (27.26) Post-test 98.69 (3.63) 59.8 (21.37) 51.18 (27.05) Word reading (out of 8, in the order of simple, CV, learned, near, medium, far) Narrow Pre-test 5.58 (1.46) 4.18 (2.08) 3.42 (1.49) 2.23 (1.61) 3.85 (2.18) 2.81 (1.77) Post-test 6.16 (1.49) 5.48 (1.71) 4.85 (1.52) 3.15 (1.84) 4.10 (2.50) 3.65 (1.86) Wide Pre-test 5.51 (1.63) 4.09 (1.94) 3.46 (1.59) 2.43 (1.88) 3.70 (2.23) 3.17 (1.83) Post-test 6.17 (1.52) 4.99 (1.99) 5.26 (1.15) 3.49 (1.86) 4.63 (1.79) 3.93 (1.83)	Akshara recognition Pre-test 95.37 (9.71) 46.3 (21.50) 31.11 (28.86) Post-test 95.68 (9.31) 50.93 (24.86) 35.56 (26.67) Word reading Pre-test 5.06 (1.74) 3.89 (1.96) 3.14 (1.53) 2.11 (1.64) 3.08(2.46) 2.93 (2.00) Post-test 6.35 (1.66) 4.96 (2.04) 4.18 (2.01) 2.85 (1.74) 4.07 (2.13) 3.17 (1.86)	*Not stated by author and not enough information to calculate.*

Spelling test (out of 6, in the order of simple, CV, learned, near medium, far)		Spelling test (out of 6)
Narrow	Wide	Pre-test
Pre-test	Pre-test	3.25 (1.34)
3.59 (1.48)	3.34 (1.37)	1.13 (1.54)
1.34 (1.46)	1.36 (1.44)	0.63 (0.81)
0.61 (0.92)	0.77 (0.92)	0.50 (0.73)
0.44 (0.82)	0.36 (0.71)	0.33 (0.76)
0.33 (0.69)	0.37 (0.81)	0.06 (0.20)
0 (0)	0 (0)	Post-test
Post-test	Post-test	4.19 (0.98)
4.03 (1.36)	4.11 (0.83)	1.94 (1.40)
2.14 (1.52)	2.24 (1.26)	0.94 (0.99)
1.11 (1.01)	1.60 (1.37)	0.57 (0.90)
0.70 (0.85)	0.71 (1.09)	0.65 (1.02)
1.03 (1.26)	0.81 (1.14)	0.11 (0.34)
0.17 (0.49)	0.20 (0.62)	Akshara construction (out of 7)
Akshara construction (out of 7)		
Narrow	Wide	Pre-test
Pre-test	Pre-test	4.89 (2.03)
4.97 (1.73)	5.06 (1.82)	Post-test
Post-test	Post-test	6.09 (1.22)
6.06 (1.16)	6.09 (1.26)	
Vocabulary (out of 24, in order of “unrelated to game,” “morphologically related to words in game,” and “learned in game”)		Vocabulary
Narrow	Wide	Pre-test
Pre-test	Pre-test	15.06 (5.66)
14.00 (5.83)	14.40 (6.36)	7.78 (5.00)
7.75 (5.41)	8.83 (6.06)	6.36 (4.98)
5.03 (4.59)	6.03 (5.81)	Post-test
Post-test	Post-test	16.61 (5.93)
15.34 (5.71)	16.57 (5.39)	9.61 (5.84)
8.75 (6.54)	9.54 (6.38)	7.47 (4.63)
6.50 (5.56)	7.66 (6.01)	

2. Çakmak & Erçetin (2018)	No difference	**Form recognition** Textual-only gloss 18.36 (4.04) **L1 meaning production** Textual-only gloss 17.09 (3.73) **L2 meaning production** Textual-only gloss 13.86 (4.22)	Pictorial-only 17.59 (4.8) Pictorial-only 15 (3.37) Pictorial-only 11.5 (3.5)	Textual-plus-pictorial gloss 16.36 (5.57) Textual-plus-pictorial gloss 17 (5.53) Textual-plus-pictorial gloss 13.41 (4.84)	No gloss **Form recognition** 8.82 (3.35) **L1 meaning production** 12.77 (4.09) **L2 meaning production** 10.18 (4.03)	**Multivariate** $\eta^2 = 0.195$. **Univariate** form recognition: $\eta^2 = 0.43$; L1 meaning production: partial $\eta^2 = 0.15$; L2 meaning production: partial $\eta^2 = 0.12$.
4. C.-M. Chen, Chen, et al. (2019)	Positive	**Vocabulary ability assessment** Experimental group Pre-test 40 (29.07) Post-test 68.19 (22.82)*			Control group Pre-test 46.08 (30.44) Post-test 58.88 (27.36)*	d=0.37
5. C.-M. Chen, Liu, et al. (2019)	Positive	**Rank-based vocabulary test adopted from TOEIC** Experimental Pre-test 14.7 Post-test 48.9* Delayed post-test 41.75*			Control Pre-test 14.7 Post-test 39.3* Delayed post-test 23.65*	d=0.86
6. de Jong et al. (2010)	Difference	**Knowledge gain on a vocabulary test.**				Treatment: r=0.79, Context filter: r=0.42, Selection method: r=0.69
12. Liu (2016)	Positive	**A quiz on Chinese characters to translate between English and Chinese.** Post-test only because zero proficiency confirmed with pre-test Group A:20.26 (1.498)*; Group B: 17.98 (3.398)*				d=0.87

	Semacode-based	Number-based	List-based	Location-based
Room filter	0.35 (0.24)	0.38 (0.20)	0.47 (0.04)	0.62 (0.13)
Object filter	0.22 (0.12)	0.37 (0.14)	0.35 (0.18)	

17. Nicholes, J. (2016)	Mixed	**A version of a U.S. university's English placement test** **Rosetta Stone** **Tell me more** **Memrise** **Reading** Pre-test 45.9(16.25) 49.11(13.21) 39.86(14.58) Post-test 44.84(18.45) 55.16(12.56) 47.42(21.12) **Writing** Pre-test 54.41(11.04)* 63.63(10.94)* 64.86(6.89)* Post-test 63.26(11.42)* 68.21(11.99)* 72.43(6.45)* **Listening** Pre-test 54.47(12.68) 63.42(13.44) 53.62(11.21) Post-test 49.74(17.91) 65.79(11.69) 54.45(10.8) **Speaking** Pre-test 50(12.25) 82.9(12.62) 64.29(8.26)* Post-test 51.58(12.48) 81.58(14.34) 68.81(9.61)* **Grammar** Pre-test 60.79(13.81)~ 70(12.07) 65.95(12.34) Post-test 44.84(18.45)~ 69.79(11.7) 68.38(10.51) **Total** Pre-test 53 (8.56) 65.84 (7.65) 57.714 (5.41)* Post-test 54.62 (11.40) 68.00(8.90) 61.57 (5.25)* **Significantly better score (paired t-test and Wilcoxon's test)* *~Significantly worse score*	N/A	*Not stated by author.* d=0.72 (Memrise total only)

22. Terantino (2016)	Positive	**A vocabulary recall test containing animal vocabulary** Pre-test: 1.71 (3.40) Post-test: 12.86 (3.34) **Listening comprehension test on animal vocabulary** Pre-test: 1.71 (3.40) Post-test: 12.86 (2.79)	N/A	Vocabulary: r^2=0.96, Listening r^2=0.97
Studies that had a "moderate" rating (3)				
10. Hsu et al. (2013)	Positive	**A reading comprehension test** **Experimental 1** Pre-test 39.76 (9.43)* Post-test 71.21 (16.68)* **Experimental 2** 40.79 (11.83)* 68.94 (15.35)*	**Control** Pre-test 39.38 (8.84)* Post-test 57.74 (23.09)*	d=0.47
15. Mohamad Ali et al. (2016)	Positive	**An oral test that only measure the pronunciation performance of the selected words** Pre-test (covariate): 22.91 Post, adjusted Mean (SE) 3D talking-head with audio and text MALL 40.70 (0.96)* 3D talking-head with audio alone MALL 36.73 (0.97)*	Post, adjusted Mean (SE) Audio and text alone MALL 36.77 (0.96)*	partial η^2=0.17
16. Moreno & Gatewood (2019)	No difference	**Five online tests on Spanish conjugation** 1: 93.06 (6.28) 2: 93.44 (4.68) 3: 94.10 (5.47) 4: 93.69 (5.41) 5: 93.07 (4.97)	Control 1: 93.12 (5.75) 2: 93.45 (5.02) 3: 93.90 (5.27) 4: 94.01 (5.37) 5: 93.75 (5.17)	N/A (no difference)

Table 8 Results: Studies that do not compare apps or app versions

Reference	Direction of effects	Outcome scores (experimental) Mean (SD) unless otherwise noted			Outcome scores (control) Mean (SD)			Effect sizes (if applicable)
Studies that had a "weak" rating (10)								
3. Cavus & Ibrahim (2017)	Positive	**A on vocabulary, comprehension, pronunciation and listening with no information given**						*Not stated by author.* Vocabulary: d=2.66 Comprehension: d=2.79 Pronunciation: d=2.28 Listening: d=0.71
		Experimental			**Control**			
			Pre-test	Post-test		Pre-test	Post-test	
		Vocabulary	22.42 (10.90)*	76.53 (7.18)*	**Vocab**	23.50 (11.11)*	51.28 (11.35)*	
		Comprehension	27.42 (9.54)*	78.84 (7.58)*	**Comp.**	21.78 (9.57)*	56.22 (8.60)*	
		Pronunciation	21.16 (11.53)*	79.74 (8.03)*	**Pron.**	23.83 (10.19)*	55.67 (12.56)*	
		Listening	26.16 (10.86)*	78.42 (7.07)*	**List.**	24.44 (8.91)*	71.61 (11.68)*	
7. García-Mencía et al. (2016)	Positive	**A researcher-designed online examination on vocabulary and listening**			N/A			*Not stated by author.* d=1.11
		Pre-test		Post-test				
		Volunteer 1: 2		Volunteer 1: 4				
		Volunteer 2: 0		Volunteer 2: 10				
		Volunteer 3: 5		Volunteer 3: 5				
		Volunteer 4: 1		Volunteer 4: 8				
		Volunteer 5: 6		Volunteer 5: 9				
		Volunteer 6: 5		Volunteer 6: 7				
		Volunteer 7: 0		Volunteer 7: 6				
		Volunteer 8: 2		Volunteer 8: 4				
		Volunteer 9: 0		Volunteer 9: 2				
		Volunteer 10: 3		Volunteer 10: 2				
		Volunteer 11: 5		Volunteer 11: 6				
		Mean: 2.64 (2.29)*		Mean: 5.73 (2.65)*				
8. Guaqueta & Castro-Garces (2018)	Positive	**A vocabulary test: identify vocabulary in context and the full mark is 100.**			N/A			d=2.67
		Pre-test		Post-test				

		F1: 32 F2: 36 F3: 58 F4: 45 F5: 46 F6: 18 F7: 22 F8: 24 F9: 42 F10: 29 F11: 34 F12: 28 F13: 65 M1: 53 M2: 31 M3: 16 M4: 47 M5: 26 M6: 48 M7: 34 Mean: 36.7 (13.4)*	F1: 62 F2: 80 F3: 89 F4: 67 F5: 78 F6: 65 F7: 72 F8: 59 F9: 72 F10: 82 F11: 76 F12: 80 F13: 92 M1: 71 M2: 65 M3: 32 M4: 84 M5: 63 M6: 86 M7: 94 Mean: 73.45 (14.19)*		
9. Hao et al. (2019)	Positive	**Two four-part unit quizzes. Each participant had two units they were most unfamiliar with but equivalent in difficulty level. Sub-scores on reading, listening, speaking, and writing were reported but only total scores on each unit are extracted here.** **Unit 1 (out of 14)** S1: 14 S2: 14 S3: 12 S4: 14 S5: 14 S6: 14 S7: 13	**Unit 2 (out of 14)** S1: 13 S2: 14 S3: 14 S4: 12 S5: 14 S6: 14 S7: 14	N/A	Not enough information to calculate (no pre-test or control group)

		S8: 12 S9: 12 S10: 13 Mean: 12.8 (1.23) Total Mean: 13 (1.08)	S8: 13 S9: 12 S10: 12 Mean: 13.2 (0.92)	
14. Martinelli (2016)	Towards negative	**Identical pre-test and post-test which tested the participants' pronunciation in single words, phrases, and passages.** **Comprehensibility: 1-9, 1 is extremely easy to understand** Pre-test Emma: 3.5 (1.91) Barbara: 4.27 (2.58) Luciana: 3.77 (2.18)* Sabrina: 4.22 (2.83) Carolina: 4.9 (2.56) Post-test Emma: 2.72 (1.70) Barbara: 4.33 (1.840) Luciana: 1.61 (0.970)* Sabrina: 2.83 (2.430) Carolina: 3.5 (2.540) **Accentedness: 1-9, 1 is no foreign accent** Pre-test Emma: 5.88 (2.08) Barbara: 7.11 (1.87) Luciana: 6.61 (1.88)* Sabrina: 6.16 (1.85) Carolina: 7.1 (2.06) Post-test Emma: 4.55 (1.65) Barbara: 7.16 (1.94) Luciana: 4.37 (2.32)* Sabrina: 5.77 (2.23) Carolina: 6.7 (1.76)	N/A	Luciana only. Comprehensibility: η^2=0.54, Accentedness: η^2=0.47
18. Ou-Yang & Wu (2017)	Positive	**An English proficiency pre-test/post-test comprised of 50 items and the total score was 100**. **Non-English major** Pre-test 27.70 (14.60) Post-test 48.38 (19.71)* **English major** Pre-test 59.20 (15.30) Post-test 75.42 (15.72)*	N/A	*Not stated by author.* Pre/post: Non-English major d=1.19, English major d=1.05. Can't calculate between groups because assumptions are not met.
19. Purgina et al. (2020)	Positive	**Two sets of researcher-developed paper-based tests to measure participants' English grammar performance over two course units (69, 70).**		*Not stated by author.* Between groups: Unit 69: d=0.64,

		G2 (experimental group) **Unit 69** **Unit 70** Pre-test Pre-test 15.90 (4.43) 4.20 (2.57) Post-test Post-test 24.20 (4.02)* 11.60 (2.84)8 *The article reports a third group in a separate study which is not included in this review because insufficient information is given.*	G1 (control group) **Unit 69** **Unit 70** Pre-test Pre-test 15.18 (5.04) 6 (2.72) Post-test Post-test 21 (5.80) 9.18 (4.17)	Unit 70: d=0.68
23. Thongsri et al. (2019)	Positive	**An English vocabulary test containing 100-item multiple-choice tasks with four choices for each item.** **Pre-test** STEM non-STEM 44.81 (8.33) 45.27 (8.12) **Post-test** STEM non-STEM 71.62 (12.01)* 63.89 (14.28)*	N/A	*Not stated by author.* Pre/post: STEM: d=0.50, non-STEM: d=0.52
31. Wu (2015b)	Positive	**An English vocabulary test with 100 words selected from the app.** **Experimental** Pre-test Post-test 50.84 (13.11)* 60.61 (11.96)*	**Control** Pre-test Post-test 49.80 (12.68) 51.08 (12.08)	*Not stated by author.* d=0.41
32. Wu (2015a)	Positive	**An English vocabulary test with 100 words selected from the app.** **Experimental** Pre-test Post-test 35.13 50.52 (10.41)*	**Control** Pre-test Post-test 35.68 43.56 (11.40)	*Not stated by author.* d=0.55
Studies that had a "moderate" rating (7)				
13. Loewen et al. (2019)	Mixed	**Duolingo progress test** 0.63 (0.48) **A Turkish 151 Test derived from a university-level, end-of- course summative achievement test for first semester L2 Turkish learners (70% is considered pass).** **Total** 48 (19) **Listening** 37 (22) **Speaking** 33 (14)	N/A	*Not stated by author. Not enough information to calculate.*

		Writing 57 (27) **Reading** 55 (36) **Lexicogrammar** 50 (16)		
21. Ramírez et al. (2018)	Negative	A paper and pencil, multiple-choice vocabulary test containing 12 items for the primary grades and 10 for the intermediate grades. **Experimental (primary)** Pre-test 81% Post-test 81% **Experimental (intermediate)** Pre-test 71% Post-test 72%	**Control (primary)** Pre-test 80%* Post-test 85%* **Control (intermediate)** Pre-test 87%* Post-test 94%*	At post test partial η^2= 0.093
24. Vesselinov & Grego (2015)	Positive	**WebCAPE (Vocabulary/Reading/Grammar)** Pre-test 243.4 (117.6), [0, 418]. 95% CI: 220.2-266.6 Post-test 314.4 (113.8), [0, 543], 95% CI: 291.9-336.9	N/A	*Not stated by author. Not enough information to calculate.*
25. Vesselinov & Grego (2019b)	Positive	**An oral proficiency test developed by an independent testing company** Pre-test: Level placement / N / %; Post-test: N / % Novice Low: 78 / 95.10% / 33 / 41.30% Novice Mid: 2 / 2.40% / 38 / 47.50% Novice High: 2 / 2.40% / 6 / 7.50% Intermediate Low: 2 / 2.50% Intermediate High: 1 / 1.30%	N/A	*Not stated by author. Not enough information to calculate.*
26. Vesselinov & Grego (2019a)	Positive	**WebCAPE** Pre-test 97.2 (106.9), 95% CI: 74.4-120.0 Post-test 229.4 (106.9), 95% CI: 204-255 **TrueNorth Test (oral proficiency, 0-10)** Pre-test 2.4 (1.4), 95% CI: 2.1-2.7 Post-test		*Not stated by author. Not enough information to calculate.*

		3.4 (1.1), 95% CI: 3.2-3.7		
28. Vesselinov & Grego (2012)	Positive	**WebCAPE** Pre-test 158.0 (115.4) [0, 405] Post-test 253.9 (110.5) [0, 539]	N/A	*Not stated by author. Not enough information to calculate.*
29. Vesselinov & Grego (2016b)	Positive	**WebCAPE** Pre-test 24.8 (122.4), 95% CI: 104.6-145.0 Post-test 269.6 (151.4), 95% CI: 244.6-294.5 **Oral Proficiency Interview by Computer® (OPIc)** Pre-test: Level placement N / %; Un-ratable 6 / 9.80%; Novice Low 30 / 49.20%; Novice Mid 19 / 31.10%; Novice High 2 / 3.30%; Intermediate Low 3 / 4.90%; Intermediate High 1 / 1.60% Post-test: N / %; Un-ratable 2 / 3.30%; Novice Low 9 / 14.80%; Novice Mid 17 / 27.90%; Novice High 21 / 34.40%; Intermediate Low 7 / 11.50%; Intermediate High 3 / 4.90%; Intermediate Mid 1 / 1.60%; Advanced Low 1 / 1.60%	N/A	*Not stated by author. Not enough information to calculate.*
Studies that had a "strong" rating (4)				
11. Liakin et al. (2015)	Towards positive	**A pronunciation test on CAN-8 VirtuaLab** **Production (item n=20)** Pre-test / Post-test ASR 7.09 (5.51)* / 10.71 (4.23)* Non-ASR 9.79 (3.98) / 11.86 (3.98) Control 8.43 (5.86) / 9.5 (5.53) **Perception (item n=15)** Pre-test / Post-test ASR 8.07 (3.64) / 9.93 (2.89) Non-ASR 9.64 (3.67) / 10.29 (3.77) Control 10.29 (2.81) / 10.93 (3.4)		*Not stated by author. Not enough information to calculate.*

20. Rachels & Rockinson-Szapkiw (2018)	No difference	**A researcher developed test based on materials, comprising of questions on vocabulary, writing (translation), and grammar. The full mark is 50.** Pre-test 11.78 (9.94) Post-test 11.78 (8.88)	**Control group** Pre-test 20.94 (9.93) Post-test 21.47 (10.20)	N/A. No difference.
27. Vesselinov & Grego (2016a)	Positive	**WebCAPE** Pre-test 106.4 (108.6), 95% CI: 96.8-120.4 Post-test 265.1 (124.2) ,95% CI: 251.5-278.7	N/A	*Not stated by author. Not enough information to calculate.*
30. Vesselinov et al. (2019)	Positive	**WebCAPE** Pre-test 42.0 (68.5), 95% CI: 31.2-53.5 Post-test 145.2 (110.0) ,95% CI: 127.4-163.7	N/A	*Not stated by author. Not enough information to calculate.*

*Note. * means that statistical significance is found between scores.*

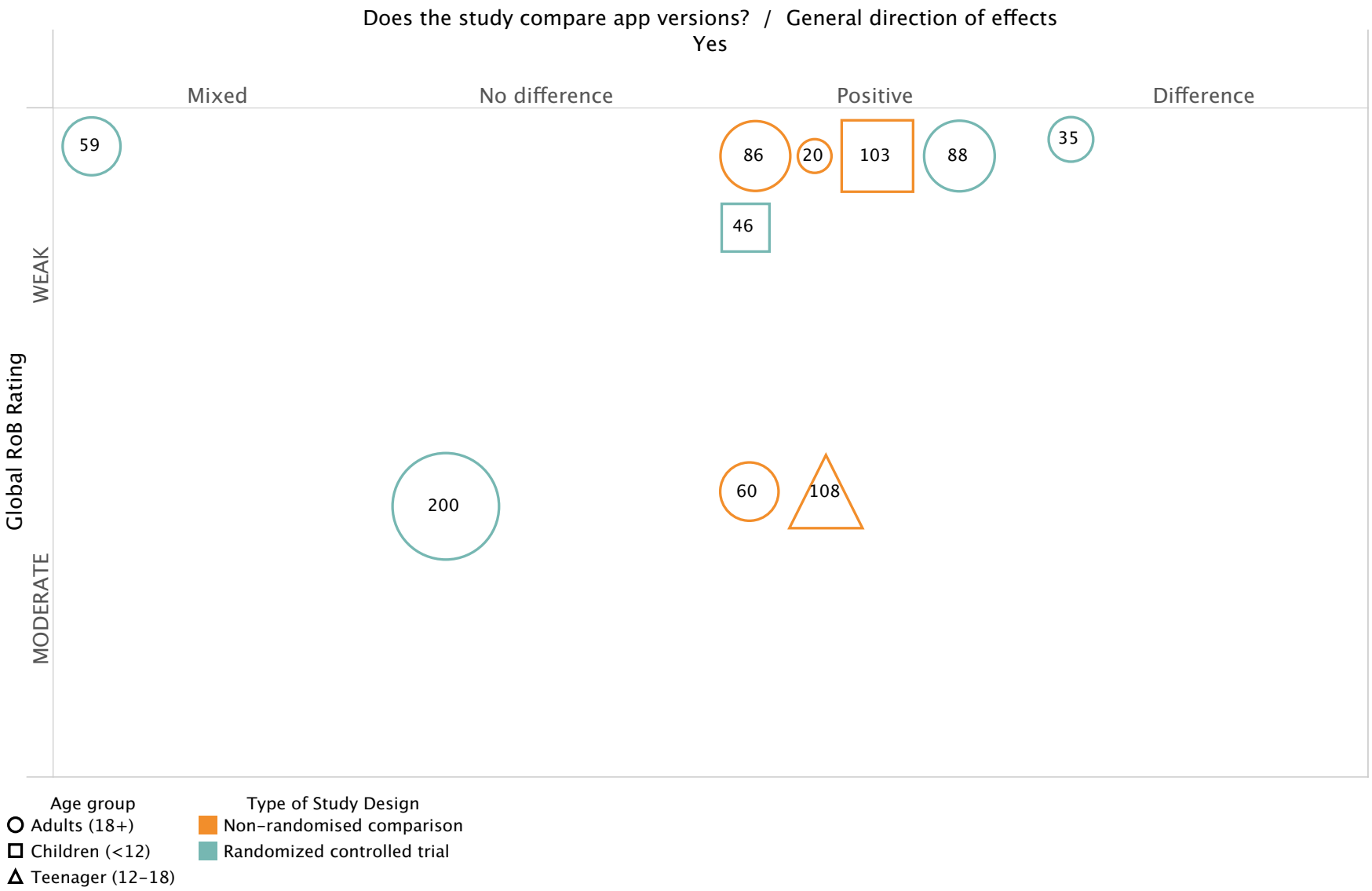
Does the study compare app versions? / General direction of effects
Yes
Mixed
No difference
Positive
Difference
Global RoB Rating
WEAK
MODERATE
59
200
86
20
103
88
35
46
60
108
Age group
Adults (18+)
Children (<12)
Teenager (12-18)
Type of Study Design
Non-randomised comparison
Randomized controlled trial

Figure 12 Studies that compare app/app versions: direction of effects, study design, age groups, sample size, and RoB

