NCAL#2035401, Vol 0, ISS 0

A systematic review examining the efficacy of commercially available foreign language learning mobile apps

Jodi M. Tommerdahl, Chrystal Sapphire Dragonflame and Amanda A. Olsen

QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left are hyperlinked to the location of the query in your paper.

The title and author names are listed on this sheet as they will be published, both on your paper and on the Table of Contents. Please review and ensure the information is correct and advise us if any changes need to be made. In addition, please review your paper as a whole for typographical and essential corrections.

Your PDF proof has been enabled so that you can comment on the proof directly using Adobe Acrobat. For further information on marking corrections using Acrobat, please visit http://journalauthors.tandf.co.uk/production/acrobat.asp; https://authorservices.taylorandfrancis.com/how-to-correct-proofs-with-adobe/

The CrossRef database (www.crossref.org/) has been used to validate the references.

AUTHOR QUERIES

Q1 Please check and confirm whether the author affiliations and corresponding details have been set correctly.

Q2 The keywords information has been imported from data supplied with the original manuscript. Please revise if incorrect.

Q3 Please provide the short biography of the author in the "Notes on contributor" section.

Q4 The disclosure statement has been inserted. Please correct if this is inaccurate.

Q5 Please provide complete details for (ILTA, 2007, JLTA, 2002, EALTA, 2006, and Purpura, 2004) in the reference list or delete the citation from the text.

Q6 There is no mention of (Clearinghouse 2012 and Purpura 2013) in the text. Please insert a citation in the text or delete the reference as appropriate.

Q7 Please provide the volume number.

Q8 Please provide the volume number.

Q9 Please provide the year.

Q10 Please provide the page range.

Q11 Please note that the ORCID section has been created from information supplied with your manuscript submission/CATS. Please correct if this is inaccurate.

COMPUTER ASSISTED LANGUAGE LEARNING
https://doi.org/10.1080/09588221.2022.2035401

Routledge
Taylor & Francis Group

A systematic review examining the efficacy of commercially available foreign language learning mobile apps

Jodi M. Tommerdahl[a,b], Chrystal Sapphire Dragonflame[a,b] and Amanda A. Olsen[a]

[a]Department of Curriculum and Instruction, University of Texas at Arlington, Arlington, TX, USA; [b]Southwest Center for Mind, Brain, and Education, University of Texas at Arlington, Arlington, TX, USA

ABSTRACT

A systematic review examining the efficacy of commercially available foreign language-learning apps (FLL) was completed. A database search of ERIC, PsychINFO, and LearnTechLib produced 1,786 journal articles. After applying specific inclusion and exclusion criteria based on Burston's seminal study (2015) requiring a minimum number of 10 participants, quantitative learning outcome data and rigorous research design, eight studies remained. These studies were categorized in terms of the app studied, year of publication, language taught, age group of participants, setting, length of study, and device(s) used. Descriptive statistics demonstrate there is a dearth of studies examining app efficacy, that English was the most commonly taught language, and that vocabulary was the most commonly tested area. Although commercial apps were found to successfully support FLL, the included studies' methods varied in ways that made direct comparison difficult.

KEYWORDS

Mobile apps; technology; foreign language-learning; mobile assisted language learning

Since the inception of Apple's app store and Android Market in 2008, mobile applications (apps) have grown exponentially in popularity, rapidly expanding to meet the enormous demand for novel varieties of apps (Pandey et al., 2019; Yu, 2019). For example, the Apple app store opened with 500 apps which by 2017 had expanded to 2.2 million. Although many apps, such as Angry Birds in 2009 (Rovio Entertainment Corporation, 2020), were created for their amusement value, others offered utility in areas such as social media, business, fitness, and education (Toto & Limone, 2019). Within the category of educational apps, the subcategory of foreign language learning (FLL) quickly emerged,

CONTACT Jodi M. Tommerdahl joditom@uta.edu Department of Curriculum and Instruction, University of Texas at Arlington, 701 S Nedderman Dr, Arlington, TX 76019, USA.

demonstrating the public's strong desire to learn foreign languages in an accessible and convenient way.

An early example of the enthusiasm and interest for FLL apps can be seen in the example of Babbel, which entered the market in 2007 (Babbel, 2020). Although they do not release their overall subscriber numbers, they reported a gain of 50 million new downloads worldwide in 2018 alone (Chan, 2019).