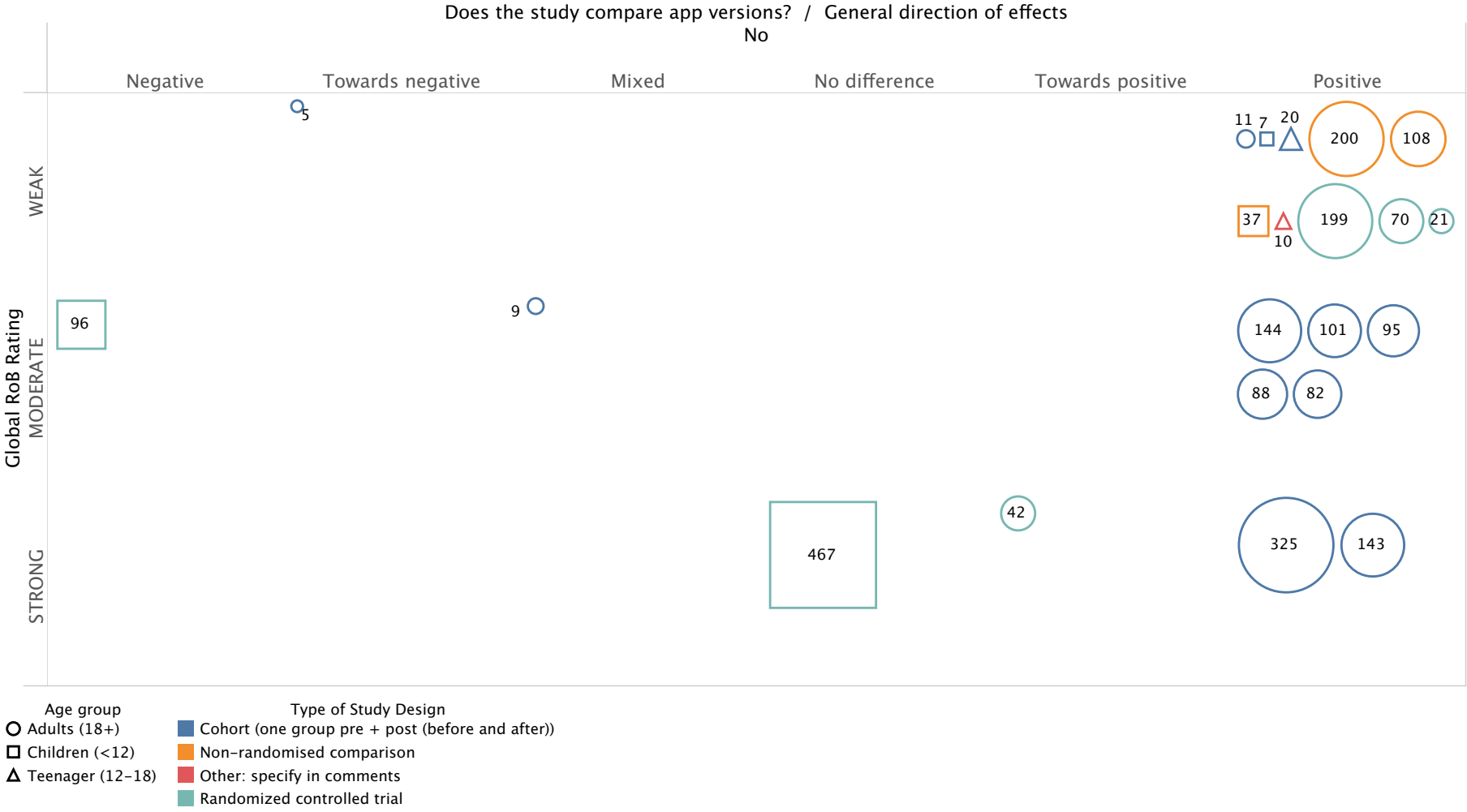

Figure 13 Studies that do not app/app versions: direction of effects, study design, age groups, sample size, and RoB

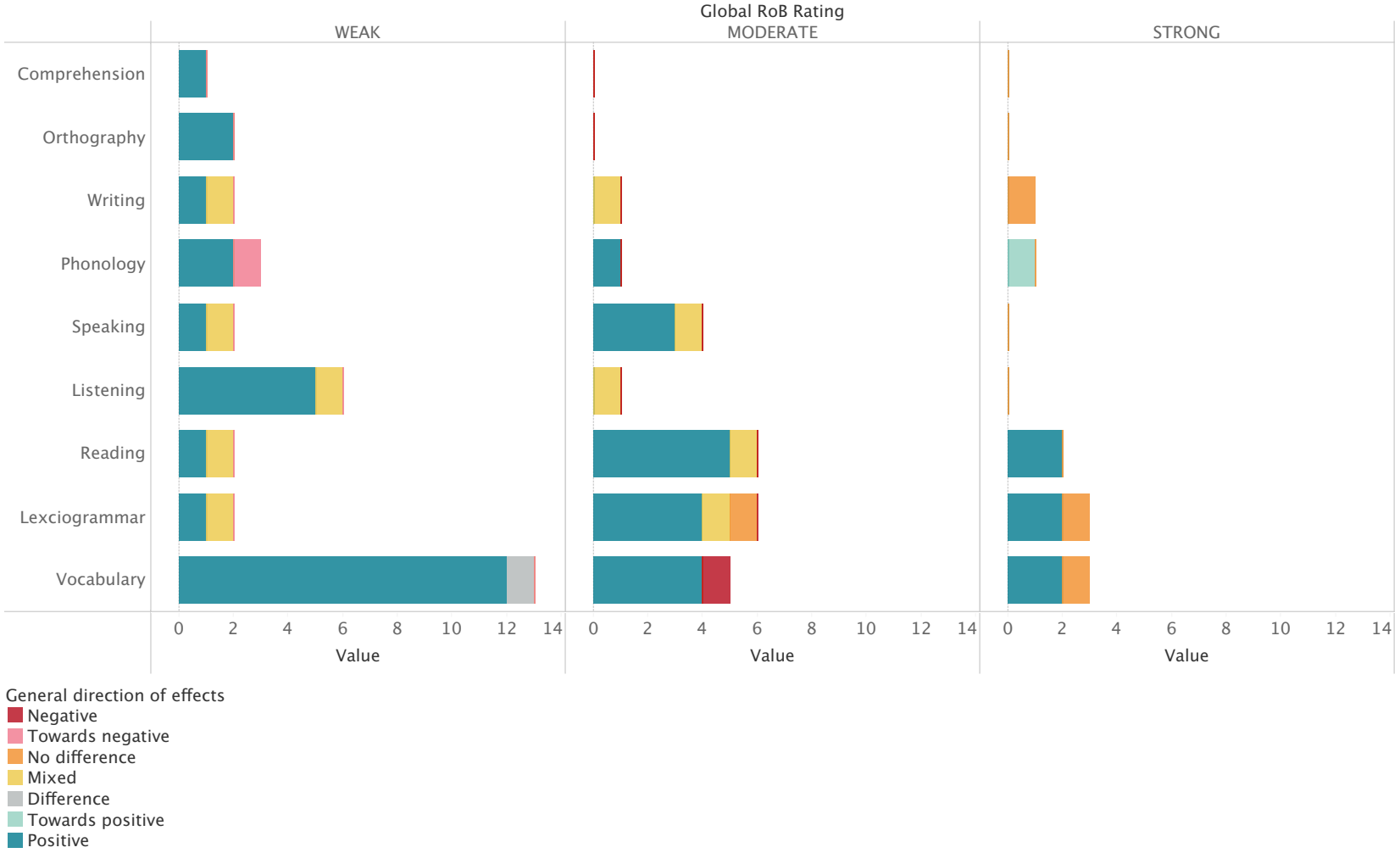

Figure 14 Direction of effects, global risk of biases, and measured language skills

4.3. Cumulative confidence across studies

Table 9 summarises the risk of bias ratings across studies (see Appendix E for a full summary). It reveals that while most studies are strong in controlling for confounders and attrition, there are significant risks in selection bias, the blinding process, and data collection practices. Most studies' designs are rated to be either "moderate" or "strong." That 56.3% of included studies are rated "weak" and 31.3% as "moderate" suggests that the overall security of findings in included studies is low.

Table 9 Summary of risk of biases

Ref. No.	Selection bias	Study design	Confounders	Blinding	Data collection	Withdrawal /dropout	Global Rating
Number of "weak"	**13**	1	6	**15**	12	3	18
Number of "moderate"	7	12	4	4	2	4	10
Number of "strong"	12	**19**	**22**	1	**18**	**25**	4

CHAPTER 5. Discussion

5.1. Synthesis and summary of findings

Aligning with the best-evidence approach (Slavin, 1986), I explore relationships between study outcomes and important study characteristics with a textual synthesis and supporting tables and figures.

5.1.1. Studies that compare apps/app versions

As seen in the last section, Tables 5 and 7, in studies that compare apps or app versions, all but one study[17] investigate the effects of specific functions of researcher-developed apps on language learning. Such function as gamification[5,16], more access to multimedia resources[2,12,15], self-regulation[4], and personalisation[10] are subjects of interest. The findings suggest that these functions are conducive to the improvement of language proficiency. This reflects inferences from the existing literature (see 2.1.2).

Seven of 10 studies found unequivocally positive effects of the treatment. Study design seems to associate with the direction of effects: all non-randomised comparisons found positive effects for the treatment, meaning that surveyed functions are more effective in improving language proficiency than their control treatments. On the other hand, the RCTs reached conclusions of a variety of directions. While the results reveal that apps, regardless of their functionality, generally helps learners gain language proficiency (see 4.2.1 and Table 7), one RCT (Moreno & Gatewood, 2019) found no difference in vocabulary gains of participants using gamified and non-gamified apps. In the original study, the authors investigated gamification's impact on motivation, usability, and satisfaction in addition to language knowledge. They found significant and positive effects of gamification on all factors but knowledge gain. This finding contradicts to the positive effects of gamification in a "weak" non-randomised comparison (C.-M. Chen, Liu, et al., 2019). These mixed findings for the effects of gamification corroborate those in Landers and Landers (2014) (see **Gamification** for the detailed discussion). Nevertheless, since Moreno and Gatewood (2019) has both lower risk of bias and a larger sample size, its results should be considered more trustworthy.

In addition, an RCT (Nicholes, 2016) compared the effectiveness of different apps and only found positive effects of one (Memrise) on overall language proficiency, whereas Rosetta Stone disadvantaged learners on English writing. This suggests that language learning apps with different pedagogical foci could have differential effects on different

learning outcomes. However, since this is the only study that compare different commercial apps, further research is needed for a robust conclusion.

In terms of sample size and demographics, Figure 12 reveals that the sample size of studies in this category ranges from 20 to 200. The majority of studies (70%) recruited adults as participants. Since the representation of age groups is dissimilar, it is inappropriate to make any inference about the relationship between the direction of effects and age of participants. At last, it is noticeable that most studies (70%) in this category are considered "weak" studies because of their study design and unsatisfactory reporting (see Table 9). Hence, more rigorous research is needed to develop a full picture of the impact of app functions on learning outcomes.

5.1.2. Studied that do not compare apps/app versions

Studies in this category either compared language learning apps to other methods of teaching or investigated the effectiveness of them in a one group, pre-test/post-test design. 18 of 22 studies found explicitly positive results by learners using apps.

With regards to study design, Figure 13 shows that non-randomised comparisons and one group (pre-test/post-test) studies generally found language learning apps effective in improving language proficiency. However, the results of RCTs are more heterogenous: three[19,31,32] (all rated "weak" on the RoB assessment) found positive evidence for language learning apps and one[11] found evidence towards the positive direction. Meanwhile, one RCT (Ramírez et al., 2018) found negative evidence against apps and another (Rachels & Rockinson-Szapkiw, 2018) found no difference between apps and other methods of teaching.

Among those RCTs, Rachels and Rockinson-Szapkiw (2018) offers the most robust results with its rigorous study design (a "strong" RoB rating) and a large sample size of 467. The authors not only controlled for teaching materials, but also learning environments and age of participants by recruiting intact classes from several grades in the same school. However, the experiment restricted the learning experience to in-classroom only, regardless of treatment. In other words, while its design removed the confounder of unequal access time to teaching materials, it also removed mobile apps' distinctive feature of "anywhere, anytime" for learners. Furthermore, participants using Duolingo still improved from pre-test to post-test despite performing equivalently with their peers in the control group. The other "strong" RCT (Liakin et al., 2015) found

partially positive results in using ASR-powered app, in that participants improved on phoneme production but not perception. The author contends that this effect might result from greater in-group variability in the ASR group (Liakin et al., 2015). As well, the authors maintain that this gap in acquisition is in line with previous research, suggesting that phoneme perception is harder to acquire due to complexities in the listening environment (Liakin et al., 2015).

The only study that found negative evidence against language learning apps in this category is an RCT with 96 participants studying an indigenous language in Canada (Ramírez et al., 2018). The data showed that only the control group (who learned with a teacher) increased their vocabulary test scores and all tests were reliable and valid. The lack of improvement could result from the insufficient engagement due to the low level of challenges in the researcher-created app, according to the interviews with participants in this study (Ramírez et al., 2018). The test scores reveal that participants already knew over 70% of the tested words before the intervention (see Table 8). Since over half of this studies in this category are commercially available apps and found positive effects, it is possible that commercial app developers have refined these features during the development stage.

Turning now to the results of non-RCT studies, the overwhelming majority of studies reached a positive conclusion. Even so, two cohort studies with small sample sizes concluded that Duolingo's effects in improving Italian pronunciation (Martinelli, 2016) or in teaching Turkish (Loewen et al., 2019) are less than ideal. In the case of Loewen et al. (2019), nine participants studied Turkish as complete beginners and used Duolingo for the recommended 34 hours of study that would equal one semester of college course. Test scores on a first-semester university Turkish course exam expose that while they were not complete beginners anymore, only one participant received a passing grade. An examination on study time and test scores show that expectedly, usage time is associated with learning results. Additionally, participants' learning journal entries reveal that while Duolingo is enjoyable, the gamification features could be so distracting that "gaining XP points" and competing against users overpowers language learning. The results highlight the fact that while language learning apps are effective in teaching *some* language to its users, its overall efficacy could be overestimated.

Concerning efficacy, it is necessary to highlight the line of research done by Vesselinov and colleagues (2019; 2012, 2015, 2016b, 2016a, 2019a, 2019b). The research group

employed similar methodologies (large cohort, pre-test/post-test) and data collection practices to evaluate the effectiveness of a variety of language learning apps, of which seven are included in this review. For all studies, Vesselinov and colleagues employed validated and reliable placement tests to evaluate the progress of learning by participants. In implementation, these studies replicated the real-world conditions in that the studies were mostly conducted online, and participants were instructed to use the apps at their discretion. After several rounds of random sampling, a minimum of two hours of study and non-use of other language courses were required for participants to be included in the final sample. The behaviours of the sample likely resembled those of actual users. With the results from their Duolingo study, Vesselinov and Grego (2012) assert than 34 hours of study on Duolingo would equal to one semester of college instruction; similar recommendations were made for other apps. Although the test measures are shown to be validated, it is unclear how comparable it is to university course assessment with the absence of an explicit comparison. Furthermore, Loewen et al. (2018) casts reasonable doubts against its validity, albeit with a small sample. At last, it is important to note that Vesselinov's research team conducted these studies under commission from the apps' developing companies. Even though a statement of academic independence was included in each report, one should take caution in interpreting these results. Funding source was not included as a criterium in the risk of bias assessment because it was not included in the EPHPP protocol.

Like studies that compare app versions, most studies in this category recruited adults as participants. Two studies that reached "negative" or "no difference" conclusions both recruited children as participants. However, the small number prevents me from drawing conclusions about the relationship between this demographic and the direction of outcomes. On the other hand, the range of sample size in this category is larger, spanning from single digit (5) to almost 500. While many studies with large sample size found positive evidence for apps, they either have "weak" risk of bias ratings or employed a one group, pre-test/post-test design.

5.1.3. Language skills

Figure 14 displays the relationship between direction of effects, risk of bias, and concerned language skills. Vocabulary, being the most investigated language skill, has an overwhelming amount of supporting evidence from studies. This phenomenon is in line

with the findings from previous reviews (Heil et al., 2016) on the trends of mobile language app development. Meanwhile, there are different directions of effects for every language skill except for comprehension and orthography. From existing evidence, it is only appropriate to conclude that while a variety of language skills were measured in studies, vocabulary, lexicogrammar, and reading have received the most attention in research on language learning apps and the evidence of their effectiveness is mostly positive.

5.2. Further discussion

Some results from the current study resemble those from previous reviews while other demonstrate new trends in the industry and in research. Similar to Burston's (2015) results on MALL, studies in language learning apps primarily use adults as participants. At the same time, vocabulary is the most widely investigated language skill, as such found Heil et al. (2016). Nevertheless, whereas Burston (2015) and Viberg and Grönlund (2012) located mostly short experiments with small samples, this review found numerous studies with large sample sizes and long interventions. This highlights the increase in attention and resources being allocated to this topic.

On the other hand, the study results reflect Burston's (2015) observation that implementation studies are rare, and the study designs are often unsound. Although the number of studies is higher in this review because of my adoption of the best-evidence approach, this number would be significantly lower if I had followed his choice of only reviewing experiments with duration longer than a month. Only 10 comparison studies would meet this requirement.

In terms of learning outcomes, the results from the current review echoes previous those of ones. Burston (2015) found that the majority of studies found positive learning outcomes especially in vocabulary acquisition and Viberg and Grönlund (2012) saw similar trends. Gangaiamaran and Pasupathi (2017) contend that mobile apps help improve listening skills the most, a finding somewhat reflected in this study as well. Contradictory to Viberg and Grönlund's (2012) finding is that, lexicogrammar follows vocabulary to be the second most popular area of interest. Comparing effect sizes across these studies is, however, inappropriate because there are no syntheses of results from the previous reviews.

The bottom-line finding of the current review is that, despite the overall inadequate quality of studies, language learning apps do offer a plausible alternative for learning

some foreign language for independent users. That only one study (Nicholes, 2016) found apps hindering writing development and another (Ramírez et al., 2018) finding no difference in vocabulary scores before and after the treatment points out the effectiveness of language learning apps.

The question is then, compared to what are apps more or less effective?

Studies like those of Rachels and Rockinson-Szapkiw(2018) found that learning with apps does not differ from learning with teacher. Others found that learning with apps are better than with other media or with teachers (Cavus & Ibrahim, 2017; Purgina et al., 2020; Wu, 2015a, 2015b). Mobile apps provide an incomparable degree of access and mobility for learners who cannot commute for face-to-face foreign language education. Furthermore, averaging £10/month for subscriptions and some free of cost, apps are considerably more affordable than teacher-led courses.

At last, it is unclear whether language learning apps suit the learning needs for learners with various levels of proficiency. Due to the heterogeneity in measured and terminologies used by studies I did not attempt to categorise initial proficiencies of participants. But the descriptions in many studies reveal that most participants are either beginners or lower level users in their foreign language (Hao et al., 2019; Loewen et al., 2019; Vesselinov & Grego, 2016b). Even for rather proficient university students who wish to improve in academic English skills, their language proficiency levels are unclear without the provision of standardised test scores (Çakmak & Erçetin, 2018; Wu, 2015a, 2015b). The current research seems to leave advanced learners out of the discussion. While different language learning apps serve various purposes to their target users, a closer examination on how different levels of users benefit from them is needed.

CHAPTER 6. Conclusions

The following research questions guided this systematic review and I summarise the answers for them below:

RQ1. What is the extent of the literature on the effectiveness of language learning applications on smart devices?

Research on language learning apps is increasing in size and numbers. With a best-evidence approach and a lenient criterium for inclusion, the systematic search identified 32 primary studies covering a dozen countries, nine languages, and all age groups. More than ten commercially available apps were surveyed along with a larger number of researcher-developed ones. The number of participants across all studies total more than 3000 people. Quantitative measures on language learning outcomes cover all language skills, ranging from vocabulary and lexicogrammar to pronunciation and orthography. In addition, scholars are interested in two types of comparisons: 1. comparing between apps or app versions to investigate some functions' effects and 2. comparing between apps and other methods of language learning.

RQ2. What is the evidence of their effects on outcomes related to language proficiency?

Using a variety of learning outcome measures, 25 of 32 studies unequivocally find language learning apps to help learners improve language skills regardless of their functions. Most studies that investigated specific app functions favoured the addition thereof, including gamification, self-regulation, and more access to multimedia resources. Cohort studies, non-randomised comparisons and RCTs all found participants using language learning apps to improve from pre-test to post-test. Reported effect sizes range from small to large. In only two studies, participants received either lower or equivalent scores after learning language with apps. Some studies found learning with an app as effective as in a face-to-face classroom. In general, the evidence is of low quality; only four of 32 studies are considered "strong" in the risk of bias assessment.

RQ2a. Is the effectiveness of these apps different for different language skills?

Vocabulary was studied by the most studies (21), followed by lexicogrammar and reading skills. Studies found predominantly positive effects in all language skills. Only one study found that participants decreased on their writing test score after using an app. In another study, participants who used an app to learn vocabulary did not improve their performance. It is inappropriate to infer correlation between

effectiveness and language skills due to the absence of meta-analyses and heterogeneity of study characteristics.

RQ2b. Is the effectiveness of these apps mediated by user characteristics like age? The majority of studies recruited adults as their participants. Seven recruited teenagers and three children under the age of 12. Two of four studies that found non-positive effects of apps recruited children as participants. However, without meta-analyses and a larger number of adequately homogenous studies, it is inappropriate to infer any relationship between age group and app effectiveness.

6.1. Implications for language learners

For language learners, apps provide economical and accessible educational resources. Learners could receive language input and organised courses and achieve learning outcomes similar to those in traditional language courses. However, mobile app users should recognise that all apps are not the same and personal usage significantly impacts learning outcomes. Whereas some apps offer full-package learning experiences, others are more valuable for improving certain language skills. Several apps even provide opportunities for social interaction with other language learners.

The current evidence suggests that beginners might benefit from using language apps. Evidence on learning outcomes by advanced learners is absent from the literature. Serious and autonomous learners, on the other hand, can seek to augment their app learning experiences with external resources. Comparison between apps is beyond the scope of this review, so I will make no recommendation on app selection. Apps are equipped with a variety of functions and users should choose the app according to their preferences.

6.2. Limitations of the review

During the review process I have tried to be as comprehensive and as methodological as possible. Nonetheless, I fully acknowledge that the review could be improved in several ways.

First, I sought to investigate the effectiveness of language learning apps in general but was negligent that many studies evaluated the effectiveness of specific functions in these apps. While this type of research provides deep insight into the theory behind why these apps work, a narrower focus would have allowed for more in-depth discussions.

Secondly, due to time constraint the second reviewer could not conduct data extraction. Therefore, data extraction criteria and instructions might need adjustment for clarity.

Finally, during this review, more relevant studies might have been published. For instance, Duolingo released a self-produced efficacy trial after the conclusion of data collection (Jiang et al., 2020). It is impossible for me to continuously include new studies during the writing of this report. However, the appearance of these new studies signals the increase of interests in this area. The current review serves as a point of departure for future projects.

6.3. Recommendations for future research

The risk of bias assessment reveals that most studies were inadequate in study design and reporting practices. Researchers should strive to produce rigorous studies with comparisons and of more control over the study conditions. Data collection measures should be validated and tested for reliability to improve the trustworthiness of their results. As for reporting, authors should try to report standardised effect sizes to allow for horizontal comparison.

This review exposes new directions for future research. First, it shows that while there is abundant evidence for adults, not enough attention has been dedicated to younger language learners. Similarly, studies have disproportionally engaged language learners with lower levels of proficiency. Future research should investigate language learning apps' effectiveness for young learners and advanced learners. At last, it is understandable that English as the global lingua franca received the most attention. But the potential of mobile language learning could be realised in other areas, such as promoting endangered and minority languages (e.g. García-Mencía et al., 2016; Ramírez et al., 2018). More research should investigate apps' contribution in improving learning outcomes in a more diverse population and in a broader range of languages.

References

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning and Technology*, *5*(1), 202–232. https://www.ncbi.nlm.nih.gov/pubmed/11516946%0Ahttps://linkinghub.elsevier.com/retrieve/pii/S0960-9822(01)00369-4%0Ahttps://www.sciencedirect.com/science/article/pii/S0960982201003694%0Ahttps://ac.els-cdn.com/S0960982201003694/1-s2.0-S0960982201003694-mai

Alter, S. (2020). Making sense of smartness in the context of smart devices and smart systems. *Information Systems Frontiers*, *22*(2), 381–393. https://doi.org/10.1007/s10796-019-09919-9

Ayres, P., & Sweller, J. (2014). The split attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 206–226). Cambridge University Press.

Baddeley, A. D. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge University Press.

Basak, S. K., Wotto, M., & Bélanger, P. (2018). E-learning, m-learning and d-learning: Conceptual definition and comparative analysis. *E-Learning and Digital Media*, *15*(4), 191–216. https://doi.org/10.1177/2042753018785180

Bhide, A. (2017). Instructional methods for promoting the development of orthographic and phonological knowledge in second language learners of Indic languages [University of Pittsburgh]. In *ProQuest Dissertations and Theses* (Issue 10645799). https://search.proquest.com/docview/1943998636?accountid=13042

Bodleian Libraries. (2020). *SOLO - Search Oxford Libraries Online (Classic): Subject searching*. Oxford LibGuides. https://libguides.bodleian.ox.ac.uk/c.php?g=423014&p=2890504

Boland, A., Cherry, G., & Dickson, R. (Eds.). (2017). *Doing a systematic review: A student's guide* (2nd ed.). SAGE Publications Ltd.

Boutron, I., Page, M. J., Higgins, J. P. T., Altman, D. G., Lundh, A., & Hróbjartsson, A. (2019). Chapter 7: Considering bias and conflicts of interest among the included

studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019)*. Cochrane. www.training.cochrane.org/handbook.