Duolingo, arguably one of the most well-known mobile apps for language learning, also achieved rapid success. Within three weeks of launching in the Apple app store in 2012, the app reached one million downloads, with no advertising except by word of mouth. By 2013, four million individuals used the app, and in 2018 alone, the app was downloaded another 50 million times (Chan, 2019). Duolingo offers over 30 different languages to study with several more in either the development or beta stage (Duolingo, 2020). Although we are unsure of how many FLL apps exist in the world's app stores, a perusal of any major app download site allows us to estimate that at least several hundred are available, although the true number could be in the thousands.

Language learning apps vary widely from each other in qualities such as their selection of languages taught, and the native language used for instruction. For instance, an app developed for teaching Norwegian to Chinese speakers would not be appropriate for English speakers learning Norwegian. This means that some apps may have several different versions for different audiences, even when teaching the same language. Apps also vary by the type of language skills they emphasize. Some apps may focus on understanding the pronunciation of the studied language, or concern themselves solely with vocabulary development, while others may attempt to blend several linguistic skills. Another variation between FLL apps is their revenue model. Revenue can be earned through advertising on the app, relying on in-app purchases, completing a one-time payment for access to the app, or through a recurring subscription fee.

Given the large number of consumers who are using and often paying for FLL apps, the need for research into their efficacy is apparent, both for the consumer who is trying to decide between them and for app designers who want to ensure their apps are effective and attractive. The need for efficacy studies has been increasingly recognized in recent years in the field of education within core subject areas such as mathematics (Doabler et al., 2019), science literacy (Goldman et al., 2019), and reading (Vaughn et al., 2019; Wanzek et al., 2011) with Toste et al. (2019, p. 46) referring to this need as "a central mission of educational research." This need is further evidenced by U.S. federal policy in the form of the Every Student Succeeds Act, a program which mandates

that federal money must be used for programs that show evidence of effectiveness (Dynarski, 2015) and the creation of the What Works Clearinghouse, a website that provides research-based efficacy guidelines and information in several areas of education (https://ies.ed.gov/ncee/wwc/). When considering efficacy, it is important to define both efficacy and an efficacy perspective. The current paper draws heavily from the Efficacy Framework put forward by American College Testing (ACT) (Mattern, 2019) which defines efficacy as "the degree to which evidence, rationales, and theory support the claim that a learning tool improves intended learner outcomes under ideal conditions" (p. 2). More specifically, the framework targets 'results/impact,' defined as "the degree to which targeted outcomes occur as a result of using the learning tool." This is important as designers of instructional tools need to base their work on specific knowledge components to optimize learning (Mattern, 2019).

The Efficacy Framework seeks to align the validity of assessments and learning product efficacy, allowing them to achieve two stated goals:

1. Develop learning solutions that are most likely to impact the intended outcome.
2. Increase our ability to detect whether a learning solution is achieving its intended outcome through thoughtful study design, methodology, and data collection.

The ACT Framework proposes that the quantification of efficacy arises from the intersection of learning and measurement, as shown in Figure 1. This requires measuring the impact of the learning tool on the intended outcome, which is in turn aligned with the Knowledge-Learning-Instruction Framework (Koedinger et al., 2012).

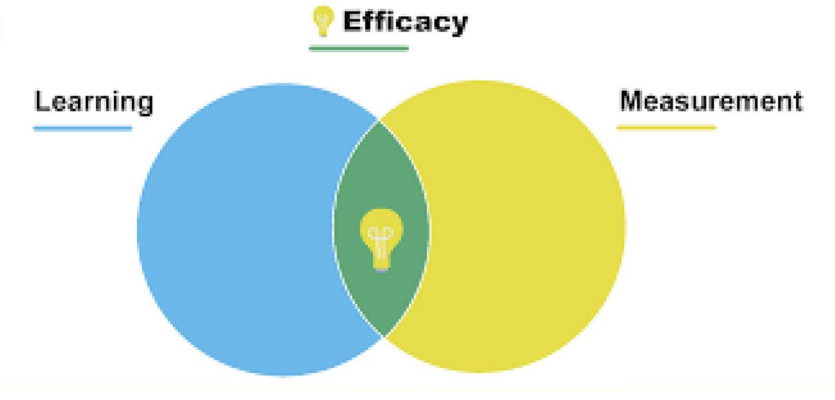

Figure 1. ACT's efficacy framework: intersection of learning and measurement.