Burston, J. (2014). The reality of MALL: still on the fringes. *CALICO Journal*, *31*(1), 103–125. https://doi.org/10.11139/cj.31.1.103-125

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*, *27*(1), 4–20. https://doi.org/10.1017/S0958344014000159

Çakmak, F., & Erçetin, G. (2018). Effects of gloss type on text recall and incidental vocabulary learning in mobile-assisted L2 listening. *ReCALL*, *30*(1), 24–47. https://doi.org/10.1017/S0958344017000155

Cavus, N., & Ibrahim, D. (2017). Learning English using children's stories in mobile devices. *British Journal of Educational Technology*, *48*(2), 625–641. https://doi.org/10.1111/bjet.12427

Chalmers, H. (2019). *Leveraging the L1: The role of EAL learners' first language in their acquisition of English vocabulary Organised a Symposium on Multiple Methods in Education Research View project* (Issue February) [Oxford Brookes University]. https://doi.org/10.13140/RG.2.2.20764.13445

Chen, C.-M., Chen, L.-C., & Yang, S.-M. (2019). An English vocabulary learning app with self-regulated learning mechanism for promoting learning performance and motivation. *Computer Assisted Language Learning*, *32*(3), 237–260. https://doi.org/10.1109/IIAI-AAI.2018.00040

Chen, C.-M., Liu, H., & Huang, H.-B. (2019). Effects of a mobile game-based English vocabulary learning app on learners' perceptions and learning performance: A case study of Taiwanese EFL learners. *ReCALL*, *31*(2), 170–188. https://doi.org/10.1017/S0958344018000228

Chen, M.-P., Wang, L.-C., Chen, H.-J., & Chen, Y.-C. (2014). Effects of type of multimedia strategy on learning of Chinese characters for non-native novices. *Computers & Education*, *70*, 41–52. http://search.ebscohost.com/login.aspx?direct=true&db=eax&AN=92873346&site=ehost-live&authtype=ip,uid

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal*, *80*(2), 183–198.

https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning.* John Wiley & Sons.

Clement, J. (2020). *Number of apps available in leading app stores 2020*.

de Groot, A. M. B., & van Hell, J. G. (2005). The learning of foreign language vocabulary. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: psycholinguistic approaches* (pp. 9–29). Oxford University Press. https://doi.org/10.1016/j.cell.2009.04.050.

de Jong, T., Specht, M., & Koper, R. (2010). A study of contextualised mobile information delivery for language learning. *Educational Technology and Society*, *13*(3), 110–125. https://search.proquest.com/docview/1287030726?accountid=13042

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification." *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek 2011*, 9–15.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004

EdSurge-Pearson. (2020). *Decoding Adaptive*. http://d3btwko586hcvj.cloudfront.net/static_assets/PearsonDecodingAdaptiveWeb.pdf

Education Endowment Foundation. (n.d.-a). *Evidence reviews*. Retrieved March 3, 2020, from https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/review-resources/

Education Endowment Foundation. (n.d.-b). *Foreign language learning and its impact on wider academic outcomes: Protocol for a rapid evidence assessment*.

Effective Public Health Practice Project. (2003). Quality assessment tool for quantitative studies. *Effective Public Health Practice Project*, 2–5. https://www.ephpp.ca/PDF/Quality Assessment Tool_2010_2.pdf

Feldman, A., & Carson, B. (2019, July 16). Next billion-dollar startups 2019. *Forbes*. https://www.forbes.com/sites/amyfeldman/2019/07/16/next-billion-dollar-startups-2019/#2e659a9c1d43

Ganan, D., Caballe, S., Conesa, J., Barolli, L., Kulla, E., & Spaho, E. (2014). A systematic review of multimedia resources to support teaching and learning in virtual environments. *Proceedings - 2014 8th International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2014*, 249–256. https://doi.org/10.1109/CISIS.2014.35

Gangaiamaran, R., & Pasupathi, M. (2017). Review on use of mobile apps for language learning. *International Journal of Applied Engineering Research*, *12*(21), 11242–11251.

García-Mencía, R., López-López, A., & Muñoz Meléndez, A. (2016). An audio-lexicon Spanish-Nahuatl: using technology to promote and disseminate a native Mexican language. In S. Papadima-Sophocleous, L. Bradley, & S. Thouësny (Eds.), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 155–159). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.554

García Botero, G., Questier, F., & Zhu, C. (2019). Self-directed language learning in a mobile-assisted, out-of-class context: do students walk the talk? *Computer Assisted Language Learning*, *32*(1–2), 71–97. https://doi.org/10.1080/09588221.2018.1485707

Georgiev, T., Georgieva, E., & Smrikarov, A. (2004). M-learning - a new stage of e-learning. *International Conference on Computer Systems and Technologies*, 1–5. https://doi.org/10.1145/1050330.1050437

Godwin-Jones, R. (2011). Emerging technologies: Mobile apps for language learning. *Language Learning and Technology*, *15*(2), 2–11.

Godwin-Jones, R. (2017). Emerging technologies: Smartphones and Language Learning. *Language Learning & Technology*, *21*(2), 3–17. http://llt.msu.edu/issues/june2017/index.html%0Ahttps://eric.ed.gov/?q=Godwin-Jones&id=EJ1142376

Gressick, J., & Langston, J. B. (2017). The guilded classroom: Using gamification to engage and motivate undergraduates. *Journal of the Scholarship of Teaching and Learning*, *17*(3), 109–123. https://doi.org/10.14434/v17i3.22119

Guaqueta, C. A., & Castro-Garces, A. Y. (2018). The use of language learning apps as a didactic tool for EFL vocabulary building. *English Language Teaching*, *11*(2), 61. https://doi.org/10.5539/elt.v11n2p61

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? - A literature

review of empirical studies on gamification. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 3025–3034. https://doi.org/10.1109/HICSS.2014.377

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers and Education*, *80*, 152–161. https://doi.org/10.1016/j.compedu.2014.08.019

Hao, Y., Lee, K. S., Chen, S. T., & Sim, S. C. (2019). An evaluative study of a mobile application for middle school students struggling with English vocabulary learning. *Computers in Human Behavior*, *95*, 208–216. https://doi.org/10.1016/j.chb.2018.10.013

Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, T. (2016). A review of mobile language learning applications: Trends, challenges, and opportunities. *The EUROCALL Review*, *24*(2), 32–50. https://doi.org/10.5267/j.msl.2018.7.002

Hsu, C. K., Hwang, G. J., & Chang, C. K. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. *Computers and Education*, *63*, 327–336. https://doi.org/10.1016/j.compedu.2012.12.004

Huang, C. S. J., Yang, S. J. H., Chiang, T. H. C., & Su, A. Y. S. (2016). Effects of Situated Mobile Learning Approach on Learning Motivation and Performance of EFL Students. *Journal of Educational Technology & Society*, *19*(1), 263–276. https://search.proquest.com/docview/1768612513?accountid=13042

Institute of Education Sciences, & Foundation, N. S. (2013). *Common guidelines for education research and development* (Issue August). https://doi.org/10.1080/00220973.2013.813364

Jiang, X., Rollinson, J., Plonsky, L., & Pajak, B. (2020). Duolingo efficacy study: Beginning-level courses equivalent to four university semesters. In *Duolingo Research Report*. https://www.duolingo.com/efficacy

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon. http://solo.bodleian.ox.ac.uk/OXVU1:oxfaleph011549664

Kukulska-Hulme, A. (2009). Will mobile learning change language learning? *ReCALL*, *21*(2), 157–165. https://doi.org/10.1017/S0958344009000202

Kukulska-Hulme, A., & Shield, L. (2008). An overview of mobile assisted language

learning: From content delivery to supported collaboration and interaction. *ReCALL*, *20*(3), 271–289. https://doi.org/10.1017/S0958344008000335

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation and Gaming*, *45*(6), 752–768. https://doi.org/10.1177/1046878114563660

Landers, R. N., & Landers, A. K. (2014). An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance. *Simulation and Gaming*, *45*(6), 769–785. https://doi.org/10.1177/1046878114563662

Lardinois, F. (2018, August 2). Duolingo hires its first chief marketing officer as active user numbers stagnate but revenue grows. *TechCrunch*. https://techcrunch.com/2018/08/01/duolingo-hires-its-first-chief-marketing-officer-as-active-user-numbers-stagnate/?guccounter=1&guce_referrer=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnLw&guce_referrer_sig=AQAAAFLVTZP6Gbsa9HPOqsi46WyzIjacGO9y7Xjodl91bLt47fnvX8_Da9J

Li, Y., Chen, W., Liu, G., Wei, L., Liu, C., Si, L., Zhang, J., & Yang, K. (2017). Analysis of quality assessment tools in Campbell systematic reviews. *Abstracts of the Global Evidence Summit*, *9 Supp 1*. https://doi.org/10.1002/14651858.cd201702

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French/y/. *CALICO Journal*, *32*(1), 1–25. https://doi.org/10.1558/cj.v32i1.25962

Liu, J. J. (2016). Designing intelligent language tutoring system for learning Chinese characters [Iowa State University]. In *ProQuest Dissertations and Theses* (Issue 10167832). https://search.proquest.com/docview/1831441077?accountid=13042

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293–311. https://doi.org/10.1017/S0958344019000065

Long, M. H. (1981). Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences*, *379*(1), 259–278. https://doi.org/10.1111/j.1749-6632.1981.tb42014.x

Long, M. H. (1985). Input and second language acquisition theory. In S. M. Gass & C. G. Madden (Eds.), *Input in Second Language Acquisition* (pp. 377–393). Newbury

House Publishers. https://weblearn.ox.ac.uk/access/content/group/cd464c28-e981-4dcc-af89-945b50a3ef48/Digitisation_HD/ALSLA/FSLAT - Foundations of SLA Theory/Zobl_1985.pdf

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*(4), 557–587. https://doi.org/10.1017/S0272263199004027

Mackey, A. (2007). *Conversational interaction in second language acquisition: A collection of empirical studies*. Oxford University Press. http://solo.bodleian.ox.ac.uk/OXVU1:LSCOP_OX:oxfaleph016693518

Martinelli, M. (2016). Effectiveness of online language learning software (Duolingo) on Italian pronunciation features: A case study [Oklahoma State University]. In *ProQuest Dissertations and Theses* (Issue 10191215). https://search.proquest.com/docview/1964383737?accountid=13042

Mayer, R. E. (2002). Multimedia learning. In *The Psychology of Learning and Motivation* (Vol. 41). Cambridge University Press. https://doi.org/doi.org/10.1017/CBO9781139164603

Mayer, R. E., & Anderson, R. B. (1991). Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology*, *83*(4), 484–490.

Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology*, *28*(5), 653–660. https://doi.org/10.1002/acp.3050

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/bm.2012.031

Memrise. (n.d.).

Meticulous Market Research Pvt. Ltd. (2020). *Online language learning market by product (SaaS, apps, tutoring), mode (consumer, government, K-12, corporate), language (English, German, Japanese, Korean, Mandarin Chinese) and geography – Global forecast to 2027*. https://www.meticulousresearch.com/product/online-language-learning-market-5025/?utm_source=globenewswire.com&utm_medium=press release&utm_campaign=paid

Mohamad Ali, A. Z., Segaran, K., & Hoe, T. W. (2016). Effects of verbal components in

3D talking-head on pronunciation learning among non-native speakers. *Journal of Educational Technology & Society*, *18*(2), 380–389. https://www.jstor.org/stable/10.2307/jeductechsoci.18.2.313

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Estarli, M., Barrera, E. S. A., Martínez-Rodríguez, R., Baladia, E., Agüero, S. D., Camacho, S., Buhring, K., Herrero-López, A., Gil-González, D. M., Altman, D. G., Booth, A., … Whitlock, E. (2016). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Revista Espanola de Nutricion Humana y Dietetica*, *20*(2), 148–160. https://doi.org/10.1186/2046-4053-4-1

Moreno, I. D., & Gatewood, K. (2019). Gamification of foreign language mobile-learning: A quantitative study on Spanish-as-a-foreign language adult learners' satisfaction, perception, motivation, and knowledge [Keiser University]. In *ProQuest Dissertations and Theses* (Issue 13897893). https://search.proquest.com/docview/2284212925?accountid=13042

Nicholes, J. (2016). Measuring the impact of language-learning software on test performance of Chinese learners of English. *TESL-EJ*, *20*(2), 1–20. http://search.ebscohost.com/login.aspx?direct=true&db=eax&AN=118290156&site=ehost-live&authtype=ip,uid

O'Dea, S. (2020). *Global market share smartphone operating systems of unit shipments 2014-2023*. https://www.statista.com/statistics/277048/global-market-share-forecast-of-smartphone-operating-systems/

Orhan Göksün, D., & Gürsoy, G. (2019). Comparing success and engagement in gamified learning experiences via Kahoot and Quizizz. *Computers and Education*, *135*(February), 15–29. https://doi.org/10.1016/j.compedu.2019.02.015

Ou-Yang, F. C., & Wu, W. C. V. (2017). Using mixed-modality vocabulary learning on mobile devices: Design and evaluation. *Journal of Educational Computing Research*, *54*(8), 1043–1069. https://doi.org/10.1177/0735633116648170

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 1–10. https://doi.org/10.1186/s13643-016-0384-4

Paramythis, A., & Loidl-Reisinger, S. (2004). Adaptive learning environments and e-learning standards. *Electronic Journal of E-Learning*, *2*(1), 181–194.

https://files.eric.ed.gov/fulltext/EJ1099144.pdf%0A

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences*. Blackwell Publishing.

Pica, T. (1994). Research on negotiation: What does it reveal about second language learning conditions, processes, and outcomes? *Language Learning*, *44*(3), 493–527. https://doi.org/10.1111/j.1467-1770.1994.tb01115.x

Pica, T. (1996). Do second language learners need negotiation. *IRAL - International Review of Applied Linguistics in Language Teaching*, *34*(1), 22.

Pica, T., Young, R., & Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly*, *21*(4), 737–758. https://doi.org/10.7575/aiac.ijalel.v.4n.4p.108

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior*, *19*(2), 221–243. https://doi.org/10.1016/S0747-5632(02)00015-8

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., & Britten, N. (2006). Narrative synthesis in systematic reviews: A product from the ESRC methods programme. *ESRC Methods Programme*, *2006*, 93. https://doi.org/10.13140/2.1.1018.4643

Purgina, M., Mozgovoy, M., & Blake, J. (2020). WordBricks: Mobile technology and visual grammar formalism for gamification of natural language grammar acquisition. *Journal of Educational Computing Research*, *58*(1), 126–159. https://doi.org/10.1177/0735633119833010

Rachels, J. R., & Rockinson-Szapkiw, A. J. (2016). The effect of gamification on elementary students' Spanish language achievement and academic self-efficacy [Liberty University]. In *ProQuest Dissertations and Theses* (Issue 10141916). https://search.proquest.com/docview/1823238499?accountid=13042

Rachels, J. R., & Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning*, *31*(1), 72–89. https://doi.org/10.1080/09588221.2017.1382536

Ramírez, G., ElMiligi, H., Walton, P., & Billy, C. (2018). Revitalization of an indigenous language: Which is better the teacher or the app? *International Conference on E-Learning*, *2018-July*, 344–350, XVII.

https://search.proquest.com/docview/2081757033?accountid=13042

Rosell-Aguilar, F. (2017). State of the app: A taxonomy and framework for evaluating language learning mobile applications. *CALICO Journal*, *34*(2), 243–258. https://doi.org/10.1558/cj.27623

Shadiev, R., Hwang, W.-Y. Y., & Huang, Y.-M. M. (2017). Review of research on mobile language learning in authentic environments. *Computer Assisted Language Learning*, *30*(3), 284–303. https://doi.org/10.1080/09588221.2017.1308383

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Altman, D. G., Booth, A., Chan, A. W., Chang, S., Clifford, T., Dickersin, K., Egger, M., Gøtzsche, P. C., Grimshaw, J. M., Groves, T., Helfand, M., … Whitlock, E. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ (Online)*, *349*. https://doi.org/10.1136/bmj.g7647

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5–11.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Technavio. (2020). *Global online language learning market 2020-2024 | 18% CAGR projection through 2024*. https://www.businesswire.com/news/home/20200227005365/en/Global-Online-Language-Learning-Market-2020-2024-18

Terantino, J. (2016). Examining the effects of independent MALL on vocabulary recall and listening comprehension: An exploratory case study of preschool children. *CALICO Journal*, *33*(2), 260–277. https://doi.org/10.1558/cj.v33i2.26072

Teske, K. (2017). Duolingo. *CALICO Journal*, *34*(3), 393–402. https://www.duolingo.com/

The Campbell Collaboration. (n.d.). *Evidence synthesis tools for Campbell authors*. Retrieved April 1, 2020, from https://campbellcollaboration.org/author-tools.html

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evid Based Nurs*, *1*(3), 176–184. https://doi.org/10.1111/j.1524-475X.2004.04006.x

Thongsri, N., Shen, L., & Yukun, B. (2019). Does academic major matter in mobile assisted language learning? A quasi-experimental study. *International Journal of Information and Learning Technology*, *36*(1), 21–37. https://doi.org/10.1108/IJILT-05-2018-0043

Unsworth, S. (2016). Quantity and quality of language input in bilingual language development. *Bilingualism across the Lifespan: Factors Moderating Language Proficiency.*, 103–121. https://doi.org/10.1037/14939-007

Usai, F., O'Neil, K. G., Newman, A. J., ONeil, K. G. R., & Newman, A. J. (2018). Design and empirical validation of effectiveness of LANGA, an online game-based platform for second language learning. *IEEE Transactions on Learning Technologies*, *11*(1), 107–114. https://search.proquest.com/docview/2034278217?accountid=13042

Valencia, J. A. Á. (2014). *Language, learning and identity in social networking sites for language learning: The case of busuu*. University of Arizona.

Verdu, E., Regueras, L. M., Jesus Verdu, M., de Castro, J., & Angeles Perez, M. (2008). Is adaptive learning effective? A review of the research. *Wseas: Advances on Applied Computer and Applied Computational Science*, 710–715.

Vesselinov, R., & Grego, J. (2012). *Duolingo Effectiveness Study*. http://comparelanguageapps.com/documentation/DuolingoReport_Final.pdf

Vesselinov, R., & Grego, J. (2015). *Efficacy of a New Language App*.

Vesselinov, R., & Grego, J. (2016a). *The Babbel Efficacy Study* (Issue September). http://comparelanguageapps.com/documentation/Babbel2016study.pdf

Vesselinov, R., & Grego, J. (2016b). *The Busuu Efficacy Study* (Issue May). http://comparelanguageapps.com/documentation/The_busuu_Study2016.pdf

Vesselinov, R., & Grego, J. (2019a). *Mango Languages Efficacy Study*. http://comparelanguageapps.com/documentation/Mango_Languages_FinalReport2019.pdf

Vesselinov, R., & Grego, J. (2019b). *Pimsleur Efficacy Study*. http://comparelanguageapps.com/documentation/Pimsleur_EfficacyStudy2019.pdf

Vesselinov, R., Grego, J., Sacco, S. J., & Tasseva-Kurktchieva, M. (2019). *The 2019 Rosetta Stone Efficacy Study*. http://comparelanguageapps.com/documentation/The2019_RS_FinalReport.pdf

Viberg, O., & Grönlund, Å. (2012). Mobile assisted language learning: A literature

review. *CEUR Workshop Proceedings*, *955*, 9–16.

Weiss, C. H. (1998). *Evaluation: methods for studying programs and policies.* (2nd ed.). Prentice-Hall.

White, L. (1987). Against comprehensible input: The input hypothesis and the development of second-language competence. *Applied Linguistics*, *8*(2), 95–110. https://doi.org/10.1093/applin/8.2.95

Wu, Q. (2015a). Designing a smartphone app to teach English (L2) vocabulary. *Computers and Education*, *85*, 170–179. https://doi.org/10.1016/j.compedu.2015.02.013

Wu, Q. (2015b). Pulling mobile assisted language learning (MALL) into the mainstream: MALL in broad practice. *PLoS ONE*, *10*(5). https://doi.org/10.1371/journal.pone.0128762

Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers and Education*, *140*(May), 103599. https://doi.org/10.1016/j.compedu.2019.103599

Zainuddin, Z. (2018). Students' learning performance and perceived motivation in gamified flipped-class instruction. *Computers and Education*, *126*(July), 75–88. https://doi.org/10.1016/j.compedu.2018.07.003

Appendix A PRISMA Protocol

1. Introduction

1.1 Working title:

Thirteen years since the first iPhone: A systematic review on the effectiveness of language learning apps on smart devices

1.2 Rationale

Using applications on smart devices to learn languages has been in the spotlight since the release of the first smart phone in 2007 and its increasing accessibility. The prosperity of the software development industry, married with increasing globalisation and online learning efforts, contributes to the rapid growth of this trend. The slogan "anytime, anywhere" appeals to language learners in the mobile internet age, and companies that develop such apps use emerging technology to market and develop their products as "world's best way to learn a new language (Teske, 2017)" or "the fastest way to learn a language (*Memrise*, n.d.)." I am curious about the empirical evidence on which these claims are based and whether the evidence is valid from a theory-informed, applied linguistic perspective.

Moreover, there is an existing body of literature on technology-enhanced learning, namely mobile-assisted language learning (MALL) and computer-assisted language learning (CALL). While both CALL and MALL have been extensively researched since the 1980s, the common definitions are rather "archaic." Particularly, the MALL literature includes all kinds of learning methods powered by any mobile device (e.g. MP3, even flashcards), usually as an aid to classroom teaching. With use of new technology and methods such as machine learning and artificial intelligence, new apps on smart devices posit innovative functions, such as course personalisation and free progression.

Given the surge in popularity of mobile apps for learning languages and the strong claims of efficacy made by their manufacturers it is clearly important that potential consumers understand the likely effects of learning with such apps. A number of individual primary studies have been conducted on the substantive effects of apps like these (e.g. García Botero, Questier, & Zhu, 2019; Rachels & Rockinson-Szapkiw, 2018; Valencia, 2014; Vesselinov & Grego, 2016). However, it appears that no systematic review synthesising the findings of all studies on this type of technology has been conducted. To fully understand the potential of this form of MALL, it is imperative that all studies be considered together.

1.3 Objectives

This study systematically reviews the literature on the effectiveness of language learning applications on smart devices. It will identify the underlying theoretical framework for these applications and synthesize the current empirical evidence on the effectiveness of these applications in language learning.

1.4 Research questions

RQ1. What is the extent of the literature on the effectiveness of language learning applications on smart devices?

RQ2. What is the evidence of their effects on outcomes related to language proficiency?

RQ2a. Is the effectiveness of these apps different for different language skills?

RQ2b. Is the effectiveness of these app mediated by user characteristics such as age?

2. Methods

To address the review's research questions, I will examine primary studies published after 2007 (the year when the first smart phones hit the market) that reported quantitative data. Articles will not be excluded based on their language and publication type. See below for a complete list of eligibility criteria with rationale.

2.1 Eligibility criteria

Item	Inclusion	Exclusion	Rationale
Complete reference	Include studies with a complete reference.	Exclude studies with an incomplete reference.	
Time	Include studies published after 2007.	Exclude studies published before 2007.	The first smartphone, the iPhone, was released in 2007. Therefore, it provides a suitable watershed moment for investigating the use of mobile apps on smartphones.
Publication status	Do not exclude studies based on their publication status.		
Language	Do not exclude studies based on the language in which they are written. However, if the main text of the study is in languages other than those I read (English and Chinese), I would make efforts to obtain the translation of the text or to ask a speaker of that language to perform data extraction.		
Duration	Include studies that are longitudinal, with any duration.	Exclude studies that are not longitudinal, for example, one-off cross-sectional studies.	Cross-sectional studies cannot show improvement.
Study design	Include primary studies that use an RCT, non-randomised comparison, or one group, pre-test/post-test design, with, at minimum, one experimental group. Include one cohort, post-test-only design. Include mixed methods studies that comprise of mentioned	Exclude observational studies, ethnographies, secondary data analyses, and studies that only collected qualitative data. Exclude single case studies where individuals are the subjects.	I want to investigate the direct relationship between language learning apps and language proficiency on the basis of substantive, measurable outcomes such as test scores. Primary studies with experimental groups are the more reliable design for making these kinds of causal inferences. I include one cohort, pre-test/post-test designs because they are prevalent in the scoping search.

	designs. Include studies that use intact groups as subjects.		The risk of bias based on study design will be evaluated.
Participants	Include studies on typically developing language learners of all ages and background. Include studies that use intact groups as subjects, if they are reasonably assumed to comprise of typically developing students.	Exclude studies that exclusively recruit non-typically developing participants, for example those with physical disabilities or Developmental Language Disorder.	Since the feature of mobile language learning is "anytime, anywhere," it's reasonable to infer that they are widely accessible to all populations of language learners. However, non-typically developing populations might exhibit traits and factors not applicable to the wide populations. Therefore, I will not include studies that only recruit participants from this population.
Intervention	Include studies that use language learning apps on smart phones or tablets as the intervention for second/foreign language learning for the experimental group. These applications have self-contained content and curriculum which allow users to progress. Examples include Duolingo and Busuu. These apps do not have to be commercially available as long as they have the potential to be used by all populations.	Exclude studies that do not use an app on mobile devices as the intervention and those that *only* use the website versions of the apps. Exclude studies that use apps that require user-generated content and those that only serve as tools to enhance learning experience (e.g. Anki, Quizlet, dictionaries). Exclude studies that use mobile devices as platforms for teaching or as an addition to other methods of teaching. Exclude studies that study first language skills, such as literacy.	Some applications have a website version with different layouts and features from their mobile counterpart, which could impose additional confounding variability to the study results. However, studies that do not limit the platform on which they use the app can be included because it is to the participants' discretion and convenience which one to use, a core feature of mobile learning. Moreover, participants in studies that use mobile devices as platforms or additional teaching tools might benefit from interaction with instructors, which presents additional confounding variability.
Outcome	Include studies that report quantitative data of post-tests on the language skills they seek to measure.	Exclude studies that don't report any quantitative data and those that report survey results as quantitative data. Exclude studies that report only conclusions drawn from quantitative data.	I want to investigate the direct relationship between language learning apps and increased language proficiency on the basis of substantive, measurable outcomes such as test scores. Access to quantitative data allows for cross-study comparison.

2.2 **Information Sources**

To ensure comprehensiveness of this study, I will search for not only peer-reviewed published articles, but also articles from other sources. These sources include but are not limited

to conference proceedings, commercial and/or government reports, trial register (if there is any), master's and doctoral theses.

Since the first smart phone was released in 2007, the literature search will only include those published after this date. Because of this date restriction, it is reasonable to assume that all relevant literature is published electronically; therefore, I will only conduct the search on electronic databases.

The electronic databases I intend to conduct the search on cover fields of education, language and linguistics, psychology, computer science and general social sciences. Moreover, I will also search in thesis, dissertation and abstract databases as well as general academic database such as Google Scholar.

3. Search strategy

3.1 List of databases

I will conduct the main search on the following databases. I identified these relevant databases on Oxford University's library system by using the list of databases by discipline and consultation with an experienced librarian.

Discipline/Category	Database(s)/Journals
Education	The Education Collection on SOLO (incluing ERIC), British Education Index, Child Development & Adolescent Studies, Education Abstracts, EThOS
Linguistics	Linguistics Collection (including LLBA), Journals: CALICO, ReCALL, Language Learning and Technology
Psychology	PsycINFO
General social sciences	ProQuest Social Science Premium Collection, ASSIA Applied Social Sciences Index and Abstracts, SCOPUS, Sociological Abstracts
Multidiscipline	Web of Science Core Collection, Google Scholar
Conference proceedings	Oxford University Research Archive, PapersFirst, ProceedingsFirst (OCLC)
Theses and dissertations	ProQuest Dissertations & Theses Global

I generated a list of databases by selecting those that are relevant on the Search Oxford Libraries Online (SOLO) system under categories of education, linguistics and social sciences. I also consulted with an experienced librarian to remove redundant and irrelevant databases, such as those indexing only books or newspapers. In addition, I will hand search key journals (ReCALL, CALICO, Language Learning and Technology) for relevant articles because of their high relevancy to the topic. I will also conduct a simple search, i.e. not using full search string, on three databases, Child Development & Adolescent Studies, ProQuest Social Science Premium Collection, and Sociological Abstracts, to determine if they could be relevant databases for a full

search. If there are relevant, preliminary results from these databases, I will also conduct a main search on them.

I will conduct an informal search on Google Scholar to include any other articles not generated in the main search on the previously identified databases.

Additionally, I will conduct backward and forward citation searches. Where I identify a study that fits all criteria for inclusion, I will search in its reference list as well as look for subsequently published research that have cited this article for eligible studies.

I will set automatic updates on searched databases for newly published articles during the review process. Given the deadline of this project, my last update on the main search will be on 15th May 2020.

In my preliminary search, I found that many commercial language learning apps such as Duolingo and busuu have commissioned their own research and those reports are publicly available. Therefore, I will also include these reports in the main search.

3.2 Search terms

I generated my search terms from consultation with an experienced research librarian as well as from the literature I read from the scoping search Four main concepts guided my generation of search terms: language, learning, mobile app, and effectiveness. After brainstorming specific search terms with the librarian, I tested the validity and efficiency of them using the SOLO Linguistics Collection. The search terms were modified so that the search results are neither too broad nor too narrow. Moreover, I also compared the search results against a candidate article generated from the scoping search that pass all inclusion criteria to validate the search terms.

I will perform the main search on databases identified above. Then, I will upload the generated citations to Rayyan, where I will resolve duplicate and conduct title and abstract screening.

3.3 Sample search string

The following boolean search string was coined based on the last section:

noft(language* OR vocabulary OR grammar OR syntax) AND
noft(learn* OR acquisition OR acquir*) AND
noft(mobile OR app OR apps OR smart?phone OR iPad OR iPhone OR Android OR iOS OR gamification OR gamified OR game?based OR m-learnig OR e-learning OR MALL OR mobile?assisted OR busuu OR duolingo OR memrise OR babbel) AND
noft(effect* OR impact OR efficacy OR efficien* OR develop*)

Boolean operators allow for combination of keywords to increase the relevance of search results. For a detailed guide on boolean search, see Bodleian Libraries's (2020) guide on Subject searching. To further increase accuracy of the main search, I applied

more search limitations. Keyword search was restricted to anything but full article if possible, and publication date was restricted to after 01/01/2007.

4. Study records

4.1 Data management

Main search results will be uploaded to Rayyan, an Internet based software program for systematic review screening and facilitates collaboration among multiple reviewers. I will use Rayyan for screenings (see 4.2) and Mendeley, a reference manager, for full article review. I will use a spreadsheet on Microsoft Excel to perform data extraction. If extracted data allow for statistical synthesis, I will perform this in SPSS v.27.

At the same time, I will keep track of my actions and records, including search strings, in a Microsoft Word document. All data management and analysis will be done on a MacBook Pro with a Mac OS Catalina 10.15.

4.2 Selection process

i) Title and Abstract Screening

Titles and abstracts generated from the main search will be checked against the inclusion/exclusion criteria. They will be marked for inclusion if they meet all inclusion criteria. They will be excluded if they match any of the exclusion criteria. Records that cannot be reliably excluded on the basis of the information in their titles and abstracts will be marked 'include' and full texts will be sought. I will try to obtain full reports for all titles that appear to meet the inclusion criteria or where there is any uncertainty.

ii) Full Text Screening

Records that cannot be excluded on the basis of the information contained in the titles and abstracts will be forwarded for full text screening. I will then try to obtain the full articles from databases identified in 3.1. If full articles are not available online in those databases, I will contact the author for a full article.

Then, I will assess the full articles for inclusion criteria. I will seek additional information from study authors where necessary to resolve questions about eligibility. Again, the articles will be labelled according to the inclusion/exclusion criteria.

In both screenings, a secondary, independent reviewer will be invited to participate in the screening process. They will screen 5-15% of the total items and the decisions on these articles will be compared with the decision of the main reviewer. I will calculate inter-rater reliability using Cohen's kappa statistic and resolve disagreement through discussion.

Since the purpose of this dual-review is to test the validity and the clarity of my eligibility criteria, and the number of dual-reviewed articles is low, I seek a high simple inter-reliability of 0.8-1.0 (McHugh, 2012). If the inter-rater reliability is below 0.8, I will rectify my eligibility criteria through discussion with the second reviewer. We will record the reasons for excluding

trials. Reviewers are blind to the decisions of one another during the review process. I will also invite my supervisor of this dissertation to Rayyan for collaboration.

4.3 Data extraction process

Using a pre-determined data extraction form, I will independently extract data items onto an Excel spreadsheet. The second, independent reviewer will also perform data extraction for the articles they read. Extracted data from the second reviewer will be compared against my data and we will resolve disagreement through discussion.

5. Outcomes and prioritization

The primary outcomes are quantitative measures of success of language learning, for example, standardized test scores.

6. Risk of bias of individual studies

Bias refers to "a systematic error, or deviation from the truth," that can positively or negatively affect the results of a study and vary in magnitude (Boutron et al., 2019, sec. 7.1). In a systematic review, sources of bias include the actions of primary study authors, conflicts of interests, study designs, and actions of the reviewer. Thus, including an assessment of the risks of bias of individual studies in a systematic review is crucial because it informs the audience of the trustworthiness of the studies' findings. A review on the risk of bias assessment tools used in Campbell systematic reviews in the social sciences found that given the lack of discipline standard, the choice of quality assessment tools should depend on the topic and needs of the review (Li et al., 2017). Reviewers in the social sciences widely adopt tools from health sciences, such as the Cochrane Risk of Bias Tool (Li et al., 2017). For this project I followed Chalmers (2019) in using the Effective Public Health Practice Project Quality Assessment Tool for Quantitative Studies (hereafter EPHPP tool; Thomas et al., 2004) because it is suitable for quantitative studies with different research designs. Moreover, Petticrew and Roberts (2006) recommend EPHPP among five other tools for systematic reviews in the social sciences.

The EPHPP tool assesses the risk of bias of a study in the following eight components (2003):

A Selection bias

A.1 Are the individuals selected to participate in the study likely to be representative of the target population?

A.2 What percentage of selected individuals agreed to participate?

B Study design

B.1 What is the study design?

B.2 Was the study described as randomised?

2a) If Yes, was the method of randomization described?

2b) If Yes, was the method appropriate?

C Confounders

C.1 Were there important differences between groups prior to the intervention?

C.2 If yes, what percentage of relevant confounders was controlled?

D Blinding as part of a controlled trial

D.1 Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?

D.2 Were the study participants aware of the research question?

E Data collection practices

E.1 Were data collection tools shown to be valid?

E.2 Were data collection tools shown to be reliable?

F Withdrawals and dropouts

F.1 Were withdrawals and drop-outs reported in terms of numbers and/or reasons per group?

F.2 what is the percentage of participants completing the study?

G Invention integrity

G.1 What percentage of participants received the allocated intervention or exposure of interest?

G.2 Was the consistency of the intervention measured?

G.3 Is it likely that subjects received an unintended intervention (contamination or co-intervention) that may influence the results?

H Analysis

H.1 What is the unit of allocation?

H.2 What is the unit of analysis?

H.3 Are the statistical methods appropriate for the study design?

H.4 Is the analysis performed by intervention allocation status (i.e. intention to treat) rather than the actual intervention received?

On all components except for G and H, the study first receives a local rating of "weak," "moderate," or "strong." Components G and H are not given ratings because while these constructs are important for understanding the results of included studies, potential deviations in these components do not constitute systematic errors caused by the researchers' design and implementation. See Thomas et al. (2004, p. 180) for a detailed explanation. Then, the study receives a global risk of bias rating of "strong," "moderate," or "weak." "Strong" studies have no weak ratings and at least four "strong" ratings; "moderate" studies have less than four "strong" ratings and/or one "weak" rating, and "weak" studies two or more "weak" ratings (Thomas et al., 2004, p. 180). "Weak," "moderate," and "strong" refer to the strength of security in the study; the results of "weak" studies are less trustworthy than those of "moderate" and "strong" studies. Like Chalmers (2019, p. 87), I made minor changes to the terminology and rating guidelines used in my assessment to make them more applicable to the topic and discipline. I will use "non-

randomised comparison" instead of "controlled clinical trial" in the Study Design component so that the term would be more applicable for educational research. For the Analysis component, I will use "learner, class/group, school, local authority/school district" in lieu of "individual, practice/office, institution/organization, community." Under Blinding, if the tests for relevant outcomes did not required human graders (such as to grade pronunciation or fluency), the answer to D.a was defaulted to "no."

The independent reviewer will also conduct this assessment, whose results will be compared against mine. We will resolve differences through discussion.

7. Data synthesis

Following guidelines from Popay et al. (2006), I will conduct a narrative synthesis with extracted data and use tools such as tables and graphs to explain the findings of included studies. This synthesis will explore relationships and findings within and between the included studies. This narrative synthesis will serve as the main component of this review project.

At the same time, if included studies use sufficiently homogeneous study designs and reported measurements, I will also conduct meta-analyses on the effect sizes of their outcomes. When effect sizes are not readily available in the included studies, I will convert outcomes to effect sizes using given information in the study. I will use the Hedge's g (corrected effect size) instead of the Cohen's d because it is more robust when heterogeneity is high between studies in terms of variances and samples sizes (Durlak, 2009; Education Endowment Foundation, n.d.-b).

7.1 Thematic analysis

I will conduct thematic analyses based on patterns found in data synthesis. The number of themes included depend on the degree of heterogeneity of included studies. Possible themes include specific type of intervention, i.e. application; participant demographics, including gender, age and location; duration of intervention

8. Confidence in cumulative estimate

The quality of evidence for all outcomes will be judged based on their risks of bias assessments.

Appendix B Data extraction sheet

DATA EXTRACTION			
Cat.	**Item**	**Response**	**Instructions**
General Information	Date form completed (dd/mm/yy)		Fill in the date of completion of this form.
	Reference		The full reference of the article.
	Source		Identify the source through which the article was specified (e.g. from an electronic database generated from the main search, or through backward or forward citation searches?)
	Language		The language in which the article is written.
	Publication status		The publication type of the document, e.g. a thesis, peer-reviewed journal article, book chapter, etc.
Setting	Time of study		Identify the date(s) the study was conducted as stated by the authors. If this information is not included, leave the date of publication and note "published"
	Duration of intervention		In the format of XX month, XX weeks, XX days, where applicable. Note as "not specified" if this information is not provided in the report.
	Location of the study		What information did the authors provide about the location(s) where the study was conducted? Note as "not specified" if this information is not provided in the report.
Methods	Research question(s)		What are the research questions of the study? Put in double quotation marks ("...") if directly quoted from the study. Note as "inferred" if they were not explicitly stated in the report.
	Study design		What study design is used for the study? For example, is it a quantitative only study or mixed-method study? Is it experimental, quasi-experimental, or pre-experimental (e.g. single group pre-test/post-test)?
	Intervention		What application is the intervention group using? Is it commercially available or developed by the researchers? If the application is not commercially available yet, what description is given about it?
			What operating system and device does the intervention run on (e.g. Android, iOS, Windows; phone, tablet)
	Comparator		What method is the comparator group using? What details are given about the method?
	App version comparison		Indicate whether the comparison is between different versions of the same app, not between app and other learning methods.
	Studied language		What is(are) the target language(s) in this study?
	Materials		What are the materials used in the intervention and comparator? Do these materials focus on specific language skills?
Participants	Source of recruitment		How were the participants recruited in this study (e.g. through a class, a public advertisement, through the app platform)? Note as "not specified" if this information is not provided in the report.

	Demographics		What demographic characters do the participants have in common? If they don't have anything in common, how is this information reported? Some key information include age, gender, first language and language learning experience.
	Sample size		n=total subjects. Fill in the total and final sample size, comprising of both number of people in the intervention group and that in the comparator.
			How big is the final sample size for the intervention group?
			How big is the final sample size for the comparator group?
	Attrition		n=initial subject. How big was the initial sample size?
			n=final subject. How big was the final sample size?
			What is the attrition rate?
	Pre-test or specified proficiency		How is the language proficiency of participants reported in the study: with test scores, certifications, or self-reports? If no specific information is given, what proficiency level do the researchers assume of the participants (note: "assumed")?
Outcomes	Language skills		What language skills do the outcomes measure?
	Outcome measurement		What tests are administered to measure learning outcomes? If the test was developed by the researchers, what details are given about it?
	Outcomes (specific scores)		What are the specific scores in the outcome measurements? Include details such as means, standard deviations, and the subsets of scores if available.
	Effect size		Record the effect size(s) as reported by the researchers, if any. Calculate using the study's information if none was given but there is enough information.
	Conclusion (researcher)		Identify the conclusion(s) as stated by the authors.
	Conclusion (general)		Identify the direction of the researcher's conclusions: positive, negative, neutral, or mixed.

Appendix C Risk of Bias Assessment Template.

Reference			
Component	Questions	Answer	Comments
A. Selection bias	Q1. Are the individuals selected to participate in the study likely to be representative of the target population?		
	Q2. What percentage of selected individuals agreed to participate?		
	Component Rating		
B. Study design	Q1. What is the study design?		
	Q2. Was the study described as randomised?		
	Q3. If Yes, was the method of randomization described?		
	Q4. If Yes, was the method appropriate?		
	Component Rating		
C. Confounders	Q1. Were there important differences between groups prior to the intervention?		
	Q2. If yes, what percentage of relevant confounders was controlled?		
	Component Rating		
D. Blinding (as a part of controlled trial)	Q1. Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?		
	Q2. Were the study participants aware of the research question?		
	Component Rating		
E. Data collection practices	Q1. Were data collection tools shown to be valid?		
	Q2. Were data collection tools shown to be reliable?		
	Component Rating		
F. Withdrawals and dropouts	Q1. Were withdrawals and drop-outs reported in terms of numbers and/or reasons per group?		
	Q2. what is the percentage of participants completing the study?		
	Component Rating		
G. Invention integrity	Q1. What percentage of participants received the allocated intervention or exposure of interest?		
	Q2. Was the consistency of the intervention measured?		
	Q3. Is it likely that subjects received an unintended intervention (contamination or co-intervention) that may influence the results?		
H. Analysis	Q1. What is the unit of allocation?		
	Q2. What is the unit of analysis?		
	Q3. Are the statistical methods appropriate for the study design?		
	Q4. Is the analysis performed by intervention allocation status (i.e. intention to treat) rather than the actual intervention received?		
GLOBAL RATING	**Number of "weak" ratings**		
	Number of "moderate" ratings		
	Number of "strong" ratings		
	GLOBAL RATING		

Note. Adapted from Effective Public Health Practice Project (2003). Data extraction was completed using Excel and answers were selected from a data-validation drop-down list. See EPHPP (2003) for answer options.

Appendix D Quality Assessment Tool for Quantitative Studies Dictionary

Note. Copied from Effective Public Health Practice Project (2003).

The purpose of this dictionary is to describe items in the tool thereby assisting raters to score study quality. Due to under-reporting or lack of clarity in the primary study, raters will need to make judgements about the extent that bias may be present. When making judgements about each component, raters should form their opinion based upon information contained in the study rather than making inferences about what the authors intended.

A) SELECTION BIAS

(Q1) Participants are more likely to be representative of the target population if they are randomly selected from a comprehensive list of individuals in the target population (score very likely). They may not be representative if they are referred from a source (e.g. clinic) in a systematic manner (score somewhat likely) or self-referred (score not likely).

(Q2) Refers to the % of subjects in the control and intervention groups that agreed to participate in the study before they were assigned to intervention or control groups.

B) STUDY DESIGN

In this section, raters assess the likelihood of bias due to the allocation process in an experimental study. For observational studies, raters assess the extent that assessments of exposure and outcome are likely to be independent. Generally, the type of design is a good indicator of the extent of bias. In stronger designs, an equivalent control group is present and the allocation process is such that the investigators are unable to predict the sequence.

Randomized Controlled Trial (RCT)

An experimental design where investigators randomly allocate eligible people to an intervention or control group. A rater should describe a study as an RCT if the randomization sequence allows each study participant to have the same chance of receiving each intervention and the investigators could not predict which intervention was next. If the investigators do not describe the allocation process and only use the words ‘random’ or ‘randomly’, the study is described as a controlled clinical trial. See below for more details. *Was the study described as randomized?*

Score YES, if the authors used words such as random allocation, randomly assigned, and random assignment.

Score NO, if no mention of randomization is made.

Was the method of randomization described?

Score YES, if the authors describe any method used to generate a random allocation sequence.

Score NO, if the authors do not describe the allocation method or describe methods of allocation such as alternation, case record numbers, dates of birth, day of the week, and any allocation procedure that is entirely transparent before assignment, such as an open list of random numbers of assignments. If NO is scored, then the study is a controlled clinical trial.

Was the method appropriate?

Score YES, if the randomization sequence allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which intervention was next. Examples of appropriate approaches include assignment of subjects by a central office unaware of subject characteristics, or sequentially numbered, sealed, opaque envelopes.

Score NO, if the randomization sequence is open to the individuals responsible for recruiting and allocating participants or providing the intervention, since those individuals can influence the allocation process, either knowingly or unknowingly. If NO is scored, then the study is a controlled clinical trial.

Controlled Clinical Trial (CCT)

An experimental study design where the method of allocating study subjects to intervention or control groups is open to individuals responsible for recruiting subjects or providing the intervention. The method of allocation is transparent before assignment, e.g. an open list of random numbers or allocation by date of birth, etc.

Cohort analytic (two group pre and post)

An observational study design where groups are assembled according to whether or not exposure to the intervention has occurred. Exposure to the intervention is not under the control of the investigators. Study groups might be nonequivalent or not comparable on some feature that affects outcome.

Case control study

A retrospective study design where the investigators gather ‘cases’ of people who already have the outcome of interest and ‘controls’ who do not. Both groups are then questioned or their records examined about whether they received the intervention exposure of interest.

Cohort (one group pre + post (before and after)

The same group is pretested, given an intervention, and tested immediately after the intervention. The intervention group, by means of the pretest, act as their own control group.

Interrupted time series

A time series consists of multiple observations over time. Observations can be on the same units (e.g. individuals over time) or on different but similar units (e.g. student achievement scores for particular grade and school). Interrupted time series analysis requires knowing the specific point in the series when an intervention occurred.

C) CONFOUNDERS

By definition, a confounder is a variable that is associated with the intervention or exposure and causally related to the outcome of interest. Even in a robust study design, groups may not be balanced with respect to important variables prior to the intervention. The authors should indicate if confounders were controlled in the design (by stratification or matching) or in the analysis. If the allocation to intervention and control groups is randomized, the authors must report that the groups were balanced at baseline with respect to confounders (either in the text or a table).

D) BLINDING

(Q1) Assessors should be described as blinded to which participants were in the control and intervention groups. The purpose of blinding the outcome assessors (who might also be the care providers) is to protect against detection bias.

(Q2) Study participants should not be aware of (i.e. blinded to) the research question. The purpose of blinding the participants is to protect against reporting bias.

E) DATA COLLECTION METHODS

Tools for primary outcome measures must be described as reliable and valid. If ‘face’ validity or ‘content’ validity has been demonstrated, this is acceptable. Some sources

from which data may be collected are described below: Self reported data includes data that is collected from participants in the study (e.g. completing a questionnaire, survey, answering questions during an interview, etc.). Assessment/Screening includes objective data that is retrieved by the researchers. (e.g. observations by investigators). Medical Records/Vital Statistics refers to the types of formal records used for the extraction of the data. Reliability and validity can be reported in the study or in a separate study. For example, some standard assessment tools have known reliability and validity.

F) WITHDRAWALS AND DROP-OUTS

Score YES if the authors describe BOTH the numbers and reasons for withdrawals and drop-outs.

Score NO if either the numbers or reasons for withdrawals and drop-outs are not reported.

The percentage of participants completing the study refers to the % of subjects remaining in the study at the final data collection period in all groups (i.e. control and intervention groups).

G) INTERVENTION INTEGRITY

The number of participants receiving the intended intervention should be noted (consider both frequency and intensity). For example, the authors may have reported that at least 80 percent of the participants received the complete intervention. The authors should describe a method of measuring if the intervention was provided to all participants the same way.

As well, the authors should indicate if subjects received an unintended intervention that may have influenced the outcomes. For example, co-intervention occurs when the study group receives an additional intervention (other than that intended). In this case, it is possible that the effect of the intervention may be over-estimated.

Contamination refers to situations where the control group accidentally receives the study intervention. This could result in an under-estimation of the impact of the intervention.

H) ANALYSIS APPROPRIATE TO QUESTION

Was the quantitative analysis appropriate to the research question being asked? An intention-to-treat analysis is one in which all the participants in a trial are analyzed according to the intervention to which they were allocated, whether they received it or

not. Intention-to-treat analyses are favoured in assessments of effectiveness as they mirror the noncompliance and treatment changes that are likely to occur when the intervention is used in practice, and because of the risk of attrition bias when participants are excluded from the analysis.

Component Ratings of Study:

For each of the six components A – F, use the following descriptions as a roadmap.

A) SELECTION BIAS

Strong: The selected individuals are very likely to be representative of the target population (Q1 is 1) and there is greater than 80% participation (Q2 is 1).

Moderate: The selected individuals are at least somewhat likely to be representative of the target population (Q1 is 1 or 2); and there is 60 - 79% participation (Q2 is 2).

'Moderate' may also be assigned if Q1 is 1 or 2 and Q2 is 5 (can't tell).

Weak: The selected individuals are not likely to be representative of the target population (Q1 is 3); or there is less than 60% participation (Q2 is 3) or selection is not described (Q1 is 4); and the level of participation is not described (Q2 is 5).

B) DESIGN

Strong: will be assigned to those articles that described RCTs and CCTs.

Moderate: will be assigned to those that described a cohort analytic study, a case control study, a cohort design, or an interrupted time series.

Weak: will be assigned to those that used any other method or did not state the method used.

C) CONFOUNDERS

Strong: will be assigned to those articles that controlled for at least 80% of relevant confounders (Q1 is 2); or (Q2 is 1).

Moderate: will be given to those studies that controlled for 60 – 79% of relevant confounders (Q1 is 1) and (Q2 is 2).

Weak: will be assigned when less than 60% of relevant confounders were controlled (Q1 is 1) and (Q2 is 3) or control of confounders was not described (Q1 is 3) and (Q2 is 4).

D) BLINDING

Strong: The outcome assessor is not aware of the intervention status of participants (Q1 is 2); and the study participants are not aware of the research question (Q2 is 2).
Moderate: The outcome assessor is not aware of the intervention status of participants (Q1 is 2); or the study participants are not aware of the research question (Q2 is 2); or blinding is not described (Q1 is 3 and Q2 is 3).
Weak: The outcome assessor is aware of the intervention status of participants (Q1 is 1); and the study participants are aware of the research question (Q2 is 1).

E) DATA COLLECTION METHODS

Strong: The data collection tools have been shown to be valid (Q1 is 1); and the data collection tools have been shown to be reliable (Q2 is 1).
Moderate: The data collection tools have been shown to be valid (Q1 is 1); and the data collection tools have not been shown to be reliable (Q2 is 2) or reliability is not described (Q2 is 3).
Weak: The data collection tools have not been shown to be valid (Q1 is 2) or both reliability and validity are not described (Q1 is 3 and Q2 is 3).

F) WITHDRAWALS AND DROP-OUTS

Strong: will be assigned when the follow-up rate is 80% or greater (Q2 is 1).
Moderate: will be assigned when the follow-up rate is 60 – 79% (Q2 is 2) OR Q2 is 5 (N/A). **Weak**: will be assigned when a follow-up rate is less than 60% (Q2 is 3) or if the withdrawals and drop-outs were not described (Q2 is 4).