To judge the efficacy of instructional tools, apps in this case, it is important that rigorous efficacy studies are completed. Then systematic reviews can help compare and contrast the varying studies to accurately evaluate the research. Although several reviews of mobile assisted language learning (MALL) exist, none to our knowledge focus exclusively on commercially available mobile phone and tablet apps designed for FLL. This study's objective was to focus on this specific area, firstly to determine the amount of rigorous research into investigating their efficacy, and secondly, to examine the results of those publications in an effort to help individuals such as language learners, foreign language instructors, and app designers to become more aware of which apps have been shown to be effective.

Literature review

Although no efficacy studies of FLL mobile apps appeared in our search of existing literature, FLL mobile apps have been studied in other contexts apart from their efficacy such as within the larger categories and MALL. The following literature review reports on these related areas not exclusive to FLL mobile app efficacy, but which nonetheless report on FLL apps.

Review of FLL: Games without efficacy measures

In a FLL review, Dehghanzadeh et al. (2019) studied the effect of games, many of which were apps, on learning English as an additional language in a digital environment. Twenty-two studies were reviewed, and all reported positive experiences for students using ESL games. However, the efficacy of language learning was not measured. Instead, the study examined engagement and motivation. The contents of the 17 of 22 games studied were mostly focused on vocabulary with five focusing on grammar, four on pronunciation, five on speaking, four on listening, and three on writing. The overall finding was that participants found the games to be "enjoyable, engaging, motivating, and fun" (Dehghanzadeh et al., 2019, p. 1).

Review specific to apps for FLL but not measuring efficacy

In a similar paper, Heil et al. (2016) reviewed the 50 most popular FLL commercial apps with regard to three questions: 1) What are the primary pedagogical focuses of popular language learning apps?; 2) Do apps adapt to individual needs, language proficiency levels, and styles of learning?; and 3) How is corrective feedback employed in these apps?

These questions stemmed from theoretical models of language which categorized different aspects of language learning such as knowledge of grammar, pragmatics, discourse, and sociolinguistics (Bachman & Palmer, 1996; Purpura, 2004). Further, Heil et al. (2016) argued that several different areas of language must be integrated for effective language learning, rather than promoting the sheer memorization of words. Their review found that 42 out of 50 commercial FLL apps taught vocabulary in isolation, whereas only 12 offered grammar instruction, demonstrating that a majority of the most popular commercial FLL apps were not appropriately equipped for teaching conversational language. This supports Dehghanzadeh et al. (2019) results which discovered an emphasis on vocabulary in language learning games.

Reviews measuring efficacy of MALL for FLL but not limited to apps

Although previous reviews have analyzed different aspects of technology and language learning, few have examined the actual efficacy of technological tools to enhance learning (Bolgün & McCaw, 2019). This was surprising, given the emphasis that researchers and government bodies have placed on understanding the efficacy of educational interventions. Furthermore, consumers who are using, and often paying for FLL apps, with the intent of learning a new language quickly and efficiently, should have research-based guidance available when making their choice. Although some data regarding the efficacy of FLL apps exists, it is typically embedded in studies that have analyzed the larger category of mobile assisted language learning (MALL). A discussion of these studies follows, including descriptions of their methodologies since they are meta-analyses, as is the current paper. Note, each of the following MALL studies included mobile apps, but their data was not analyzed separately from the other types of MALL, thereby not allowing readers to measure the efficacy of mobile apps specifically.

For example, Sung et al. (2015) completed a meta-analysis analyzing the efficacy of mobile devices for teaching FLL, extending beyond apps to include items such as text messaging, social media, global positioning systems, and video capture. They also examined commonalities found within MALL articles published between 1993 and 2013, a timeframe notably beginning before mobile apps, by examining the type of participants, hardware and software, teaching and learning methods, settings, language skills, target languages, and intervention duration. From an initial search returning 721 results including journal articles, conference papers, and doctoral dissertations, two criteria were applied: 1) the research question was required to be about mobile device use in foreign language learning, and 2) only experimental and quasi-experimental

studies that allowed for the calculation of effect sizes were included. Screening using these criteria reduced the final number of articles to 44. Their results reported that the most common age group of participants was elementary-school students, the most common timeframe for experiments using learning-oriented software was one to six months, the most frequently studied learning skill was vocabulary, and the most common language being learned was English.