Appendix E Risk of bias assessment: summary

Ref. No.	Selection bias	Study design	Confoun-ders	Blinding	Data collection	Withdrawal /dropout	Global Rating
1. Bhide (2017)	Weak	Strong	Strong	Weak	Weak	Weak	Weak
2. Çakmak & Erçetin (2018)	Weak	Strong	Strong	Weak	Strong	Strong	Weak
3. Cavus & Ibrahim (2017)	Weak	Strong	Strong	Weak	Weak	Strong	Weak
4. C.-M. Chen, Chen, et al. (2019)	Strong	Strong	Strong	Weak	Weak	Strong	Weak
5. C.-M. Chen, Liu, et al. (2019)	Weak	Strong	Strong	Weak	Strong	Strong	Weak
6. de Jong et al. (2010)	Weak	Strong	Strong	Moderate	Weak	Strong	Weak
7. García-Mencía et al. (2016)	Weak	Moderate	Weak	N/A	Weak	Strong	Weak
8. Guaqueta & Castro-Garces (2018)	Weak	Moderate	Weak	N/A	Weak	Strong	Weak
9. Hao et al. (2019)	Weak	Weak	Strong	N/A	Strong	Strong	Weak
10. Hsu et al. (2013)	Moderate	Strong	Moderate	Moderate	Strong	Strong	Moderate
11. Liakin et al. (2015)	Moderate	Strong	Strong	Strong	Strong	Strong	Strong
12. Liu (2016)	Weak	Strong	Weak	Weak	Strong	Strong	Weak
13. Loewen et al. (2019)	Weak	Moderate	Strong	N/A	Strong	Moderate	Moderate
14. Martinelli (2016)	Weak	Moderate	Moderate	N/A	Weak	Strong	Weak
15. Mohamad Ali et al. (2016)	Strong	Strong	Strong	Weak	Moderate	Strong	Moderate
16. Moreno & Gatewood (2019)	Moderate	Strong	Strong	Weak	Strong	Strong	Moderate
17. Nicholes, J. (2016)	Strong	Strong	Weak	Weak	Weak	Strong	Weak
18. Ou-Yang & Wu (2017)	Strong	Strong	Weak	Weak	Strong	Strong	Weak
19. Purgina et al. (2020)	Moderate	Strong	Strong	Weak	Weak	Strong	Weak
20. Rachels & Rockinson-Szapkiw (2018)	Moderate	Strong	Strong	Moderate	Strong	Strong	Strong
21. Ramírez et al. (2018)	Moderate	Strong	Moderate	Weak	Moderate	Strong	Moderate
22. Terantino (2016)	Weak	Moderate	Weak	Moderate	Weak	Strong	Weak
23. Thongsri et al. (2019)	Weak	Strong	Moderate	Weak	Strong	Strong	Weak

24. Vesselinov & Grego (2015)	Strong	Moderate	Strong	N/A	Strong	Weak	Moderate
25. Vesselinov & Grego (2019b)	Strong	Moderate	Strong	N/A	Strong	Moderate	Moderate
26. Vesselinov & Grego (2019a)	Strong	Moderate	Strong	N/A	Strong	Moderate	Moderate
27. Vesselinov & Grego (2016a)	Strong	Moderate	Strong	N/A	Strong	Strong	Strong
28. Vesselinov & Grego (2012)	Strong	Moderate	Strong	N/A	Strong	Weak	Moderate
29. Vesselinov & Grego (2016b)	Strong	Moderate	Strong	N/A	Strong	Moderate	Moderate
30. Vesselinov et al. (2019)	Strong	Moderate	Strong	N/A	Strong	Strong	Strong
31. Wu (2015b)	Strong	Strong	Strong	Weak	Weak	Strong	Weak
32. Wu (2015a)	Moderate	Strong	Strong	Weak	Weak	Strong	Weak
Number of "weak"	**13**	1	6	**15**	12	3	18
Number of "moderate"	7	12	4	4	2	4	10
Number of "strong"	12	**19**	**22**	1	**18**	**25**	4

Appendix F Data extraction sheet and risk of bias assessment for individual studies

1.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	18/07/2020
	Reference	Bhide, A. (2017). Instructional methods for promoting the development of orthographic and phonological knowledge in second language learners of Indic languages [University of Pittsburgh]. In *ProQuest Dissertations and Theses* (Issue 10645799). https://search.proquest.com/docview/1943998636?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Doctoral thesis
Setting	Time of study	April, 2017 (published)
	Duration of intervention	4 weeks
	Location of the study	At a large all-girl private school in Bangalore, Karnakata, India
Methods	Research question(s)	Does a mobile game increase akshara knowledge? More specifically, does it teach student to recognive complex akshara in isolation and in word contexts, orthographically and phonologically? (inferred)
	Study design	An experimental study with 2 experimental groups (with different interventions) and 1 unseen control group. A matched DID design. "The three groups were selected to match as closely as possible on spelling pre-test scores, with no significant differences on the other pre-test measures."
	Intervention	Researcher developed. An unnamed mobile game with two versions, narrow spacing and wide spacing. "[it] was programmed using e-Chimera, a visual end-use programming environment that can design experiments for mobile devices." Spacing refers to the manner in which problems about the same akshara are distributed.
		Android
	Comparator	Unseen control group. No information was given. Inferred: zero intervention.
	App version comparison	Yes
	Studied language	Hindi akshara
	Materials	Two types of problems about aksahara: 1. akshara decomposition (orthographic), 2. spelling (orthographic + phonological knowledge). "All of the words were chosen from either CBSE or ICSE 3rd-5th standard textbooks." CBSE and ICSE are two popular curricula in India. These materials focus on teaching the learners orthographic and phonological knowledge of akshara.
Participant	Source of recruitment	A co-author reached out to the school

	Demographics	"The participants were all in the 4th standard at a large all-girls private school in Bangalore, Karnataka, India." They are speak Hindi as an additional language and not at home (natively).
	Sample size	103
		Narrow spacing: 32, Wide spacing: 35
		36
	Attrition	108
		103
		4.63%
	Pre-test or specified proficiency	There were five language-related tests administered to the participants before and after the intervention: akshara (simple, CV, complex) recognition, word reading, spelling, word construction, and vocabulary.
Outcomes	Language skills	Vocabulary, Orthography, Phonology
	Outcome measurement	See above for the test type. "**Akshara recognition:** Children were presented with 20 akshara (9 simple, 6 CV, 5 complex) and were asked to read them aloud as quickly as possible." "**Word reading:** The children were asked to read 48 words, presented in 6 lists of 8 words each" (the lists differ in difficulty). "**Spelling**: The children were asked to spell 30 words" of six types upon hearing the word. "**Akshara construction:** The akshara construction assessment measured students' knowledge of ligaturing rules and their ability to apply those rules in novel contexts" (made up akshara). "**Vocabulary:** The children were asked to define 24 Hindi words." It is implied that the pretests and the posttests are identical and both contain materials that are used in the experimental interventions.
	Outcomes (specific scores)	Akshara recognition (in %, in the order of simple, CV, Complex) Narrow Pre-test 94.44 (9.78) 46.88 (21.77) 36.25 28.93) Post-test 98.96 (3.29) 57.29 (20.71) 53.13 (29.89) Wide Pre-test 94.29 (15.33) 55.71 (22.49) 34.29 (27.26) Post-test 98.69 (3.63) 59.8 (21.37) 51.18 (27.05) Word reading (out of 8, in the order of simple, CV, learned, near, medium, far) Narrow Pre-test 5.58 (1.46) 4.18 (2.08) 3.42 (1.49) 2.23 (1.61) 3.85 (2.18) 2.81 (1.77) Post-test Wide Pre-test 5.51 (1.63) 4.09 (1.94) 3.46 (1.59) 2.43 (1.88) 3.70 (2.23) 3.17 (1.83) Post-test

6.16 (1.49)	6.17 (1.52)
5.48 (1.71)	4.99 (1.99)
4.85 (1.52)	5.26 (1.15)
3.15 (1.84)	3.49 (1.86)
4.10 (2.50)	4.63 (1.79)
3.65 (1.86)	3.93 (1.83)
Spelling test (out of 6, in the order of simple, CV, learned, near medium, far)	
Narrow	Wide
Pre-test	Pre-test
3.59 (1.48)	3.34 (1.37)
1.34 (1.46)	1.36 (1.44)
0.61 (0.92)	0.77 (0.92)
0.44 (0.82)	0.36 (0.71)
0.33 (0.69)	0.37 (0.81)
0 (0)	0 (0)
Post-test	Post-test
4.03 (1.36)	4.11 (0.83)
2.14 (1.52)	2.24 (1.26)
1.11 (1.01)	1.60 (1.37)
0.70 (0.85)	0.71 (1.09)
1.03 (1.26)	0.81 (1.14)
0.17 (0.49)	0.20 (0.62)
Akshara construction (out of 7)	
Narrow	Wide
Pre-test	Pre-test
4.97 (1.73)	5.06 (1.82)
Post-test	Post-test
6.06 (1.16)	6.09 (1.26)
Vocabulary (out of 24, in order of “unrelated to game,” “morphologically related to words in game,” and “learned in game”)	
Narrow	Wide
Pre-test	Pre-test
14.00 (5.83)	14.40 (6.36)
7.75 (5.41)	8.83 (6.06)
5.03 (4.59)	6.03 (5.81)
Post-test	Post-test
15.34 (5.71)	16.57 (5.39)
8.75 (6.54)	9.54 (6.38)
6.50 (5.56)	7.66 (6.01)
Control	
Akshara recognition	
Pre-test	
95.37 (9.71)	
46.3 (21.50)	
31.11 (28.86)	
Post-test	
95.68 (9.31)	
50.93 (24.86)	
35.56 (26.67)	
Word reading	
Pre-test	

	5.06 (1.74) 3.89 (1.96) 3.14 (1.53) 2.11 (1.64) 3.08(2.46) 2.93 (2.00) Post-test 6.35 (1.66) 4.96 (2.04) 4.18 (2.01) 2.85 (1.74) 4.07 (2.13) 3.17 (1.86) Spelling test (out of 6) Pre-test 3.25 (1.34) 1.13 (1.54) 0.63 (0.81) 0.50 (0.73) 0.33 (0.76) 0.06 (0.20) Post-test 4.19 (0.98) 1.94 (1.40) 0.94 (0.99) 0.57 (0.90) 0.65 (1.02) 0.11 (0.34) Akshara construction (out of 7) Pre-test 4.89 (2.03) Post-test 6.09 (1.22) Vocabulary Pre-test 15.06 (5.66) 7.78 (5.00) 6.36 (4.98) Post-test 16.61 (5.93) 9.61 (5.84) 7.47 (4.63)
Effect size	Not enough information to calculate.
Conclusion (researcher)	Overall, the game improved students' akshara knowledge. In terms of pre/post-tests, the benefit of the game was most apparent on the akshara recognition assessment. There were also gains on the reading and spelling tasks, although they were not as robust as those with akshara recognition.
Conclusion (general)	Positive
Risk of Bias Assessment (EPHPP Adaptation)	

Component	Questions	Answer
A. Selection bias	A1. Are the individuals selected to participate in the study likely to be representative of the target population?	Not Likely
	A2. What percentage of selected individuals agreed to participate?	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1. What is the study design?	Non-randomised comparison
	B2. Was the study described as randomised?	No
	B2a. If Yes, was the method of randomization described?	Not Applicable
	B2b. If Yes, was the method appropriate?	Not Applicable
	Component Rating	Strong
C. Confounders	C1. Were there important differences between groups prior to the intervention?	No
	C2. If yes, what percentage of relevant confounders was controlled?	Not applicable
	Component Rating	Strong
D. Blinding (as a part of controlled trial)	D1. Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?	Can't tell
	D2. Were the study participants aware of the research question?	Can't tell
	Component Rating	Weak
E. Data collection practices	E1. Were data collection tools shown to be valid?	No
	E2. Were data collection tools shown to be reliable?	No
	Component Rating	Weak
F. Withdrawals, dropouts	F1. Were withdrawals and drop-outs reported in terms of numbers and/or reasons per group?	No
	what is the percentage of participants completing the study?	80-100%
	Component Rating	Weak

G. Invention integrity	G1. What percentage of participants received the allocated intervention or exposure of interest?	80-100%
	G2. Was the consistency of the intervention measured?	No
	G3. Is it likely that subjects received an unintended intervention (contamination or co-intervention) that may influence the results?	Yes
H. Analysis	H1. What is the unit of allocation?	learner
	H2. What is the unit of analysis?	learner
	H3. Are the statistical methods appropriate for the study design?	Yes
	H4. Is the analysis performed by intervention allocation status (i.e. intention to treat) rather than the actual intervention received?	Yes
GLOBAL RATING	**Number of "weak" ratings**	4
	Number of "moderate" ratings	0
	Number of "strong" ratings	2
	GLOBAL RATING	**WEAK**

2.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	10/07/2020
	Reference	Çakmak, F., & Erçetin, G. (2018). Effects of Gloss Type on Text Recall and Incidental Vocabulary Learning in Mobile-Assisted L2 Listening. *ReCALL*, *30*(1), 24–47.
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2014
	Duration of intervention	80 minutes
	Location of the study	At a state university in Turkey
Methods	Research question(s)	1. Does access to glosses facilitate text recall and incidental vocabulary learning when learners are engaged in listening to a story through mobile phones? If yes, are there differences between single-mode (i.e., textual-only, pictorial-only) and dual-mode (i.e., textual-plus-pictorial) glosses? 2. Are there differences between single-mode glosses (i.e., textual-only, pictorial-only) and dual-mode (i.e., textual-plus-pictorial) glosses in terms of frequency of access to glosses and time spent on task? 3. Are frequency of access to glosses and time spent on task related to text recall and incidental vocabulary learning in single-mode glosses (i.e., textual-only, pictorial- only) and dual-mode (i.e., textual-plus-pictorial) glosses?
	Study design	A quantitative, experimental study with four conditions and participants randomly assigned to their condition.
	Intervention	Researcher developed. (2) textual-only gloss group, (3) pictorial-only gloss group, and (4) textual-plus-pictorial gloss group. A mobile-assisted listening application was developed and optimized for Samsung Galaxy Mini devices. The application connected to a web service which was developed with PHP language, MySQL database and JSON data interchange standard, and downloaded the experimental materials. With the help of a web control panel written with HTML and PHP languages, the researcher could change the system settings, activate the conditions before the experi- ment started and terminate them when the experiment was over, and download the parti- cipants' data as spreadsheets including their names, the treatment condition, total time spent on task, and glosses referenced.

		Android: Samsung Galaxy Mini GT-S5570 mobile device
	Comparator	Control group with no gloss
	App version comparison	Yes
	Studied language	English
	Materials	"The listening text involved an audio file that was a 13.56-minute- long story taken from the website of Voice of America, an official American broadcast for non-native speakers of English."
Participants	Source of recruitment	Recruited through enrolled students at the university. One of the researchers was the students' English instructor.
	Demographics	The participants were registered freshmen students studying Public Administration, Management, and Economics in the Faculty ofManagement and Economics.
	Sample size	88
		22
		22 each group
	Attrition	88
		88
		0.00%
	Pre-test or specified proficiency	elementary level of proficiency. The participants had completed one year of English Language Preparatory Program prior to the treatment
Outcomes	Language skills	Vocabulary, Listening
	Outcome measurement	1. Free recall task, 2. vocabulary tests (form recognition, L2 meaning production and L1 meaning production)
	Outcomes (specific scores)	**Form recognition** Textual-only gloss 18.36 (4.04) Pictorial-only 17.59 (4.8) Textual-plus-pictorial gloss 16.36 (5.57) **L1 meaning production** Textual-only gloss 17.09 (3.73) Pictorial-only 15 (3.37) Textual-plus-pictorial gloss 17 (5.53) **L2 meaning production** Textual-only gloss 13.86 (4.22) Pictorial-only 11.5 (3.5) Textual-plus-pictorial gloss 13.41 (4.84) **No gloss** Form recognition 8.82 (3.35) L1 meaning production 12.77 (4.09) L2 meaning production 10.18 (4.03)
	Effect size	Multivariate: $\eta 2 = 0.195$. Univariate form recognition: $\eta 2 = 0.43$;

	L1 meaning production: partial η2 = 0.15; L2 meaning production: partial η2 = 0.12.
Conclusion (researcher)	The results indicate that the participants in the no-gloss condition did not have a significantly lower performance in terms of text recall compared to those in the gloss conditions, suggesting that the facilitative effect of access to glosses was not confirmed in terms of text recall. However, the hypothesis was confirmed in terms of vocabulary learning since the participants in the gloss conditions had significantly higher means than those in the control group.
Conclusion (general)	No difference between versions + positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Can't tell
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes

GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	0
	Number of "strong" ratings	4
	GLOBAL RATING	**WEAK**

3.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	11/07/2020
	Reference	Cavus, N., & Ibrahim, D. (2017). Learning English using children's stories in mobile devices. *British Journal of Educational Technology*, *48*(2), 625–641.
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	April, 2014
	Duration of intervention	0.5 hours/day for 4 weeks
	Location of the study	A government college in Cyprus
Methods	Research question(s)	1. Are there any significant differences between the Pre-Test and Post-Test results of the experimental group students? 2. Are there any significant differences between the Pre-Test and Post-Test results of the control group students? 3. Are there any significant differences between the Post-Test results of the experimental group and control group students, and if so, how can the implications of these findings improve practice?
		Method
	Study design	An experimental study with an experimental group and a control group; didn't specify if randomised but pre-tested and matched.
	Intervention	Researcher developed. Near East University Childrens' Story Teller (NEU-CST). This application was designed to improve the vocabulary, comprehension, pronunciation and listening skills of young students (age group 12–13) learning English as a second language using interactive children's stories.
		Android
	Comparator	Children's story book
	App version comparison	No
	Studied language	English
	Materials	A children's story book content for both
Participants	Source of recruitment	Not specified
	Demographics	Aged 12, first year students at the college
	Sample size	37

		19
		18
	Attrition	37
		37
		0.00%
	Pre-test or specified proficiency	No significant (p>.05 for all tests) differences between the two groups. **Vocabulary** Pre-Test_Experimental 22.42 (10.90) Pre-Test_Control 23.50 (11.11) Mean difference: -1.08 **Comprehension** Pre-Test_Experimental 27.42 (9.54) Pre-Test_Control 21.78 (9.57) Mean difference: 5.64 **Pronunciation** Pre-Test_Experimental 21.16 (11.53) Pre-Test_Control 23.83 (10.19) Mean difference: -2.68 **Listening** Pre-Test_Experimental 26.16 (10.86) Pre-Test_Control 24.44 (8.91) Mean difference: 1.71
Outcomes	Language skills	Vocabulary, Comprehension, Phonology, Listening
	Outcome measurement	**Comparator (lab 2):** Students were asked to put on their headphones and the story was played to them for 3 minutes. Then, 10 oral questions were asked to these students to test their comprehension skills. The listening skills were tested by asking them five questions related with the story. To test their pronunciation skills, 10 words used in the story were shown to them and they were asked to give the correct pronunciation of these words. The vocabulary skills were tested by showing them pictures of 10 objects used in the story and asking them to write down thenames of theseobjects. **The implementation for the experimental group** (also in lab) was similar to the one in lab 2, but here the story and all the ques- tions were presented to the participants on the developed application (Figures 3 and 4). Also, the results were evaluated automatically by the developed application.
	Outcomes (specific scores)	**Experimental** Pre-test / Post-test **Vocabulary** 22.42 (10.90)* / 76.53 (7.18)* **Comprehension** 27.42 (9.54)* / 78.84 (7.58)* **Pronunciation** 21.16 (11.53)* / 79.74 (8.03)* **Listening** 26.16 (10.86)* / 78.42 (7.07)* **Control** Pre-test / Post-test **Vocab** 23.50 (11.11)* / 51.28 (11.35)* **Comp.** 21.78 (9.57)* / 56.22 (8.60)* **Pron.** 23.83 (10.19)* / 55.67 (12.56)*

	List.	24.44 (8.91)*	71.61 (11.68)*

Effect size	Not stated by author. Vocabulary: d=2.66 (p<.001) Comprehension: d=2.79 (p<.001) Pronunciation: d=2.28 (p<.001) Listening: d=0.71 (p<.05)
Conclusion (researcher)	The results of the experimental study clearly indicated that the developed mobile application has helped students while learning English as a second language. English learning skills such as vocabulary, pronunciation, listening and comprehension have improved in both groups. However, these skills of the experimental group showed much bigger improvement as a result of using the developed application.
Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Can't tell
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Non-randomised comparison
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Can't tell
H. Analysis	H1	learner
	H2	learner

	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	3
	Number of "moderate" ratings	0
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**

4.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	11/07/2020
	Reference	Chen, C.-M., Chen, L.-C., & Yang, S.-M. (2019). An English vocabulary learning app with self-regulated learning mechanism to improve learning performance and motivation. *Computer Assisted Language Learning*, *32*(3), 237–260. https://search.proquest.com/docview/2213927869?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2019 (published)
	Duration of intervention	2 weeks
	Location of the study	Taoyuan city, Taiwan
Methods	Research question(s)	(1) Does the developed EVLAPP-SRLM (English vocabulary learning app with a self-regulated learning mechanism) significantly improve the English vocabulary learning per- formance and motivation of learners? (2) Does the developed EVLAPP- SRLM provide significantly different benefits to learners of different gen- ders with respect to their English vocabulary performance and motivation? (3) Does the developed EVLAPP-SRLM provide significantly different ben- efits to learners with different cognitive styles with respect to their English vocabulary learning performance and motivation?
	Study design	quasi-experimental design with an experimental group and a control group. There were a pretest and a posttest.
	Intervention	Researcher developed. EVLAPP-SRLM (English vocabulary learning app with a self-regulated learning mechanism): VOC 4 FUN. The app that was developed in this work supports English vocabulary learning and a self-learning management mechanism to help learners learn English vocabulary. Modules include: SRL setting, English vocabulary learning, quiz, note, and goal reminder.
		Not specified
	Comparator	EVLAPP-NSRLM (English vocabulary learning app without a self-regulated learning mechanism)
	App version comparison	Yes
	Studied language	English

	Materials	800 English words published by the Ministry of Education, Taiwan, for the purpose of English vocabulary learning by elementary and junior high school students
Participants	Source of recruitment	Not specified
	Demographics	Elementary school students aged 10-11 years old
	Sample size	46
		21
		25
	Attrition	46
		46
		0.00%
	Pre-test or specified proficiency	A vocabulary test. Experimental group: 40 (29.07) Control group: 46.08 (30.44)
Outcomes	Language skills	Vocabulary
	Outcome measurement	Vocabulary ability assessment
	Outcomes (specific scores)	Vocabulary ability assessment Experimental group Pre 40 (29.07) Post 68.19 (22.82) Control group Pre 46.08 (30.44) Post 58.88 (27.36)
	Effect size	d=0.37
	Conclusion (researcher)	This result indicates that the learners who used the proposed EVLAPP-SRLM exhibited better learning performance than those who used EVLAPP-NSRLM.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
	C1	No

C. Confounders	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Yes
	Component Rating	Weak
E. Data collection practices	E1	No
	E2	No
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Yes
H. Analysis	H1	class/group
	H2	class/group
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	0
	Number of "strong" ratings	4
	GLOBAL RATING	**WEAK**

5.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	18/07/2020
	Reference	Chen, C.-M., Liu, H., & Huang, H.-B. (2019). Effects of a mobile game-based English vocabulary learning app on learners' perceptions and learning performance: A case study of Taiwanese EFL learners. *ReCALL : The Journal of EUROCALL*, *31*(2), 170–188. http://search.ebscohost.com/login.aspx?direct=true&db=bri&AN=135758218&site=ehost-live&authtype=ip,uid
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2019 (published)
	Duration of intervention	4 weeks
	Location of the study	College of Liberal Arts at National Chengchi University (NCCU)
Methods	Research question(s)	1. whether significant differences exist in vocabulary acquisition, vocabulary retention, and learners' perceptions between the experimental group (MEVLA-GF) and the control group (MEVLA-NGF), 2. whether learners' involvement in gamified functions is correlated with vocabulary learning performance.
	Study design	A mixed methods study with an experiment comprising of an experimental group and a control group, followed by questionnaires and interviews.
	Intervention	Researcher developed. MEVLA-GF: PHONE Words app; the app simultaneously provides traditional assessment and gamified assessment with competition mechanism that can be optionally selected to support learning English vocabulary. The six main functions provided by the PHONE Words app are the word list, customizedword list, pre-established learning path, traditional assessment, gamified assessment, and ranking among learning peers.
		Not specified
	Comparator	MEVLA-NGF: PHONE words app without the gamified assessment with ranking among learning peers, but the other functions provided by the two apps are the same.
	App version comparison	Yes
	Studied language	English
	Materials	English vocabulary on such formal vocabulary tests as TOEIC and TOEFL
Participants	Source of recruitment	this study issued a call via the Internet for volunteers who were willing to pass the TOEIC test

	Demographics	Taiwanese sophomores studying at the College of Liberal Arts at National Chengchi University (NCCU).
	Sample size	20
		10
		10
	Attrition	20
		20
		0.00%
	Pre-test or specified proficiency	14.7
Outcomes	Language skills	Vocabulary
	Outcome measurement	Questions adapted from TOEIC, identical to pretest (adjusted order)
	Outcomes (specific scores)	Rank-based vocabulary test adopted from TOEIC Experimental Pre-test 14.7 Post-test 48.9 Delayed post-test 41.75 Control Pre-test 14.7 Post-test 39.3 Delayed post-test 23.65
	Effect size	d=0.86
	Conclusion (researcher)	This study confirmed that the learners using MEVLA-GF, the experimental group, as an assistive tool markedly outperformed learners using MEVLA-NGF, the control group, in terms of vocabulary acquisition and vocabulary retention.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Non-randomised comparison
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Yes
	Component Rating	Weak
	E1	Yes

E. Data collection practices	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Yes
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	0
	Number of "strong" ratings	4
	GLOBAL RATING	**WEAK**

6.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	18/07/2020
	Reference	de Jong, T., Specht, M., & Koper, R. (2010). A study of contextualised mobile information delivery for language learning. *Educational Technology & Society*, *13*(3), 110–125. http://search.ebscohost.com/login.aspx?direct=true&db=eax&AN=508131938&site=ehost-live&authtype=ip,uid
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2010 (published)
	Duration of intervention	30 minutes
	Location of the study	Not specified
Methods	Research question(s)	1. whether there are differences between the efficiency of learning support provided by object-based and location-based information delivery. 2. if so, are there any circumstances in which either of these granularities proves more efficient. Hypothesis 1: learners using an object filter (SOF, NOF, LOF) will have a higher knowledge gain (KG) than those using a room filter. Hypothesis 2: learners using a selection method that requires fewer actions (SRF, SOF, LORF) to access content will have a higher knowledge gain (KG) than those requiring more actions.
	Study design	a between-groups design with seven treatment groups and no unseen control.
	Intervention	Researcher developed. A mobile phrase book for learning Hindi that uses contextual information to filter the learning content. The phrase book contained learning content consisting of a picture of an object, a textual representation of the Hindi word for the object, and an audio fragment for the word created by a native speaker. Moreover, the learner could view an enlarged version of the picture with a higher level of detail. For each of the treatments in table 1 another variation of the mobile language learning software was developed. The software was developed in PHP and the learning content was adapted to be rendered on small screens.
		iOS (iPhone 3G)
	Comparator	None
	App version comparison	Yes

	Studied language	Hindi				
	Materials	A Hindi phrase book				
Participants	Source of recruitment	Not specified				
	Demographics	No obvious shared charateristics, possibly place of residence in the Netherlands. Most of the participants spoke Dutch as their native language (n = 26), however some spoke German (n = 6), Chinese/Cantonese (n = 1), Tamil (n = 1), and Spanish (n = 1). Two participants had some knowledge of Hindi.				
	Sample size	35				
		5 each				
		N/A				
	Attrition	35				
		35				
		0.00%				
	Pre-test or specified proficiency	Pre-test was given to participants who presumably didn't have any knowledge of Hindi (except for 2).				
Outcomes	Language skills	Vocabulary				
	Outcome measurement	A vocabulary test (asking for definition) identical to the pretest				
	Outcomes (specific scores)		Semacode-based	Number-based	List-based	Location-based
		Room filter	0.35 (0.24)	0.38 (0.20)	0.47 (0.04)	0.62 (0.13)
		Object filter	0.22 (0.12)	0.37 (0.14)	0.35 (0.18)	
	Effect size	Treament: r=0.79, Context filter: r=0.42, Selection method: r=0.69				
	Conclusion (researcher)	Furthermore, the results show that all participants had similar scores on the pre-test, and self-evaluated their language abilities similarly. Therefore, it can be safely assumed that the participant's language expertise was evenly distributed over the treatment groups and any differences measured were caused by the experimental manipulations.				
	Conclusion (general)	Differences between versions				

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Randomized controlled trial
	B2	Yes

	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	No
	Component Rating	Moderate
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Can't tell
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	1
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**

7.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	11/07/2020
	Reference	García-Mencía, R., López-López, A., & Muñoz Meléndez, A. (2016). An Audio-Lexicon Spanish-Nahuatl: Using Technology to Promote and Disseminate a Native Mexican Language. In S. Papadima-Sophocleous, L. Bradley, & S. Thouësny (Eds.), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 155–159). Research-publishing.net. https://search.proquest.com/docview/1895981358?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Conference proceedings
Setting	Time of study	2016 (published)
	Duration of intervention	1 week
	Location of the study	Not specified
Methods	Research question(s)	how useful is the ALEN app for reinforcing and acquiring vocabulary of the Nahuatl language
	Study design	A single group pretest/posttest pre-experimental design.
	Intervention	Researcher developed. The ALEN application has a simple interface, which allows the user to enter text in Nahuatl or Spanish. There are two buttons, one used to find the written word in the text box and the other to reproduce the pronunciation of the word when found. At the center of the screen, an image illustrating the concept related to the word is displayed. The application was created using Android Studio, the official IDE for mobile application development based on IntelliJ IDEA4. Overall, the ALEN applicationruns locally and once installed, it does not create files or require access to services and resources such as the phone network, WiFi or Bluetooth connectivity.
		Android
	Comparator	None
	App version comparison	No
	Studied language	Nahuatl
	Materials	132 Nahuatl words
Participants	Source of recruitment	Not specified
	Demographics	ages 29.6 ± 7.3 years. All the volunteers were native Spanish speakers
	Sample size	11
		11
		0

	Attrition	11
		11
		0.00%
	Pre-test or specified proficiency	An examination of knowledge of Nahuatl comprised 15 questions. The first ten were aimed to achieve word pair association; in the first column Nahuatl words were listed, and in a second column Spanish words were presented, sorted in a different order. The latter five questions were intended to assess listening skills. Each question included an audio recording of a native Nahuatl speaker, and a window for writing the word heard by the user. Volunteers received one point for each question properly answered and there were no time constraints to complete the examination.
Outcomes	Language skills	Vocabulary, Listening
	Outcome measurement	Same test as pre-test, but the words were randomly generated
	Outcomes (specific scores)	A researcher-designed online examination on vocabulary and listening Pre-test Volunteer 1: 2 Volunteer 2: 0 Volunteer 3: 5 Volunteer 4: 1 Volunteer 5: 6 Volunteer 6: 5 Volunteer 7: 0 Volunteer 8: 2 Volunteer 9: 0 Volunteer 10: 3 Volunteer 11: 5 Mean: 2.64 (2.29) Post-test Volunteer 1: 4 Volunteer 2: 10 Volunteer 3: 5 Volunteer 4: 8 Volunteer 5: 9 Volunteer 6: 7 Volunteer 7: 6 Volunteer 8: 4 Volunteer 9: 2 Volunteer 10: 2 Volunteer 11: 6 Mean: 5.73 (2.65)
	Effect size	Not stated by author. d=1.11
	Conclusion (researcher)	Most users achieved clear improvement in their Nahuatl vocabulary after one week using the mobile application. In effect, 82% of volunteers improved their knowledge of Nahuatl.
	Conclusion (general)	Positive
Risk of Bias Assessment (EPHPP Adaptation)		

Component	Questions	Answer
A. Selection bias	A1	Can't tell
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Can't tell
	Component Rating	Weak
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Can't tell
	G3	Can't tell
H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	3
	Number of "moderate" ratings	1
	Number of "strong" ratings	1
	GLOBAL RATING	**WEAK**

8.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	11/07/2020
	Reference	Guaqueta, C. A., & Castro-Garces, A. Y. (2018). The Use of Language Learning Apps as a Didactic Tool for EFL Vocabulary Building. *English Language Teaching*, *11*(2), 61–71. https://search.proquest.com/docview/2013524764?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2018 (published)
	Duration of intervention	6 months
	Location of the study	Roncesvalles (Tolima), Colombia
Methods	Research question(s)	investigating the effects of using language learning apps (LLAs) as a didactic tool to foster vocabulary building in an EFL context.
	Study design	a mixed-methods approach with a concurrent design, single group pretest/posttest
	Intervention	Commercially available. Duolingo
		Android, iOS
	Comparator	None
	App version comparison	No
	Studied language	English
	Materials	Eight chosen topics on Duolingo
Participants	Source of recruitment	Not specified
	Demographics	10th grade students age 14 to 17 years old at a rural school in Tolima
	Sample size	20
		20
		0
	Attrition	20
		20
		0.00%
	Pre-test or specified proficiency	initial diagnosis test to assess the level of vocabulary knowledge: 36.7 (13.4). The test included 10 sections for students to identify vocabulary in context. Students were asked to answer two reading comprehension exercises, with multiple choice of vocabulary; match word and image; answer true/false; fill in the blanks; circle the correct word; and match word and definition. The vocabulary included in the test was part of what would be learned through the apps.

	Language skills	Vocabulary
Outcomes	Outcome measurement	Same test as pretest
	Outcomes (specific scores)	**A vocabulary test: identify vocabulary in context and the full mark is 100.** Pre-test / Post-test F1: 32 / F1: 62 F2: 36 / F2: 80 F3: 58 / F3: 89 F4: 45 / F4: 67 F5: 46 / F5: 78 F6: 18 / F6: 65 F7: 22 / F7: 72 F8: 24 / F8: 59 F9: 42 / F9: 72 F10: 29 / F10: 82 F11: 34 / F11: 76 F12: 28 / F12: 80 F13: 65 / F13: 92 M1: 53 / M1: 71 M2: 31 / M2: 65 M3: 16 / M3: 32 M4: 47 / M4: 84 M5: 26 / M5: 63 M6: 48 / M6: 86 M7: 34 / M7: 94 Mean: 36.7 (13.4)* / Mean: 73.45 (14.19)*
	Effect size	**Two four-part unit quizzes. Each participant had two units they were most unfamiliar with but equivalent in difficulty level. Sub-scores on reading, listening, speaking, and writing were reported but only total scores on each unit are extracted here.** **Unit 1 (out of 14)** / **Unit 2 (out of 14)** S1: 14 / S1: 13 S2: 14 / S2: 14 S3: 12 / S3: 14 S4: 14 / S4: 12 S5: 14 / S5: 14 S6: 14 / S6: 14 S7: 13 / S7: 14 S8: 12 / S8: 13 S9: 12 / S9: 12 S10: 13 / S10: 12 Mean: 12.8 (1.23) / Mean: 13.2 (0.92) Total Mean: 13 (1.08)
	Conclusion (researcher)	The outcomes presented in the FDT allowed us to infer that the means adopted to help students improve their vocabulary was successful for most of them who took this as a learning opportunity.
	Conclusion (general)	Positive
Risk of Bias Assessment (EPHPP Adaptation)		

Component	Questions	Answer
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Can't tell
	Component Rating	Weak
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	Can't tell
H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	3
	Number of "moderate" ratings	1
	Number of "strong" ratings	1
	GLOBAL RATING	**WEAK**

9.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	11/07/2020
	Reference	Hao, Y., Lee, K. S., Chen, S.-T., & Sim, S. C. (2019). An evaluative study of a mobile application for middle school students struggling with English vocabulary learning. *Computers in Human Behavior*, *95*, 208–216. http://search.ebscohost.com/login.aspx?direct=true&db=eax&AN=135257557&site=ehost-live&authtype=ip,uid
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2019 (published)
	Duration of intervention	40 minutes
	Location of the study	Northern Taiwan
Methods	Research question(s)	1. What are the learning outcomes of the junior high school lowachieving students who used the APP to learn English vocabulary? 2. What is the attitude of the students toward MAPP? 3. What are the students' perceptions of MAPP design?
	Study design	A mixed methods study with a single group pre-test/post-test design
	Intervention	Researcher developed. Detective ABC: This app is a series of mission-oriented puzzle-solving game stories designed with different content levels. Through the stories, players are provided with authentic situations. The main presentation method uses colour pictures with Chinese characteristics and English letters so that students can clearly select the unit of which they want to be challenged, thus increasing their attention. There are 11 learning units, and each unit has 4 vocabularies. Each unit has different story contexts, including emotional units (two), weather units (two), diet units (two), clothing units (one), home units (two), and traffic units (two).
		Android
	Comparator	None
	App version comparison	No
	Studied language	English
	Materials	The selected content of the APP application included lessons of listening, speaking, reading, and writing, based on their textbooks and levelled according to the levels of students. The teacher and English teaching experts selected and levelled the vocabulary to compile the final version.
Participants	Source of recruitment	Not specified

	Demographics	seventh-grade low-achieving students in an urban Catholic boys' middle school in north Taiwan
	Sample size	10
		10
		0
	Attrition	10
		10
		0.00%
	Pre-test or specified proficiency	No pretest. Low achievement assumed: students were selected according to the performance results of the students for the last 35% of the original class
Outcomes	Language skills	Reading, Writing, Listening, Speaking
	Outcome measurement	Two four part unit quizzes with the same level of difficulty; the dates of the two quizzes were one day apart. **Based on the selection of the vocabulary, the researchers had the students experience the most unfamiliar learning units and measure their learning results**. The content of the quizzes consisted of four parts: reading Chinese characters and circling the correct English words (4 questions), circling the correct English words according to pictures (3 questions), listening (3 questions) and speaking (4 questions).
	Outcomes (specific scores)	Sub-scores on reading, listening, speaking, and writing were reported but only total scores on each unit are extracted here. Unit 1 (out of 14) S1 14 S2 14 S3 12 S4 14 S5 14 S6 14 S7 13 S8 12 S9 12 S10 13 **Unit 1 Mean: 12.8 (1.23)** Unit 2 (out of 14) S1 13 S2 14 S3 14 S4 12 S5 14 S6 14 S7 14 S8 13 S9 12 S10 12 Unit 2 Mean: 13.2 (0.92) Total Mean: 13 (1.08)
	Effect size	Not enough information to calculate (no pretest or control group)

	Conclusion (researcher)	Overall, student vocabulary tests scores improved. Namely, the APP helped the students listen, read and write English vocabulary, but had less effect on speaking.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Other: specify in comments
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Weak
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	0
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**

10.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	12/07/2020
	Reference	Hsu, C.-K., Hwang, G.-J., & Chang, C.-K. (2013). A Personalized Recommendation-Based Mobile Learning Approach to Improving the Reading Performance of EFL Students. *Computers & Education*, *63*, 327–336.
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2013 (published)
	Duration of intervention	4 weeks
	Location of the study	Taiwan
Methods	Research question(s)	(1) Are there significant differences between the learning achievements of the three groups of students after the learning activity? (2) Do the three diverse treatments result in different cognitive loads for the students? (3) What are the students' perceptions of the learning treatments in terms of perceived satisfaction, perceived usefulness and perceived ease of use?
	Study design	A quasi-experimental study with two experimental groups and a comparator, with intact classes as group.
	Intervention	Researcher developed . Exp 1: Personalized recommendation-based mobile learning system in individual annotation mode; Exp 2: Personalized recommendation-based mobile learning system in shared annotation mode.
		Not specified
	Comparator	Mobile learning system with standard articles and individual annotation mechanism
	App version comparison	Yes
	Studied language	English
	Materials	a total ofninety-five articles suggested by the teachers were in the reading material database. The articles were categorized into three levels ofdifficulty (i.e., 26 elementary-level articles, 41 intermediate- level articles, and 28 intermediate-high-level articles) based on the definition of the General English Proficiency Test (GEPT) commissioned by the Ministry of Education in Taiwan (Roever
Participants	Source of recruitment	Not specified

	Demographics	Students at a senior high school in Taiwan, averagely 18 years old. They had identical learning process in the English course, were taught by the same teacher and had learned English for 6 years.
	Sample size	108
		33, 33
		42
	Attrition	108
		108
		0.00%
	Pre-test or specified proficiency	Control: 39.38 (8.84), Exp 1: 39.76 (9.43), Exp 2: 40.79 (11.83). The pre-test consisted of three groups of questions to examine the reading comprehension level of the students. It consisted of five multiple-choice items at the elementary level, ten multiple-choice items at the intermediate level, and ten multiple-choice items at the intermediate-high level, with a perfect score of 100.
Outcomes	Language skills	Reading
	Outcome measurement	The post-test consisted of four reading texts, including one at the elementary level, two at the intermediate level, and one at the intermediate-high level. Each text included four multiple-choice items for assessing the students' reading comprehension. The perfect score of the post-test was 100.
	Outcomes (specific scores)	A reading comprehension test with items taking from General English Proficiency Test commissioned by the Ministry of Education in Taiwan Experimental 1 Pre 39.76 (9.43) Post 71.21 (16.68) Experimental 2 Pre 40.79 (11.83) Post 68.94 (15.35) Control Pre 39.38 (8.84) Post 57.74 (23.09)
	Effect size	d=0.47

Conclusion (researcher)	The experimental results show that students in both of the experimental groups who learned with the adaptive articles (i.e., the ones recommended by the learning system based on their preferences and reading proficiency levels) achieved significantly better reading comprehension in comparison with the students who read non-adaptive reading materials in the control group, implying that the personalized recommendation mechanism was helpful to the students in motivating them and improving their reading effectiveness. On the other hand, no significant difference was found between the learning achievements of the two experimental groups, implying that the shared annotation mechanism did not result in significantly different effects on students' learning performance in comparison with the individual vocabulary annotation in this study.
Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Moderate
B. Study design	B1	Non-randomised comparison
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	60-79% (some)
	Component Rating	Moderate
D. Blinding	D1	No
	D2	Can't tell
	Component Rating	Moderate
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	Can't tell
H. Analysis	H1	class/group
	H2	class/group
	H3	Yes

	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	3
	Number of "strong" ratings	3
	GLOBAL RATING	**MODERATE**

11.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	12/07/2020
	Reference	Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*(1), 1–25.
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2015 (published)
	Duration of intervention	5 weeks, 20 minutes/week
	Location of the study	Montreal, Canada
Methods	Research question(s)	1. Does ASR-based pronunciation practice using a mobile device im- prove French L2 /y/ production? 2. Does ASR-based pronunciation practice using a mobile device im- prove French L2 /y/ perception?
	Study design	A randomized controlled experimental study with two experimental groups and a control group.
	Intervention	Commercially available. 1. ASR on an iPod Touch or iPhone using a commercial (but free) ASR application: Nuance Dragon Dictation, 2. The 'Non-ASR Group completed the same activities that the ASR participants did: They read aloud the same words and phrases in individual, weekly 20-minute ses- sions with a French teacher who provided immediate oral feedback on their pronunciation using recast and repetitions.
		iOS
	Comparator	weekly individual 20-minute meetings with the goal of practicing their conversation skills with a French teacher who provided no feedback on /y/ pronunciation.
	App version comparison	No
	Studied language	French
	Materials	five 20-minute pronunciation activities that consisted of reading aloud the target words and phrases
Participants	Source of recruitment	Not specified
	Demographics	adult French learners, either native English speakers or had native-like proficiency in the language. All had elementary-level proficiency in French (CEFR-A2 level) and, accordingly, had not yet fully acquired the target phoneme /y/.
	Sample size	42
		14, 14
		14

	Attrition	42
		42
		0.00%
	Pre-test or specified proficiency	Production (n=20) ASR 7.09 (5.51) Non-ASR 9.79 (3.98) Control 8.43 (5.86) Perception (n=15) ASR 8.07 (3.64) Non-ASR 9.64 (3.67) Control 10.29 (2.81)
Outcomes	Language skills	Phonology
	Outcome measurement	CAN-8 VirtuaLab, 'an interactive, multimedia tool used for the instruction of modern languages. The production task consisted of reading words and phrases aloud, which were recorded automatically using CAN-8 in the university's language lab, without the presence of the researcher or teacher (20 targets). In the perception task, we employed pseudowords to avoid frequency and familiarity effects; this was based on the assumption that the productivity of a pattern is often determined by its frequency in the input (15 targets).
	Outcomes (specific scores)	Production (n=20) ASR 10.71 (4.23)* Non-ASR 11.86 (3.98) Control 9.5 (5.53) Perception (n=15) ASR 9.93 (2.89) Non-ASR 10.29 (3.77) Control 10.93 (3.4)
	Effect size	Production
	Conclusion (researcher)	The present study revealed a significant improvement in /y/ production by the group that trained in an ASR-based environment. The overall success of this group on the production measures suggests that this type of learning environ- ment is propitious for the development of L2 French /y/ and, we speculate, for the development of other related segmental features.
	Conclusion (general)	Towards positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Moderate
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	No
	Component Rating	Strong

C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	No
	D2	No
	Component Rating	Strong
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Can't tell
	G3	Can't tell
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	1
	Number of "strong" ratings	5
	GLOBAL RATING	**STRONG**

12.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	12/07/2020
	Reference	Liu, J. (2016). Designing intelligent language tutoring system for learning Chinese characters [Iowa State University]. In *ProQuest Dissertations and Theses* (Issue 10167832). https://search.proquest.com/docview/1831441077?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Masters thesis
Setting	Time of study	2016 (published)
	Duration of intervention	20-30 minutes
	Location of the study	Iowa State University
Methods	Research question(s)	"1 . Is there a difference between the learning of students who use the system with metaphor pedagogy and the students using the system without it? 2. Will beginning Chinese-as-a-second-language learners' interest be increased after using the system? 3. Will the interface be user-friendly and are there any usability issues? 4. Will users using the system with the metaphor pedagogy achieve better performances (higher scores in the quiz), more interest increased, better usability assessment than users using the system without metaphor pedagogy?"
	Study design	A pre-experimental study with two intervention groups, pre-test and post-test.
	Intervention	Researcher developed. Developed in Axure RP. Version A's features include structure (navigation, flow), idea of how to build characters (orthographic decomposition and composition), metaphors, pictures, etymology and explanation, practice feedback, and instructions. Version B's includes structure and idea of how to build characters.
		iOS (iPad with 9.7-inch screen)
	Comparator	None
	App version comparison	Yes
	Studied language	Chinese
	Materials	34 characters and phrases derived from 8 characters (characters of "person", "tree", "sun", "moon", "mountain", "mouth" and "door"). The instructional content was composed of 8 mini-courses, each based on one of these eight Chinese characters.

Participants	Source of recruitment	A convenient sample through mass email to college students, staff and personal contacts at Iowa State University.
	Demographics	Students or staff at Iowa State University with no or completely forgotten knowledge of Chinese.
	Sample size	86
		43, 43
		0
	Attrition	86
		86
		0.00%
	Pre-test or specified proficiency	No knowledge of Chinese
Outcomes	Language skills	Orthography
	Outcome measurement	A quiz included in the post-survey. There were 16 questions in total. Some questions were about translating the characters presented, and some were about select the translations of a given character or English word. At the end, three questions asked participants to guess the meaning of characters they had not been taught ("guess questions")
	Outcomes (specific scores)	Post-test only because zero proficiency confirmed with pre-test Group A: 20.26 (1.498); Group B: 17.98 (3.398)
	Effect size	d=0.87
	Conclusion (researcher)	In a conclusion, the results shown were almost consistent with the predicted outcomes. Metaphor pedagogy and quiz feedback of Version A did help learning and give better user experience to users.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Non-randomised comparison
	B2	No
	B2a	No
	B2b	No
	Component Rating	Strong
C. Confounders	C1	Can't tell
	C2	Not applicable

	Component Rating	Weak
D. Blinding	D1	No
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	3
	Number of "moderate" ratings	0
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**

13.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	12/07/2020
	Reference	Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293–311. https://doi.org/10.1017/S0958344019000065
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	Early 2016
	Duration of intervention	Phase 1: at least 1 hour/week for 12 weeks. Phase 3: to achieve 34 learning hours in total (if it wasn't reached in Phase 1) (2 participants didn't reach 34 hours)
	Location of the study	Michigan State University
Methods	Research question(s)	1. How effective is Duolingo in developing L2 knowledge in ab initio learners of Turkish? 2. What are the experiences ofab initio learners ofTurkish using Duolingo to study an L2?
	Study design	A mixed methods study that examines the L2 development and experiences ofnine individuals who chose to study Turkish using Duolingo.
	Intervention	Commercially available. Duolingo
		iOS or Android or Desktop
	Comparator	None
	App version comparison	No
	Studied language	Turkish
	Materials	Duolingo curriculum
Participants	Source of recruitment	Convenience sample through a graduate seminar on instructed SLA
	Demographics	Students and professor with applied linguistics and second language acquisition background. All had previous language learning experience.
	Sample size	9
		7
		0
	Attrition	9
		7
		22.22%
	Pre-test or specified proficiency	Zero knowledge of Turkish

	Language skills	Listening, Speaking, Writing, Reading, Lexicogrammar
Outcomes	Outcome measurement	1. **Duolingo progress test**: a language proficiency measurement offered within the Duolingo platform. The test is claimed by Duolingo to be adaptive (i.e. increasing or decreasing in difficulty depending upon user performance), and the test maintains an item format similar to the instructional component. Upon completion, test takers receive a score from 0.00 to 5.00. No explanation of this score or additional performance feedback is provided by Duolingo. 2. **A Turkish 151 Test** was derived from a university-level, end-of- course summative achievement test for first semester L2 Turkish learners. It's divided into 10 sections covering aspects of listening, reading, lexicogrammar, speaking, and writing. For Turkish 151, a cut score of 70% represented a "pass" on the exam. According to the instructor, most first-year students achieved a 90% or above on the test after one semester in the Turkish L2 program.
	Outcomes (specific scores)	**Duolingo progress test** 0.63 (0.48) **Turkish 151** Total 48 (19) Listening 37 (22) Speaking 33 (14) Writing 57 (27) Reading 55 (36) Lexicogrammar 50 (16)
	Effect size	Not enough information to calculate (no raw data, pre-test or control group)
	Conclusion (researcher)	In response to the first research question, "How effective is Duolingo in developing L2 knowledge in ab initio learners of Turkish?", the results are somewhat mixed. participants knew more Turkish at the end of the study than when they began. Most notably, however, even after 34 hours of study, only one participant received a score that would be considered a passing grade in the university's first semester Turkish course.
	Conclusion (general)	Towards positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No

	B2a	No
	B2b	
	Component Rating	Moderate
C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	60-79%
	Component Rating	Moderate
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	1
	Number of "moderate" ratings	2
	Number of "strong" ratings	2
	GLOBAL RATING	**MODERATE**

14.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	19/07/2020
	Reference	Martinelli, M. (2016). Effectiveness of Online Language Learning Software (Duolingo) on Italian Pronunciation Features: A Case Study [Oklahoma State University]. In *ProQuest Dissertations and Theses* (Issue 10191215). https://search.proquest.com/docview/1964383737?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Masters thesis
Setting	Time of study	2016 (published)
	Duration of intervention	2 weeks
	Location of the study	Not specified
Methods	Research question(s)	(1) How effective is Duolingo for the acquisition of the pronunciation macro skills of accentedness and comprehensibility?; (2) How effective is Duolingo for the acquisition of the pronunciation micro skills of geminate contrast in word-medial position and voice onset time (VOT) for voiceless stops in word initial position for Italian as a second language learners?; (3) What are the learners' perceptions (user experience) of the effectiveness of Duolingo for the acquisition of pronunciation skills?
	Study design	Multiple case study
	Intervention	Commercially available, Duolingo
		iOS, Android
	Comparator	None
	App version comparison	No
	Studied language	Italian
	Materials	Duolingo curriculum
Participants	Source of recruitment	Purposeful selection, participants are selected "who can best inform the research questions and enhance understanding of the phenomenon under study"
	Demographics	The participants range in age from 23 to 51 years old, are all females, and with at least a Master's degree. Four out of five participants are native-speakers of English, while the fourth is an English as a second language (ESL) speaker.
	Sample size	5
		5
		0