When analyzed statistically, 70% of participants using mobile devices for FLL outperformed their control group, providing evidence of FLL mobile device effectiveness. Further findings revealed that adults and school children had similar results from using MALL, handhelds had a larger effect size than laptops, use of mobile devices in multiple learning settings was more effective than the more restricted settings of the outdoors or the classroom, interventions of 1-6 months were more effective than shorter or longer time periods, and finally, using mobile devices for vocabulary or mixed skills produced higher achievement than those which focused on individual skills such as listening or reading. Despite the value of Sung et al. (2015) findings, the research does not provide data specific to mobile apps although it includes technologies such as social media and video capture.

Additionally, Burston (2015) completed a rigorous systematic review of the learning outcomes of MALL implementation projects. This included 291 articles published between 1994 and 2012, again mainly predating apps, but examining a range of technologies including video playback, flashcards, and speech recordings. Inclusion criteria applied to these 291 articles included having a minimum of 10 participants using the mobile device for a minimum of one-month, quantitative learning outcome data, tracing the amount of usage of the mobile device, adequate control group descriptions, absence of uncontrolled and confounding variables, and adequate statistical analyses. After excluding papers that did not meet this threshold, only 19 of the 291 studies could meaningfully evaluate the efficacy of MALL usage.

Out of the 19 studies, children and adult participants were equally represented with vocabulary being the most studied area. The most common treatment duration was from four to six weeks with the most common number of participants being in the 25- to 49-person range. An attempt to calculate effect sizes to compare across papers resulted in nine more papers being excluded due to their failure to report the language level of the participants. The final 10 papers were split between five different areas of language, vocabulary, reading, speaking, listening, and writing, making a comparison of effect sizes impossible. Despite the large number of MALL papers published, this review draws attention to how few reliable studies were available regarding MALL efficacy. It

is particularly interesting to notice that these two reviews restricted inclusion criteria to quantitative, experimental and quasi-experimental studies, therefore, the initial number of studies found decreased greatly, showing that most research was limited in its ability to provide quality statistical results of efficacy. Sung et al. (2015) research included only 6.1% of the initial articles found while Burston's (2015) included only 3.4%, likely due to the stricter criteria including number of participants, minimum amount of time of study length, and quality statistical analyses among others.

The limited amount of research on the efficacy of MALL, long pre-dating 2008, emphasizes the question of how much research has been completed on the efficacy of FLL apps that are available to consumers in app stores. While FLL apps have been included in larger studies of efficacy, they have not been examined independently as a group. Although commercial FLL apps generally claim to use effective pedagogical concepts in their design (Toto & Limone, 2019), it is unclear whether research evidence to support their efficacy exists, and if so, for which apps and for what language areas. Without this knowledge, it is impossible for consumers to make informed decisions on whether using an app for FLL is beneficial, and if so, which are the most effective. The intent of the present systematic review was to answer the following questions:

1. How many scientifically reputable studies exist on FLL app efficacy showing learning outcomes?
2. What general trends exist in the research (number of studies per year and age groups studied)?
3. What languages do they teach and to what linguistic audiences?
4. What specific linguistic skills have been measured in these studies?
5. What are the efficacy outcomes of the studies?

Methods

This study was part of a larger study examining the efficacy of all FLL apps, whether commercially available or not (Olsen et al., under review). Although a close examination of FLL apps that are not commercially available may be of interest for professional app developers, it was felt that the public's need for information would be better served by focusing on obtainable apps. Similarly, language teachers looking for mobile apps to assist their students may be interested in the efficacy of commercial apps compared to those that are unpublished and unavailable.

To answer the five research questions, a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses

for Systematic Reviews Protocols (PRISMA-P) (Moher et al., 2015; Shamseer et al., 2015) was completed in March of 2020. Moher et al. (2015) explained that systematic reviews should "build on a protocol that describes the rationale, hypothesis, and planned methods of the review" (p.1). The PRISMA-P is a list consisting of 17 items ensuring that reviews are prepared and reported in a rigorous and robust manner and is often used in healthcare. A systematic review was chosen for this study since studies may have different outcomes, used different participants, and/or different mobile apps, meaning a meta-analysis would not be appropriate (Sriganesh et al., 2016).

After carrying out an extensive search for existing systematic reviews on the efficacy of FLL apps and finding none to exist, a list of criteria was developed, based on Burston's (2015) article. These criteria are listed below.

Criteria for Inclusion

1. Published in peer-reviewed journals from 2008 onwards
2. Written in English
3. Stated an intention to examine the efficacy of an app in relation to FLL outcomes
4. Implemented a research methodology (qualitative, quantitative or mixed methods)
5. Used mobile apps that were designed to teach a foreign language
6. Used mobile apps that were commercially available in an app store

Criteria for Exclusion

1. Fewer than 10 participants in the study
2. Research design shortcomings such as the existence of uncontrolled variables, lack of a control group or inadequate statistical analyses

Although most criteria were identical to Burston's, two exclusion criteria were dropped and one inclusion criterion was added. Specifically, Burston required the amount of time interacting with an app to be a minimum of one month; however, it was decided that an app showing efficacy in language learning in a shorter period of time would be of interest to the public. Burston's limitation of articles to those that included an in-app tracking of time was also dropped, as the authors felt that the amount of time an app was open on a participant's mobile device did not necessarily reflect focus and attention given to the app. The decision to omit two of Burston's criteria allowed us the potential to include more studies. Regarding inclusion criteria, this study added

the requirement that apps studied be commercially available in an app store, thereby allowing the public and FLL educators to make decisions regarding apps that are actually available to them instead of having to sort through numerous studies where apps were designed specifically by researchers for a study and were not available to the public.

With the assistance of a university librarian, strings of search terms were developed and entered into the databases. The specific search strings used are provided in Table 1.

Databases searched included the Educational Resources Information Center (ERIC), PsychINFO, and LearnTechLib. This search produced 1,786 studies which were uploaded into Covidence, a software specifically designed for the development of systematic reviews. Using this software and the rules of PRISMA-P, the inclusion and exclusion criteria were applied to each article by the first (A.O.) and third (J.T.) authors. Reviews were completed in two stages, the abstract review, followed by a full-text review when appropriate. At each stage the author accepted or rejected each paper blindly, without the input of the other author until all articles had been reviewed and categorized. Only at that point were conflicts addressed. These conflicts were resolved through the close examination of each article and discussion between the authors. Each remaining article was categorized according to several variables of interest by the final author (J.T.). The second author (C.D.) also re-categorized 10% of the articles that were randomly selected in order to ensure reliability. Agreement across the variables of interest was over 95%.

Results

The first and most striking result was that only eight studies met the inclusion criteria, even with the easing of some of Burston's restrictions (see Figure 2).

This stands in stark contrast to what was initially expected given that a search of "Duolingo" on academic databases shows numerous results. Although it appears that a great deal is being published about FLL apps, very few are rigorous efficacy studies interested in learning outcomes. Figures 3–13 display general trends of this limited group of studies.

None of the papers in this review were published until 2016, eight years after the introduction of FLL apps to app stores. Research of this type does not appear to be increasing since that time, with an average of approximately two papers a year since 2016. No developed body of work in peer reviewed journals exists with regard to any individual FLL app. In fact, the only app to have been effectively studied for efficacy in more than one paper was Duolingo which was represented here only twice (25%).

Table 1. The protocol executed in each database.