	Attrition	5
		5
		0.00%
	Pre-test or specified proficiency	Identical pre- and post-tests, which tested their Italian pronunciation skills. The test contained three sections: one testing pronunciation in words, one in phrases, and one in a longer passage. In the test, the participants were exposed to both words and phrases from Duolingo and words and phrases that did not appear in the software, which were taken from a corpus of written Italian called CORIS. Stastical data such as vowel length are reported but only comprehensibility and accentedness scores are extracted here.
Outcomes	Language skills	Phonology
	Outcome measurement	Identical to pre-test
	Outcomes (specific scores)	Comprehensibility: 1-9, 1 is extremely easy to understand Pre-test Emma: 3.5 (1.91) Barbara: 4.27 (2.58) Luciana: 3.77 (2.18)* Sabrina: 4.22 (2.83) Carolina: 4.9 (2.56) Post-test Emma: 2.72 (1.70) Barbara: 4.33 (1.840) Luciana: 1.61 (0.970)* Sabrina: 2.83 (2.430) Carolina: 3.5 (2.540) Accentedness: 1-9, 1 is no foreign accent Pre-test Emma: 5.88 (2.08) Barbara: 7.11 (1.87) Luciana: 6.61 (1.88)* Sabrina: 6.16 (1.85) Carolina: 7.1 (2.06) Post-test Emma: 4.55 (1.65) Barbara: 7.16 (1.94) Luciana: 4.37 (2.32)* Sabrina: 5.77 (2.23) Carolina: 6.7 (1.76) *Significantly different score (paired two-tailed t-test)
	Effect size	Luciana only. Comprehensibility: $\eta 2=0.54$, Accentedness: $\eta 2=0.47$
	Conclusion (researcher)	The ratings from the native speakers show that Luciana achieved statistically significant improvement on both comprehensibility and accentedness, while the rest of the participants did not.
	Conclusion (general)	Mixed
Risk of Bias Assessment (EPHPP Adaptation)		

Component	Questions	Answer
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Weak
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	No
	B2b	
	Component Rating	Moderate
C. Confounders	C1	No
	C2	60-79% (some)
	Component Rating	Moderate
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	No
	E2	No
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	2
	Number of "strong" ratings	1
	GLOBAL RATING	**WEAK**

15.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	13/07/2020
	Reference	Mohamad Ali, A. Z., Segaran, K., & Hoe, T. W. (2015). Effects of Verbal Components in 3D Talking-head on Pronunciation Learning among Non-native Speakers. *Journal of Educational Technology & Society*, *18*(2), 313–322. http://search.ebscohost.com/login.aspx?direct=true&db=eax&AN=102557877&site=ehost-live&authtype=ip,uid
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2015 (published)
	Duration of intervention	15 minutes
	Location of the study	Not specified
Methods	Research question(s)	to investigate the benefit of inclusion of various verbal elements in 3D talking-head on pronunciation learning among non-native speakers
	Study design	An experimental study with three experimental groups with pretest and posttest.Random stratified sampling technique was used to group them equally based on gender.
	Intervention	Researcher developed. Group 1: 3D talking-head with spoken text and on-screen text, Group 2: 3D talking-head with spoken text alone, Group 3: spoken text with on-screen text. There would be instructions on the top of the practice screen to inform the students to select the options. Essentially, the 3D talking-head character design was developed into non-realistic facial proportion yet with proper modeled features of human to show the movement of the lip. The 3D talking-head animation was only limited to lip syncing and basic facial expressions. The automated-lip-sync technic using the special plug-in in the 3D software was utilized to synchronize the lip movement with audio.
		Not specified
	Comparator	None
	App version comparison	Yes
	Studied language	English
	Materials	Ten words: aegis, archipelago, cache, cavalry, foliage, mischievous, pronunciation, ubiquitous, suite, voluptuous
Participants	Source of recruitment	Not specified
	Demographics	college students, whose age ranged from 18 to 20, enrolled in the Diploma in Multimedia Application program and undergoing similar Essential English Communication course
	Sample size	60

		20, 20, 20
		0
	Attrition	60
		60
		0.00%
	Pre-test or specified proficiency	Pre-test and post-test were oral test that only measure the pronunciation performance of the selected words. The tests required students to read the word list given loudly within 15 seconds. Three English lecturers who are expert in linguistic were appointed to assess the students' performance to ensure the reliability of the result. The examiners assessed the students' performance based on the oral test score sheet and schema which was adapted from the existing National Education Certificate oral test score sheet. Pre: 22.91
Outcomes	Language skills	Phonology
	Outcome measurement	Identical to pre-test
	Outcomes (specific scores)	An oral test that only measure the pronunciation performance of the selected words Pre-test (covariate): 22.91 Post, adjusted Mean (SE) 3D talking-head with audio and text MALL 40.70 (0.96) 3D talking-head with audio alone MALL 36.73 (0.97) Post, adjusted Mean (SE) Audio and text alone MALL 36.77 (0.96)
	Effect size	partial eta squared = 0.17
	Conclusion (researcher)	the finding showed that students in the 3D talking-head with spoken text and on-screen text MALL learning condition, outperformed students in the 3D talking-head with spoken text alone MALL, and spoken text with on-screen text MALL, which appears to be the traditional language learning setting
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Non-randomised comparison
	B2	Yes
	B2a	No
	B2b	
	Component Rating	Strong
C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Can't tell
	Component Rating	Weak

E. Data collection practices	E1	Can't tell
	E2	Yes
	Component Rating	Moderate
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	1
	Number of "moderate" ratings	1
	Number of "strong" ratings	4
	GLOBAL RATING	**MODERATE**

16.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	13/07/2020
	Reference	Moreno, I. D., & Gatewood, K. (2019). Gamification of Foreign Language Mobile-Learning: A Quantitative Study on Spanish-as-a-Foreign Language Adult Learners' Satisfaction, Perception, Motivation, and Knowledge [Keiser University]. In *ProQuest Dissertations and Theses* (Issue 13897893). https://search.proquest.com/docview/2284212925?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Doctoral thesis
Setting	Time of study	January, 2017
	Duration of intervention	4 to 6 weeks
	Location of the study	Online
Methods	Research question(s)	1. What is the effect of the gamification of a Spanish-as-a-foreign language learning app on adult learners' satisfaction in terms of usability, motivation and knowledge? 2. Is there evidence of a relationship between adult learners' perception of the usefulness of game elements and usability, motivation and knowledge?
	Study design	A quantitative, randomized, experimenetal study with an experimental group and a control group
	Intervention	Researcher developed. a gamified mobile app prototype. It had features including game goals, instructional goals, core dynamics, game mechanics and game elements (aesthetics, competition, time, reward structure, feedback, levels, storytelling, resources and chance).
		iOS, Android or Windows
	Comparator	a non-gamified mobile app prototype
	App version comparison	Yes
	Studied language	Spanish
	Materials	Spanish conjugation in the present tense of the indicative mode within the context of daily activities and routines
Participants	Source of recruitment	Snowball sampling, social media searcher and by personal contacts
	Demographics	Native American English speakers (30 years old+) with little or no knowledge of Spanish
	Sample size	200
		100
		100
	Attrition	200
		200
		0.00%
	Pre-test or specified proficiency	Zero knowledge of Spanish

Outcomes	Language skills	Lexicogrammar (Spanish conjugation)
	Outcome measurement	Five online tests on Spanish conjugation
	Outcomes (specific scores)	Five online tests on Spanish conjugation Experimental 1: 93.06 (6.28) 2: 93.44 (4.68) 3: 94.10 (5.47) 4: 93.69 (5.41) 5: 93.07 (4.97) Control 1: 93.12 (5.75) 2: 93.45 (5.02) 3: 93.90 (5.27) 4: 94.01 (5.37) 5: 93.75 (5.17)
	Effect size	N/A (no difference)
	Conclusion (researcher)	Gamification has no effect on adult learners' knowledge acquisition.
	Conclusion (general)	No difference

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Moderate
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	No
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Can't tell

H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	1
	Number of "moderate" ratings	1
	Number of "strong" ratings	4
	GLOBAL RATING	**MODERATE**

17.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	13/07/2020
	Reference	Nicholes, J. (2016). Measuring the Impact of Language-Learning Software on Test Performance of Chinese Learners of English. *TESL-EJ*, *20*(2), 1–20. http://search.ebscohost.com/login.aspx?direct=true&db=eax&AN=118290156&site=ehost-live&authtype=ip,uid
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2016 (published)
	Duration of intervention	15 week
	Location of the study	a private Chinese college in a cross-border program with a Midwestern U.S. state-comprehensive university (SCU)
Methods	Research question(s)	How does Rosetta Stone, Tell Me More, Memrise, and ESL WOW impact the English performance on a high-stakes standardized test meant to measure reading, writing, listening, speaking, and grammar?
	Study design	A pre-experimental design with four experimental groups with four different interventions. Only three are relevant (WOW is excluded from this review)
	Intervention	Commercially available. Rosetta Stone (RS), Tell Me More (TMM), Memrise (MEM). WOW (ESL writing online workshop) is excluded for analysis because it doesn't meet the intervention criterium.
		iOS, Android, or Windows
	Comparator	None
	App version comparison	No
	Studied language	English
	Materials	App content
Participants	Source of recruitment	nonprobability sampling approach of convenience representative of classroom-based research before the semester began
	Demographics	Chinese college students, English learners
	Sample size	59
		RS=19, TMM=19, MEM=21
		0
	Attrition	59
		59
		0.00%

	Pre-test or specified proficiency	A version of a placement exam used at the researcher's U.S. university served as the pre- and posttest. The exam included, (a) two 30-minute reading sections with a total of 15 questions related to two 300-word passages; (b) a 45-minute writing section that asked participants to compose an argument in the form of an essay; (c) a 30-minute listening section of 20 questions in response to a 3-minute audio dialog; (d) an approximately 5-to-7-minute speaking section of a one-on-one conversation with the researcher; and (e) a 30-minute grammar section of 50 discrete-item and integrative questions.
Outcomes	Language skills	Reading, Writing, Listening, Speaking, Lexicogrammar
	Outcome measurement	See above, identical to pretest
	Outcomes (specific scores)	**A version of a U.S. university's English placement test** **Rosetta Stone / Tell me more / Memrise** **Reading** Pre-test 45.9(16.25) / 49.11(13.21) / 39.86(14.58) Post-test 44.84(18.45) / 55.16(12.56) / 47.42(21.12) **Writing** Pre-test 54.41(11.04)* / 63.63(10.94)* / 64.86(6.89)* Post-test 63.26(11.42)* / 68.21(11.99)* / 72.43(6.45)* **Listening** Pre-test 54.47(12.68) / 63.42(13.44) / 53.62(11.21) Post-test 49.74(17.91) / 65.79(11.69) / 54.45(10.8) **Speaking** Pre-test 50(12.25) / 82.9(12.62) / 64.29(8.26)* Post-test 51.58(12.48) / 81.58(14.34) / 68.81(9.61)* **Grammar** Pre-test 60.79(13.81)~ / 70(12.07) / 65.95(12.34) Post-test 44.84(18.45)~ / 69.79(11.7) / 68.38(10.51) **Total** Pre-test 53 (8.56) / 65.84 (7.65) / 57.714 (5.41)* Post-test 54.62 (11.40) / 68.00(8.90) / 61.57 (5.25)* **Significantly better score (paired t-test and Wilcoxon's test), ~Significantly worse score*
	Effect size	Not stated by author. d=0.72

	Conclusion (researcher)	Although RS and TMM groups both improved only on writing, and although the RS group's writing improvement proved statistically highly significant, the RS group's starting out at a lower level seems to have offered more room for fundamental improvement. Still, at least among this group of learners, RS failed to motivate participants to approach the target of 3 hours per week, 45 hours total over the 15-week duration of the study.
	Conclusion (general)	Mixed

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	Yes
	C2	Can't tell
	Component Rating	Weak
D. Blinding	D1	Yes
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Yes
H. Analysis	H1	class/group
	H2	class/group
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	3
	Number of "moderate" ratings	0
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**

18.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	
	Reference	Ou-Yang, F. C., & Wu, W.-C. V. (2017). Using Mixed-Modality Vocabulary Learning on Mobile Devices: Design and Evaluation. *Journal of Educational Computing Research*, *54*(8), 1043–1069. https://search.proquest.com/docview/1969019983?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2017 (published)
	Duration of intervention	2 weeks
	Location of the study	Taiwan
Methods	Research question(s)	"1. Will a mobile learning system using mixed-modality vocabulary learning enhance the learning effectiveness of two groups of university students (English majors and non-English majors)? Will students of different profi- ciency levels have different learning effectiveness? 2. Which PLSs (visual, auditory, kinesthetic, tactile, and group/individual) pro- vide the best learning effectiveness using the mobile learning system? 3. Do university students with different learning attitudes and proficiencies have different preferences regarding the mobile learning system? 4. What are learners' use behaviors regarding English vocabulary learning using the mobile learning system?"
	Study design	A pre-experimental study with two experimental groups, pre-test and post-test: English majors and none-English majors.
	Intervention	Researcher developed. MyEVA Mobile: The smartphone MyEVA users learned target words using four kinds of VLSs, word card, flash card, Chinese assonance, and imagery.
		Android
	Comparator	None
	App version comparison	No
	Studied language	English
	Materials	Fifty words taken from the common TOEIC vocabu- lary were chosen, including words of different degrees of difficulty and syntactical functions
Participants	Source of recruitment	Not specified
	Demographics	university students from two groups in Taiwan, non-English majors in the College of Management, Chien Hsin University of Science and Technology, and English majors in the Department of English Language, Literature, and Linguistics at Providence University.

	Sample size	108
		53, 55
		0
	Attrition	108
		108
		0.00%
	Pre-test or specified proficiency	The pretest mean scores of the non-English majors were 27.70 (M=27.70, SD=14.60, SEM=2.01). English major pretest mean scores were 59.20 (M=59.20, SD=15.30, SEM=2.06).
Outcomes	Language skills	Vocabulary
	Outcome measurement	an English proficiency pretest/posttest. There were 50 vocabulary words for the participants to learn and they were tested on all the 50 words accordingly. The proficiency test was designed and verified by five language instructors/scholars whose expertise is on English vocabulary acquisition to ensure its validity. The test was comprised of 50 items and the total score was 100. There were three parts to the proficiency test: Part 1 (10 items) was an English translation into Chinese (the item was an English word and the participants were required to write the Chinese translation); Part 2 (20 items) was multiple choice (participants selected the appropriate Chinese meaning of the provided English word); Part 3 (20 items) was fill in the blank (the participants filled the correct missing words in sentences according to context).
	Outcomes (specific scores)	Non-English major Pre-test 27.70 (14.60) Post-test 48.38 (19.71) English major Pre-test 59.20 (15.30) Post-test 75.42 (15.72)
	Effect size	Not stated by author. Non-English major: d=1.19, English major: d=1.05. Can't calculate between groups because the assumptions are not met.
	Conclusion (researcher)	The mixed-modality vocabulary mobile learning system significantly enhanced non-English major and English major learning achievement.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Not Likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Non-randomised comparison
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	Yes
	C2	Can't tell
	Component Rating	Weak
D. Blinding	D1	Yes

	D2	Yes
	Component Rating	Weak
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Can't tell
	G3	Can't tell
H. Analysis	H1	class/group
	H2	class/group
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	0
	Number of "strong" ratings	4
	GLOBAL RATING	**WEAK**

19.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	15/07/2020
	Reference	Purgina, M., Mozgovoy, M., & Blake, J. (2020). WordBricks: Mobile Technology and Visual Grammar Formalism for Gamification of Natural Language Grammar Acquisition. Journal of Educational Computing Research, 58(1), 126–159. https://doi.org/10.1177/0735633119833010
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2016 (conference proceeding)
	Duration of intervention	Not specified
	Location of the study	Japan
Methods	Research question(s)	"gather initial responses from the teacher and the students, and assess the feasibility of the chosen approach"
	Study design	a pre- test/post-test design with a control group and experimental group. A quasi-experimental design with two intact classes randomly assigned to experimental conditions.
	Intervention	Researcher developed. WordBricks
		Android
	Comparator	taught with an English grammar textbook in a traditional way (teacher-centered, grammar focused)
	App version comparison	No
	Studied language	English
	Materials	units 69 and 70 from the same *English Grammar in Use* textbook (Murphy, 2012) with the same teacher. Both units are about countable and uncountable nouns. However, Unit 70 seems to be more demanding than introductory Unit 69, since it is dedicated to more advanced grammatical points.
Participants	Source of recruitment	Not specified
	Demographics	predominantly male sophomore computer science students at a public Japanese university. The students were enrolled in an elective advanced English grammar course, and ranged in age from 19 to 21 years. First,
	Sample size	21
		10
		11
	Attrition	21
		21
		0.00%
	Pre-test or specified proficiency	A separate diagnostic grammar test from the pre-test (out of 100).

		Control: 65.25 (7.50) Experimental: 65.90 (7.99) **Unit 69** Experimental (G2) Pre-test 15.90 (4.43) Control (G1)Pre-test 15.18 (5.04) **Unit 70**Experimental (G2) Pre-test 4.20 (2.57) Control (G1) Pre-test 6 (2.72)
	Language skills	Lexicogrammar
	Outcome measurement	two sets of researcher-developed paper-based English grammar tests to measure participants' English grammar performance over two course units. For Unit 69 pre-/post-test, participants were asked to correct given sentences focusing on the nouns of the sentences. For Unit 70 pre-/post-test, they were asked to complete sentences using correct noun form
Outcomes	Outcomes (specific scores)	Unit 69 Experimental (G2) Pre-test 15.90 (4.43) Post-test 24.20 (4.02) Control (G1) Pre-test 15.18 (5.04) Post-test 21 (5.80) Unit 70 Experimental (G2) Pre-test 4.20 (2.57) Post-test 11.60 (2.84) Control (G1) Pre-test 6 (2.72) Post-test 9.18 (4.17) *The article reports a third group in a separate study which is not included in this review because insufficient information is given.*
	Effect size	Not stated by author. Between groups: Unit 69: d=0.64, Unit 70: d=0.68
	Conclusion (researcher)	The diagrams show that in both groups teaching was effective, and students in general were able to improve their test scores. Most progress was made by the participants having lower initial scores, which is not surprising. It is also noticeable that the difference in pre-test and post-test scores is larger in the WordBricks group.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
	A1	Somewhat Likely

A. Selection bias	A2	80 - 100% agreement
	Component Rating	Moderate
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Can't tell
	G3	Can't tell
H. Analysis	H1	class/group
	H2	class/group
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	1
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**

20.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	15/07/2020
	Reference	Rachels, J. R., & Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. Computer Assisted Language Learning, 31(1), 72–89. https://doi.org/10.1080/09588221.2017.1382536
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2016 (dissertation publish date)
	Duration of intervention	12 weeks, 40 minutes/week
	Location of the study	South Florida, USA
Methods	Research question(s)	"(1) What is the effect of a foreign language mobile gamification application on ele- mentary students' Spanish language achievement, while controlling for a Span- ish language achievement pretest? (2) Does elementary students' academic self- efficacy differ based on the type of foreign language instruction provided (i.e. traditional vs. a foreign language mobile gamification application), while con- trolling for an academic self-efficacy pretest?"
	Study design	A quasi-experimental, pretest-posttest, non-equivalent control group design with 11 intact classes randomly assigned to experimental conditions. Two third-grade and three fourth-grade classes served as the treatment group, and the other three third-grade and the other three fourth-grade classes served as the control group
	Intervention	Commercially available, Duolingo. (The teacher was present in the classroom to answer questions and provide instruction about the application as needed. Not about the materials so it's ok)
		iOS (iPad)
	Comparator	traditional face-to-face Spanish instruction, which was adapted to cover the same vocabulary and material as covered in the Duolingo® application.
	App version comparison	No
	Studied language	Spanish
	Materials	first 60 lessons of Duolingo® Spanish course
Participants	Source of recruitment	a convenience sample
	Demographics	third and fourth grade students from a private school in South Florida, age from seven to ten years
	Sample size	467
		77
		88
	Attrition	187
		167
		10.70%

	Pre-test or specified proficiency	treatment group 11.78 (9.94) control group 11.78 (8.88)
Outcomes	Language skills	Vocabulary, Writing (translation), lexicogrammar
	Outcome measurement	Researcher developed test based on materials: The test contained 50 questions and was comprised of 20 single-word, vocabulary questions, 20 questions in which phrases had to be translated, and 10 questions built around identifying correct grammar. Of the 50 questions, 25 required the student to translate Spanish to English and 25 required the student to translate English to Spanish.
	Outcomes (specific scores)	Treatment group 20.94 (9.93) control group 21.47 (10.20)
	Effect size	N/A. No difference.
	Conclusion (researcher)	These results indicate that the students using the Duolingo application achieve similarly to those receiving tradi- tional face-to-face classroom instruction.
	Conclusion (general)	No difference

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Moderate
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	No
	Component Rating	Moderate
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	class/group
	H2	class/group
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	2

Number of "strong" ratings	4
GLOBAL RATING STRONG	

21.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	15/07/2020
	Reference	Ramírez, G., ElMiligi, H., Walton, P., & Billy, C. (2018). Revitalization of an Indigenous Language: Which is Better the Teacher or the App? International Conference on E-Learning, 2018-July, 344-350,XVII. https://search.proquest.com/docview/2081757033?accountid=13042
	Source	Main search from an electronic database
	Language	English
	Document type	Conference proceedings
Setting	Time of study	2018 (published)
	Duration of intervention	Three 15 to 20 minute sessions
	Location of the study	an Indigenous community school located in the interior of British Columbia, Canada that served primarily Indigenous children
Methods	Research question(s)	"What is the effectiveness of using apps to teach Secwepemctsín to elementary school children?"
	Study design	a quantitative approach with an experimental design with two conditions (control and experimental), and participants were randomly assigned to a condition
	Intervention	Researcher developed. 1: Dress me Up app: This app was designed for children in Kindergarten to Grade 3. It taught children the words for glasses, shirt, hat, pants, shoes, and skirt through interaction (drag and drop, animation) and recorded voices. Children could choose a character (male or female) and a setting (kitchen, living room, bedroom, or dining room). The app had two levels of difficulty. 2. Matching UP: This clothing-themed matching game app was aimed at children in Grades 4 to Grade 7. It taught 12 words: belt, earrings, glasses, necklace, purse, ring, scarf, socks, underwear, and vest. Users progressed through seven levels of increasing numbers of vocabulary items populated with randomly selected elements.
		Android
	Comparator	Same target words
	App version comparison	No
	Studied language	Secwepemctsín
	Materials	DMU: words for glasses, shirt, hat, pants, shoes, and skirt through interaction (drag and drop, animation) and recorded voices, MU: 12 words: belt, earrings, glasses, necklace, purse, ring, scarf, socks, underwear, and vest.
Participants	Source of recruitment	Not specified
	Demographics	The participants were children from k to 7 (N = 96), 60% were boys and 40% girls
	Sample size	96
		48

		48
	Attrition	96
		96
		0.00%
	Pre-test or specified proficiency	Zero knowledge (assumed)
Outcomes	Language skills	Vocabulary
	Outcome measurement	A paper and pencil, multiple-choice vocabulary test containing 12 items for the primary grades (in the first six items they just heard the target words and for the second six they heard the command accompanied by the word, e.g., choose the hat) and 10 for the intermediate grades. There were two tests. The first one was administered after 20 minutes of exposure to the words (through apps or teacher) and the second on a different day after 40 minutes of exposure, on two different days during 20 minutes on each occasion.
	Outcomes (specific scores)	Experimental (primary) Pre-test 81% Post-test 81% Experimental (intermediate) Pre-test 71% Post-test 72% Control (primary) Pre-test 80%* Post-test 85%* Control (intermediate) Pre-test 87%* Post-test 94%*
	Effect size	Time 2: partial eta squared = .093
	Conclusion (researcher)	The main findings were that overall, children learned more words from the teacher than from the app. Also, whereas student in the teacher condition continued to grow in their vocabulary knowledge form Time 1 to Time 2, children in the apps condition stagnated.
	Conclusion (general)	Negative

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Moderate
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Moderate
D. Blinding	D1	Can't tell
	D2	Can't tell
	Component Rating	Weak

E. Data collection practices	E1	Can't tell
	E2	Yes
	Component Rating	Moderate
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Can't tell
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	1
	Number of "moderate" ratings	3
	Number of "strong" ratings	2
	GLOBAL RATING	**MODERATE**

22.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	
	Reference	Terantino, J. (2016). Examining the Effects of Independent MALL on Vocabulary Recall and Listening Comprehension: An Exploratory Case Study of Preschool Children. CALICO Journal, 33(2), 260–277.
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2016 (published)
	Duration of intervention	at least 15 minutes each day over a six-month period
	Location of the study	USA
Methods	Research question(s)	1. Where, when, and how do preschool-aged children participate in independent iPad-based language learning? 2. Among the free Spanish apps available for the iPad, which do preschool- aged children prefer for independent language learning? What are the common characteristics of these apps? 3. To what extent are preschool-aged children capable of demonstrat- ing increases in animal vocabulary recall as a result of participating in independent iPad-based language learning? 4. To what extent are preschool-aged children capable of demonstrat- ing increases in listening comprehension for animal vocabulary as a result of participating in independent iPad-based language learning?
	Study design	An exploratory case study method
	Intervention	Commercially Available. Spanish Smash, LinguPinguin, Busuu Kids, Bilingual Child Bubbles, Bilingual Child
		iOS (iPad)
	Comparator	None
	App version comparison	Yes (all different apps)
	Studied language	Spanish
	Materials	
Participants	Source of recruitment	A total of 22 children (enrolled in the same preschool class) were invited to participate in the study. Ultimately, the participants of this research were purposively selected based on the parents' willingness to participate.
	Demographics	preschool-aged children (four to five years of age) enrolled at a private preschool in the United States.
	Sample size	7
		7
		0
	Attrition	7