Search Order	Database	Protocol
1	ERIC	TX (mobile applications or apps or smartphone or iphone or iOS or android or lingualift or duolingo or hellotalk or mindsnacks or busuu or babbel or triplingo or mosalingua or cell phones or ipods or mobile devices or mobile phones or memrise or hinative or "(How to) Pronounce" or mondly or lirica or drops or pimsleur or beelinguapp or clozemaster or fluentu or accellastudy essential apps or tandem or lyrics training or "mobile-assisted language learning" or MALL) or [DE]Handheld devices
2	ERIC	((DE "Second Language Learning" OR DE "Bilingual Education" OR DE "Bilingual Instructional Materials" OR DE "Bilingualism" OR DE "College Second Language Programs" OR DE "Communicative Competence (Languages)" OR DE "English (Second Language)" OR DE "Language Enrichment" OR DE "Second Language Instruction" OR DE "Second Language Programs") OR TX "second language acquisition" or TX "mobile-assisted language learning"
3	ERIC	(DE "Academic Achievement" OR DE "Educational Attainment" OR DE "Student Promotion" OR DE "Academic Ability" OR DE "Academic Aptitude" OR DE "Academic Aspiration" OR DE "Academic Failure" OR DE "Academic Probation" OR DE "Achievement Rating" OR DE "College Readiness" OR DE "Educational Indicators" OR DE "Excellence in Education" OR DE "Grades (Scholastic)" OR DE "Grading" OR DE "Growth Models" OR DE "Instructional Effectiveness" OR DE "Knowledge Level" OR DE "Performance" OR DE "Reading Achievement" OR DE "School Effectiveness" OR DE "Student Evaluation" OR DE "Vocabulary Development" OR DE "Verbal Development" OR DE "Language Acquisition" OR DE "tests" OR DE "Scores" OR DE "Success" DE "Performance Based Assessment" OR DE "Language Aptitude" OR DE "Language Fluency" OR DE "Language Proficiency" OR DE "Language Skills" OR DE "Language Tests") OR TX (student success or GPA or grade point average or academic performance or learning performance or test scores OR vocabulary learning or vocabulary acquisition)
4	ERIC	#1 AND #2 AND #3
5	ERIC	#5 Limits: English, Scholarly (Peer Reviewed) Journals, Academic Journals, Publication Date 2008-
1	PsychINFO	#1 TX (apps or iphone or iOS or android or lingualift or duolingo or hellotalk or mindsnacks or busuu or babbel or triplingo or mosalingua or cell phones or ipods or mobile devices or mobile phones or memrise or hinative or "(How to) Pronounce" or mondly or lirica or drops or pimsleur or beelinguapp or clozemaster or fluentu or accellastudy essential apps or tandem or lyrics training or "mobile-assisted language learning" or MALL) or DE "Smartphones" OR DE "Mobile Phones" OR DE "Digital Gaming" OR DE "Mobile Applications" OR DE "Smartphone Use"
2	PsychINFO	#2 DE "Foreign Language Learning" OR DE "Bilingual Education" OR DE "Foreign Languages" OR DE "English as Second Language" OR DE "Language Development" OR DE "Language Proficiency" OR DE "Lexical Access" OR DE "Linguistics" OR DE "Foreign Language Education" OR DE "Vocabulary" OR TX (second language learning or second language acquisition or "mobile-assisted language learning")
3	PsychINFO	DE "Academic Achievement" OR DE "Academic Overachievement" OR DE "Academic Underachievement" OR DE "College Academic Achievement" OR DE "Academic Achievement Motivation" OR DE "Academic Achievement Prediction" OR DE "Academic Aptitude" OR DE "Academic Failure" OR DE "Verbal Comprehension" OR DE "Listening Comprehension" OR DE "Reading Comprehension" OR DE "Sentence Comprehension" OR TX (student success or GPA or grade point average or academic performance or learning performance or test scores OR vocabulary learning or vocabulary acquisition)
4	PsychINFO	#1 AND #2 AND #3
5	PsychINFO	Limits: English, Scholarly (Peer Reviewed) Journals, Academic Journals, Publication Date 2008-

(*Continued*)

Table 1. Continued.