		7
		0.00%
	Pre-test or specified proficiency	two of the participants had minimal prior exposure to Spanish before participating in the study. The remaining five participants had no prior exposure. The average pretest score was 1.71 out of 23 (7.43%, SD = 3.40). As Figure 4 indicates, five of the seven participants were unable to accurately identify any of the 23 vocabulary items. It should also be men- tioned that one of the two children (participant 1) with prior linguistic expo- sure scored notably higher on the pretest (9 out of 23 items) in
Outcomes	Language skills	Vocabulary, listening
	Outcome measurement	an identical vocabulary recall posttest containing the same 23 animal vocabulary words and images
	Outcomes (specific scores)	A vocabulary recall test containing animal vocabulary Pre-test: 1.71 (3.40) Post-test: 12.86 (3.34) Listening comprehension test on animal vocabulary Pre-test: 1.71 (3.40) Post-test: 12.86 (2.79)
	Effect size	Vocabulary: r^2=0.96, Listening r^2=0.97
	Conclusion (researcher)	the children demonstrated gains in vocabulary recall and listening comprehension
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	Less than 60% agreement
	Component Rating	Weak
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	Yes
	C2	Less than 60% (few or none)
	Component Rating	Weak
D. Blinding	D1	Can't tell
	D2	Yes
	Component Rating	Moderate
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Yes
H. Analysis	H1	learner

	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	3
	Number of "moderate" ratings	2
	Number of "strong" ratings	1
	GLOBAL RATING	**WEAK**

23.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Thongsri, N., Shen, L., & Yukun, B. (2019). Does academic major matter in mobile assisted language learning? A quasi-experimental study. International Journal of Information and Learning Technology, 36(1), 21–37.
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	2019 (published)
	Duration of intervention	3 weeks
	Location of the study	Wuhan, China
Methods	Research question(s)	H1. STEM students' rating on CSE (computer self-efficacy) is higher than non-STEM students. H2a. MALL has a significant positive effect on learning score for STEM more than non-STEM students. Mobile assisted language learning H2b. MALL has a significant positive effect on learning satisfaction for STEM more than non-STEM students. H3a. CSE has a significant positive effect on learning score. H3b. CSE has a significant positive effect on learning satisfaction. H4a. The effect of CSE on learning score will be stronger for STEM students than non-STEM ones. H4b. The effect of CSE on learning satisfaction will be stronger for STEM students than non-STEM ones.
	Study design	a quasi-experimental, pre-test/post-test design with two groups with the same condition but different on independent variables, non-randomised comparison
	Intervention	Commercially available. BW Vocabulary.
		Android or iOS
	Comparator	None
	App version comparison	No
	Studied language	English
	Materials	the vocabulary list covered general topics and not specifically related to a specific major.
Participants	Source of recruitment	Deliberate sampling

	Demographics	undergraduate students who took part in a short-term English training workshop. The participants were chosen from departments including electrical engineering, management, computer science, telecommunication, public administration, business and education
	Sample size	200
		STEM 100, non-STEM 100
		0
	Attrition	200
		200
		0.00%
	Pre-test or specified proficiency	English vocabulary test containing 100-item multiple-choice tasks with four choices for each item. STEM: 44.81 (8.33), non-STEM: 45.27 (8.12)
	Language skills	Vocabulary
	Outcome measurement	Identical to the pre-test
Outcomes	Outcomes (specific scores)	An English vocabulary test containing 100-item multiple-choice tasks with four choices for each item. Pre-test STEM 44.81 (8.33) non-STEM 45.27 (8.12) Post-test STEM 71.62 (12.01) non-STEM 63.89 (14.28)
	Effect size	Not stated by author. Pre/post: STEM: d=0.50, non-STEM: d=0.52
	Conclusion (researcher)	MALL has a significant positive effect on learning score (for STEM more than non-STEM students)
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Not Likely
	A2	Can't tell
	Component Rating	Weak
B. Study design	B1	Non-randomised comparison
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Strong
	C1	Yes

C. Confounders	C2	60-79% (some)
	Component Rating	Moderate
D. Blinding	D1	Yes
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Yes
H. Analysis	H1	class/group
	H2	class/group
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	1
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**

24.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Vesselinov, R., & GREGO, J. (2015). Efficacy of a New Language App (Issue May).
	Source	Forward citation
	Language	English
	Document type	Organisation report
Setting	Time of study	February to April 2015
	Duration of intervention	2 months
	Location of the study	Online
Methods	Research question(s)	evaluate the efficacy of a new language app (The actual name of the language app tested in this study was replaced by thegeneric “Language App” and “LA” because the report was not officially made public.)
	Study design	A pre-experimental, pre-test/post-test design with a single intervention group
	Intervention	Commercially available. A new language app, not detailed.
		Not specified
	Comparator	None
	App version comparison	No
	Studied language	Spanish
	Materials	Not specified
Participants	Source of recruitment	LA provided to the Research team the e-mail addresses of their new subscribers and we drew a representative sample based on the pool of eligible users.
	Demographics	- Willing to participate in the study; - Studying Spanish as a foreign language; - At least 18 years of age; - Not of Hispanic origin; - Not living in a Spanish-speaking country; - Not advanced users of Spanish. The
	Sample size	101
		101
		0
	Attrition	231
		101
		56.28%
	Pre-test or specified proficiency	Web Based Computer Adaptive Placement Exam (WebCAPE test). Pre-test: 243.4 (117.6), [0, 418]. 95% CI: 220.2-266.6
Outcomes	Language skills	Lexicogrammar, Reading, Vocabulary

	Outcome measurement	WebCAPE
	Outcomes (specific scores)	314.4 (113.8), [0, 543], 95% CI: 291.9-336.9
	Effect size	Not enough information to calculate (no correlation or raw data)
	Conclusion (researcher)	Beginner users of Spanish gain on average about 18 test points per one hour of study.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	Less than 60%
	Component Rating	Weak
G. Invention integrity	G1	Less than 60%
	G2	Yes
	G3	Yes
H. Analysis	H1	learner
	H2	learner
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	1
	Number of "moderate" ratings	1
	Number of "strong" ratings	3
	GLOBAL RATING	**MODERATE**

25.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Vesselinov, R., & GREGO, J. (2019). Pimsleur Efficacy Study (Issue April).
	Source	Forward citation
	Language	English
	Document type	Organisation report
Setting	Time of study	February to April 2019
	Duration of intervention	2 months
	Location of the study	USA
Methods	Research question(s)	evaluate the efficacy of The Pimsleur® Language Program
	Study design	A pre-experimental, pre-test/post-test design with a single intervention group
	Intervention	Commercially available. Pimsleur
		iOS, Android
	Comparator	None
	App version comparison	No
	Studied language	Spanish
	Materials	Pimsleur curriculum
Participants	Source of recruitment	The random sample for this study was drawn from existing or new Pimsleur users residing in the U.S.
	Demographics	- Be willing to study Spanish using only Pimsleur for two months; - Take two sets of oral proficiency language tests; - Be at least 18 years of age; - Be novice or beginner learners of Spanish.
	Sample size	82
		82
		0
	Attrition	120
		82
		31.67%
	Pre-test or specified proficiency	The language test used in the study was designed and developed by an external independent testing company. An oral proficiency test. No details are given about the test. Proficiency: novice low: 78/82, 95.1% novice mid: 2/82, 2.4% novice high: 2/82, 2.4%
Outcomes	Language skills	Speaking
	Outcome measurement	Identical to pre-test

	Outcomes (specific scores)	Proficiency: novice low: 33/80, 41.3%novice mid: 38/80, 47.5%novice high: 6/80, 7.5%intermediate low: 2/80, 2.5%intermediate mid: 1/80, 1.3%
	Effect size	Not enough information to calculate (no correlation or raw data)
	Conclusion (researcher)	The main goal of measuring the efficacy of Pimsleur was achieved with this study. The results show that, on average, 83% of the users who follow the Pimsleur method of half an hour per lesson and complete 30 lessons (Level 1) improve their oral proficiency level at least by one and up to 3 levels. The 95% confidence interval for this efficacy measure is between 63% and 94%.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	60-79%
	Component Rating	Moderate
G. Invention integrity	G1	60-79%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	2
	Number of "strong" ratings	3

GLOBAL RATING	**MODERATE**

26.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Vesselinov, R., & GREGO, J. (2019). Mango Languages Efficacy Study.
	Source	Forward citation
	Language	English
	Document type	Organisation report
Setting	Time of study	July to September 2019
	Duration of intervention	2 months
	Location of the study	Not specified
Methods	Research question(s)	evaluate the efficacy of Mango languages app
	Study design	A pre-experimental, pre-test/post-test design with a single intervention group
	Intervention	Commercially available. Mango languages.
		iOS, Android
	Comparator	None
	App version comparison	No
	Studied language	Spanish
	Materials	Mango curriculum
Participants	Source of recruitment	In July 2019, emails were sent to existing Mango Languages users with an invitation to participate in a Spanish language study for two months. They were directed to an online survey designed by the Research Team. This survey collected demographic information, and self-evaluation of their language proficiency level.
	Demographics	- be willing to study Spanish using only Mango Languages for two months. The minimum -requirement was a total of 2 study hours for the two-month period. take two sets of written and oral proficiency language tests; - be at least 18 years of age; - be novice or beginner learners of Spanish.
	Sample size	95
		95
		0
	Attrition	149
		95
		36.24%
	Pre-test or specified proficiency	1. WebCAPE: Written Proficiency: Vocabulary/Reading/Grammar.

		97.2 (106.9), 95% CI: 74.4-120.0 2. TrueNorth test: oral procieicncy 2.4 (1.4), 95% CI: 2.1-2.7
Outcomes	Language skills	Vocabulary, Reading, Lexicogrammar, Speaking
	Outcome measurement	WebCAPE, TrueNorth
	Outcomes (specific scores)	WebCAPE (n=87) 229.4 (106.9), 95% CI: 204-255 TNT 3.4 (1.1), 95% CI: 3.2-3.7
	Effect size	Not enough information to calculate (no correlation or raw data)
	Conclusion (researcher)	• On average participants gained 18 written proficiency WebCAPE points per one hour of study; • On average participants gained 0.13 points of the TNT oral proficiency test per one hour of study;
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	60-79%
	Component Rating	Moderate
G. Invention integrity	G1	60-79%
	G2	Yes
	G3	No

H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	2
	Number of "strong" ratings	3
	GLOBAL RATING	**MODERATE**

27.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Vesselinov, R., & Grego, J. (2016). The Babbel Efficacy Study (Issue September).
	Source	Forward citation
	Language	English
	Document type	Organisation report
Setting	Time of study	June to August 2016
	Duration of intervention	8 weeks
	Location of the study	Online; Berlin, Germany; New York City, Usa
Methods	Research question(s)	evaluate the efficacy of Babbel app
	Study design	A pre-experimental, pre-test/post-test design with a single intervention group
	Intervention	Commercially available. babbel
		iOS, Android
	Comparator	None
	App version comparison	No
	Studied language	Spanish
	Materials	Babbel curriculum
Participants	Source of recruitment	The random sample for this study was selected from existing or new Babbel users in the US and Germany. People who lived close to Berlin or New York City (NYC) were asked to take a proctored test in Babbel's offices in Berlin and NYC.
	Demographics	- Willing to study Spanish using only Babbel for two months, and come to the testing location for two sets of language tests;- At least 18 years of age; - Not advanced learners of Spanish.
	Sample size	325
		325
		0
	Attrition	391
		325
		16.88%
	Pre-test or specified proficiency	WebCAPE Pre-test 106.4 (108.6), 95% CI: 96.8-120.4
Outcomes	Language skills	Vocabulary, Reading, Lexicogrammar
	Outcome measurement	WebCAPE.
	Outcomes (specific scores)	265.1 (124.2), 95% CI: 251.5-278.7
	Effect size	Not enough information to calculate (no correlation or raw data)

	Conclusion (researcher)	• Overall 92% of the participants improved their language proficiency. • Babbel users need on average 21 hours of study in a two-month period to cover the requirements for one college semester of Spanish.• Truly novice users with no knowledge of Spanish need on average 15 hours of study in a two-month period to cover the requirements for one college semester of Spanish.
	Conclusion (general)	Positive
Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals/ dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	1
	Number of "strong" ratings	4
	GLOBAL RATING	**STRONG**

28.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Vesselinov, R., & Grego, J. (2012). Duolingo Effectiveness Study. In *City University of New York, USA* (Issue December 2012).
	Source	Forward citation
	Language	English
	Document type	Organisation report
Setting	Time of study	September - Novermber 2012
	Duration of intervention	8 weeks
	Location of the study	Online
Methods	Research question(s)	evaluate the efficacy of Duolingo app
	Study design	A pre-experimental, pre-test/post-test design with a single intervention group
	Intervention	Commercially available. Duolingo
		iOS, Android
	Comparator	None
	App version comparison	No
	Studied language	Spanish
	Materials	Duolingo curriculum
Participants	Source of recruitment	The Duolingo study on effectiveness was announced on the web. Duolingo included a link advertising the new study in Spanish on its website and put out a Google ad. In Duolingo the link was only visible to users who were logged in and were studying Spanish. People who were interested in participating in the study were asked to click on the link and go to the invitation page.
	Demographics	- Willing to participate in the study; - Studying Spanish as a foreign language; - At least 18 years of age; - Native speakers of English; - Residing in the U.S.; - Not of Hispanic origin; - Not advanced users of Spanish.
	Sample size	88
		88
		0
	Attrition	196
		88
		55.10%
	Pre-test or specified proficiency	webCAPE. 158.0 (115.4) [0, 405]
Outcomes	Language skills	Vocabulary, Reading, Lexicogrammar
	Outcome measurement	WebCAPE
	Outcomes (specific scores)	253.9 (110.5) [0, 539]
	Effect size	Not enough information to calculate (no correlation or raw data)

	Conclusion (researcher)	• Overall the average improvement in language abilities was 91.4 points and the improvement was statistically significant. • The effectiveness measure showed that on average participants gained 8.1 points per one hour of study with Duolingo. • The study estimated that a person with no knowledge of Spanish would need between 26 and 49 hours (or 34 hours on average) to cover the material for the first college semester of Spanish. This result is based on the language test's cut-?-off point for the second college semester and the 95% Confidence Interval of the effectiveness measure.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	Less than 60%
	Component Rating	Weak
G. Invention integrity	G1	60-79%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	1
	Number of "moderate" ratings	1
	Number of "strong" ratings	3
	GLOBAL RATING	**MODERATE**

29.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Vesselinov, R., & Grego, J. (2016). The busuu efficacy study (Issue May).
	Source	Forward citation
	Language	English
	Document type	Organisation report
Setting	Time of study	February to April 2016
	Duration of intervention	8 weeks
	Location of the study	London, UK; New York, US
Methods	Research question(s)	evaluate the efficacy of busuu app
	Study design	A pre-experimental, pre-test/post-test design with a single intervention group
	Intervention	Commercially available. busuu
		iOS, Android
	Comparator	None
	App version comparison	No
	Studied language	Spanish
	Materials	Busuu curriculum
Participants	Source of recruitment	E-mail messages were sent out to busuu clients with an invitation to participate in the research study. If they accepted the invitation they were asked to complete the online Entry Survey with some demographic questions and questions about their knowledge of Spanish.
	Demographics	- Willing to study Spanish using only busuu for two months, and come to the testing location for two sets of language tests; - At least 18 years of age; - Not of Hispanic origin; - Not advanced learners of Spanish.
	Sample size	144
		144
		0
	Attrition	196
		144
		26.53%
	Pre-test or specified proficiency	1. WebCAPE, 24.8 (122.4), 95% CI: 104.6-145.0 2. Oral Proficiency Interview by Computer® (OPIc)
Outcomes	Language skills	Vocabulary, Reading, Lexicogrammar, Speaking
	Outcome measurement	WebCAPE, OPIc
	Outcomes (specific scores)	WebCAPE: 269.6 (151.4), 95% CI: 244.6-294.5 OPIc: At the beginning of the study the truly novice users (Un-Ratable and Novice Low) were almost 60% of the sample while at the end their numbers decreased to about 18%. This is a completely different level of oral proficiency.

	Effect size	Not enough information to calculate (no correlation or raw data)
	Conclusion (researcher)	Written Proficiency Gain: • Overall 84% of the participants improved their written proficiency. • busuu users need on average 22.5 hours of study in a two-month period to cover the requirements for one college semester of Spanish. Oral Proficiency Gain: • Over 75% of busuu users increased their oral proficiency by at least one level.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals dropouts	F1	Yes
	F2	60-79%
	Component Rating	Moderate
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	2
	Number of "strong" ratings	3
	GLOBAL RATING	**MODERATE**

30.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Vesselinov, R., Grego, J., Sacco, S. J., & Tasseva-Kurktchieva, M. (2019). The 2019 Rosetta Stone Efficacy Study (Issue January). http://comparelanguageapps.com/documentation/The2019_RS_FinalReport.pdf
	Source	Forward citation
	Language	English
	Document type	Organisation report
Setting	Time of study	June to August 2018
	Duration of intervention	2 months
	Location of the study	Online
Methods	Research question(s)	evaluate the efficacy of Rosetta Stone app
	Study design	A pre-experimental, pre-test/post-test design with a single intervention group
	Intervention	Commercially available. Rosetta Stone
		iOS, Android
	Comparator	None
	App version comparison	No
	Studied language	Spanish
	Materials	Rosetta Stone curriculum
Participants	Source of recruitment	The random sample for this study was drawn from existing or new Rosetta Stone users residing in the U.S.
	Demographics	- Be willing to study Spanish using only Rosetta Stone for two months with at least two hours of study; - Take two sets of language placement tests; - Be at least 18 years of age; - Be novice or beginner learners of Spanish.
	Sample size	143
		143
		0
	Attrition	175
		143
		18.29%
	Pre-test or specified proficiency	WebCAPE Pre-test 42.0 (68.5), 95% CI: 31.2-53.5
Outcomes	Language skills	Lexicogrammar, Reading, Vocabulary
	Outcome measurement	WebCAPE

Outcomes (specific scores)	145.2 (110.0) ,95% CI: 127.4-163.7
Effect size	Not enough information to calculate (no correlation or raw data)
Conclusion (researcher)	• Overall 92% of the participants improved their language proficiency. • Babbel users need on average 21 hours of study in a two-month period to cover the requirements for one college semester of Spanish.• Truly novice users with no knowledge of Spanish need on average 15 hours of study in a two-month period to cover the requirements for one college semester of Spanish.
Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Very likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Cohort (one group pre + post (before and after))
	B2	No
	B2a	Not Applicable
	B2b	Not Applicable
	Component Rating	Moderate
C. Confounders	C1	No
	C2	Not applicable
	Component Rating	Strong
D. Blinding	D1	Not Applicable
	D2	Not Applicable
	Component Rating	Not Applicable
E. Data collection practices	E1	Yes
	E2	Yes
	Component Rating	Strong
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	Yes
	G3	No
H. Analysis	H1	learner
	H2	learner
	H3	No

	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	0
	Number of "moderate" ratings	1
	Number of "strong" ratings	4
	GLOBAL RATING	**STRONG**

31.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Wu, Q. (2015). Pulling mobile assisted language learning (MALL) into the mainstream: MALL in broad practice. PLoS ONE, 10(5).
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	Spring semester of 2012-2013 academic year and the spring semester of 2013-14 academic year
	Duration of intervention	1 semester
	Location of the study	Jiujiang University, China
Methods	Research question(s)	1. Explore the effectiveness ofusing smartphones as a tool for learning English vocabulary in a natural environment. It was based on the assumption that the participants with Word Learning-CET4 installed in their smartphones would have more encounters with these words due to its accessibility, 2. Introduce a pedagogical example ofhow to build a smartphone learning project for learning vocabulary.
	Study design	A quasi-experimental, pre-test/post-test design with three pairs of intact classes, in which three was randomly assigned to the intervention condition and the other one as control
	Intervention	Researcher developed. Word Learning-CET 4 Word, a B4A application with touchscreen commands including search command, glossary command, unknown word command, sample test command, touchscreen selection, and exit
		Android
	Comparator	All students in the three pairs of intact classes received printed copies of the 3,402 words in a wordlist to study. The university did not change textbook and the three instructors taught the three pairs ofclasses with the same cur- ricular textbook, NewHorizon College English: Reading andWriting, book 4, 2nd edition
	App version comparison	No
	Studied language	English
	Materials	3,402 words in the CET-4 vocabulary books
Participants	Source of recruitment	Not specified
	Demographics	Sophomore students at a mid-ranked university in China, who were in "good classes." The researcher as- sumed the three pairs ofclasses had no significant difference in IQ and learning ability
	Sample size	199
		101
		98

	Attrition	199
		199
		0.00%
	Pre-test or specified proficiency	An English vocabulary test with 100 words selected from the app materials. Pre-test Experimental group 50.84 (13.11) Control group 49.80 (12.68)
Outcomes	Language skills	Vocabulary
	Outcome measurement	Identical to pre-test
	Outcomes (specific scores)	An English vocabulary test with 100 words selected from the app. Experimental Pre-test 50.84 (13.11) Post-test 60.61 (11.96) Control Pre-test 49.80 (12.68) Post-test 51.08 (12.08)
	Effect size	d=0.41
	Conclusion (researcher)	Using Word Learning-CET4 the participants in the experimental group remembered 8.49% more words than the participants in the control group. This is a significant result and this researcher believes there are three reasons for it.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Strong
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	No
	E2	No
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong
G. Invention integrity	G1	80-100%
	G2	No
	G3	Yes
H. Analysis	H1	class/group

	H2	class/group
	H3	Yes
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	0
	Number of "strong" ratings	4
	GLOBAL RATING	**WEAK**

32.

Cat.	Item	Response
General Information	Date form completed (dd/mm/yy)	17/07/2020
	Reference	Wu, Q. (2015). Designing a smartphone app to teach English (L2) vocabulary. Computers & Education, 85, 170–179. https://doi.org/10.1016/j.compedu.2015.02.013
	Source	Main search from an electronic database
	Language	English
	Document type	Peer-reviewed journal article
Setting	Time of study	July to August 2014
	Duration of intervention	55 days
	Location of the study	Jiujiang University, China
Methods	Research question(s)	1. Explore the effectiveness of using smartphones as a tool for teaching/learning English vocabulary in a natural environment. It was based on the assumption that the participants with Word Learning-CET6 installed in their smartphones would have more encounters with these words due to its accessibility, 2. Introduce a pedagogical example of employing a smartphone app for teaching/learning vocabulary.
	Study design	An experimental, pre-test/post-test design with an experimental group and a control group
	Intervention	Researcher developed. Basic4Android: the researcher drew the concepts and structures of Word Learning-CET6 and designed it. The framework was to mimic a paper word list or vocabulary book, attaching two important functions of Unknown Words and Sample Test, which are difficult to be accomplished by a learner with a paper word list. Word Learning-CET6 was framed to execute three simple functions: a) Displaying words alphabetically; b) Selecting (or de-selecting) words from word database and building a new word pool to increase the efficiency of learning; c) Performing sample test to evaluate learning effectiveness.
		Android
	Comparator	Each of the 70 participants received a printed wordlist copy of these 1274 words.
	App version comparison	No
	Studied language	English
	Materials	1,274 words at CET-6 level
Participants	Source of recruitment	Not specified
	Demographics	4th year medical school students of a 5-year college system in Jiujiang University
	Sample size	70
		35
		35
	Attrition	70
		70

		0.00%
	Pre-test or specified proficiency	An English vocabulary test with 100 words selected from the app. Experimental Pre-test 35.13 Control Pre-test 35.68
Outcomes	Language skills	Vocabulary
	Outcome measurement	Identical to pre-test
	Outcomes (specific scores)	An English vocabulary test with 100 words selected from the app. Experimental Pre-test 35.13 Post-test 50.52 (10.41) Control Pre-test 35.68 Post-test 43.56 (11.40)
	Effect size	d=0.55
	Conclusion (researcher)	The results show that participants in the test group remembered 88.67 more words than participants in the control group at the end of the research. The difference was significant. Considering that there was no significant difference between groups in the pretest and all participants were treated equally except for that the participants in the test group had Word Learning-CET6 app in their smartphones, the advantage of using smartphones with Word Learning-CET6 for EFL learners to learn English vocabulary is attested.
	Conclusion (general)	Positive

Risk of Bias Assessment (EPHPP Adaptation)		
Component	**Questions**	**Answer**
A. Selection bias	A1	Somewhat Likely
	A2	80 - 100% agreement
	Component Rating	Moderate
B. Study design	B1	Randomized controlled trial
	B2	Yes
	B2a	No
	B2b	Not Applicable
	Component Rating	Strong
C. Confounders	C1	No
	C2	80-100% (most)
	Component Rating	Strong
D. Blinding	D1	Can't tell
	D2	Can't tell
	Component Rating	Weak
E. Data collection practices	E1	Can't tell
	E2	Can't tell
	Component Rating	Weak
F. Withdrawals, dropouts	F1	Yes
	F2	80-100%
	Component Rating	Strong

G. Invention integrity	G1	80-100%
	G2	No
	G3	Yes
H. Analysis	H1	learner
	H2	learner
	H3	No
	H4	Yes
GLOBAL RATING	**Number of "weak" ratings**	2
	Number of "moderate" ratings	1
	Number of "strong" ratings	3
	GLOBAL RATING	**WEAK**