Search Order	Database	Protocol
1	LearnTechLib (full texts and abstracts)	(mobile applications OR apps OR smartphone OR iphone OR iOS OR android OR lingualift OR duolingo OR hellotalk OR mindsnacks OR busuu OR babbel OR triplingo OR mosalingua OR cell phones OR ipods OR mobile devices OR mobile phones OR memrise OR hinative OR "(How to) Pronounce" OR mondly OR lirica OR drops OR pimsleur OR beelinguapp OR clozemaster OR fluentu OR accellastudy essential apps OR tandem OR lyrics training OR "mobile-assisted language learning" OR MALL OR handheld devices) AND (foreign language learning OR bilingual education OR foreign language OR english as second language OR language development OR language proficiency OR lexical access OR linguistics OR foreign language education OR vocabulary) AND (academic achievement OR academic overachievment OR academic underachievment OR academic ability OR academic aptitude OR academic aspiration OR academic failure OR academic probation OR educational indicators OR excellence in education OR grades OR grading OR instructional effectiveness OR knowledge level OR performance OR reading achievement OR student evaluation OR vocabulary development OR verbal development OR language acquisition OR tests OR scores OR success OR language aptitude OR language fluency OR language proficiency OR language skills OR language tests OR student success OR GPA OR grade point average OR academic performance OR learning performance OR vocabulary learning OR vocabulary acquisition)
2	LearnTechLib (journal articles)	(MAINSUBJECT("Cell Phones") OR "smartphones" OR "tablet computers" OR "portable computers" OR "handheld computers" OR "apps" OR "iphone" OR "iOS" OR "android" OR "lingualift" OR "duolingo" OR "hellotalk" OR "mindsnacks" OR "busuu" OR "babbel" OR "triplingo" OR "mosalingua" OR "cell phones" OR "ipods" OR "mobile devices" OR "mobile phones" OR "memrise" OR "hinative" OR "(How to) Pronounce" OR "mondly" OR "lirica" OR "drops" OR "pimsleur" OR "beelinguapp" OR "clozemaster" OR "fluentu" OR "accellastudy essential apps" OR "tandem" OR "lyrics training" OR "mobile-assisted language learning" OR "MALL") AND (MAINSUBJECT("Second Language Learning") OR MAINSUBJECT("Bilingual Education") OR MAINSUBJECT("Second Languages") OR MAINSUBJECT("Linguistic Interference") OR MAINSUBJECT("Direct Method of Language Teaching") OR MAINSUBJECT("Audiolingual Language Teaching") OR "second language learning" OR "second language acquisition" OR "mobile-assisted language learning" OR "MALL" OR "language teaching methods" OR "language acquisition" OR "foreign language learning") AND (MAINSUBJECT("Learning") OR MAINSUBJECT("Achievement Tests") OR MAINSUBJECT("Academic Achievement") OR MAINSUBJECT("Tests") OR MAINSUBJECT("Reading Comprehension") OR MAINSUBJECT("Reading Ability") OR MAINSUBJECT("Reading Tests") OR MAINSUBJECT("Reading Acquisition") OR MAINSUBJECT("Reading Achievement") OR MAINSUBJECT("Verbal Learning") OR MAINSUBJECT("Listening Comprehension") OR MAINSUBJECT("Language Tests") OR MAINSUBJECT("Pronunciation Accuracy") OR MAINSUBJECT("Language Proficiency") OR MAINSUBJECT("English Proficiency") OR "student success" OR GPA OR "grade point average" OR "academic performance" OR "learning performance" OR "test scores" OR "vocabulary learning" OR "vocabulary acquisition" OR "language proficiency" OR "pronunciation accuracy" OR "reading achievement") AND stype.exact("Scholarly Journals") AND at.exact("Article") AND la.exact("English") AND PEER(yes)

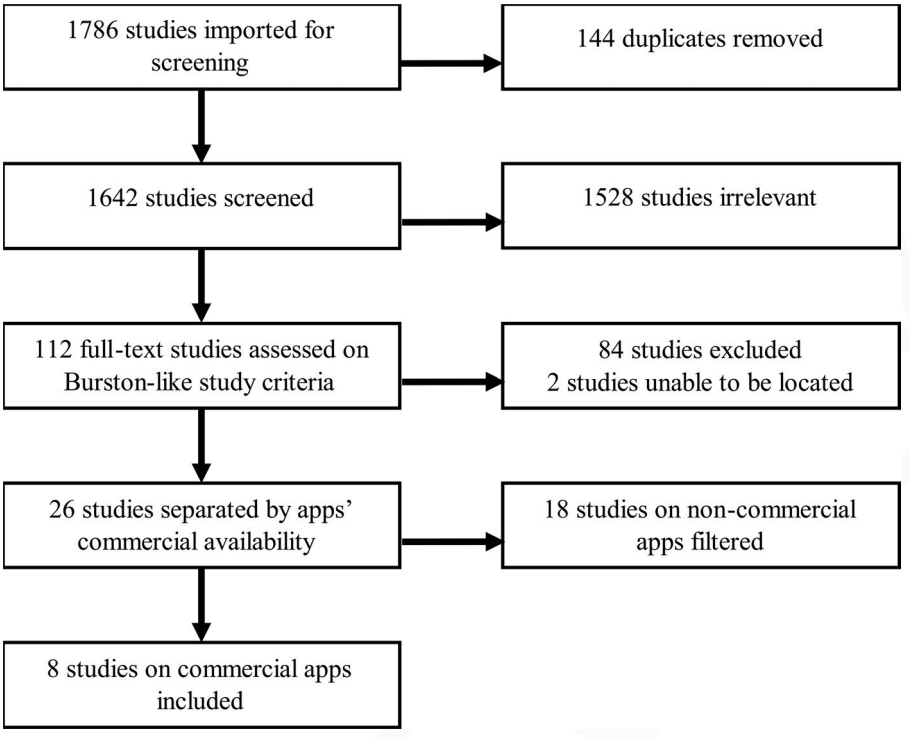

Figure 2. PRISMA diagram for foreign language mobile apps published between 2008 and 2020.

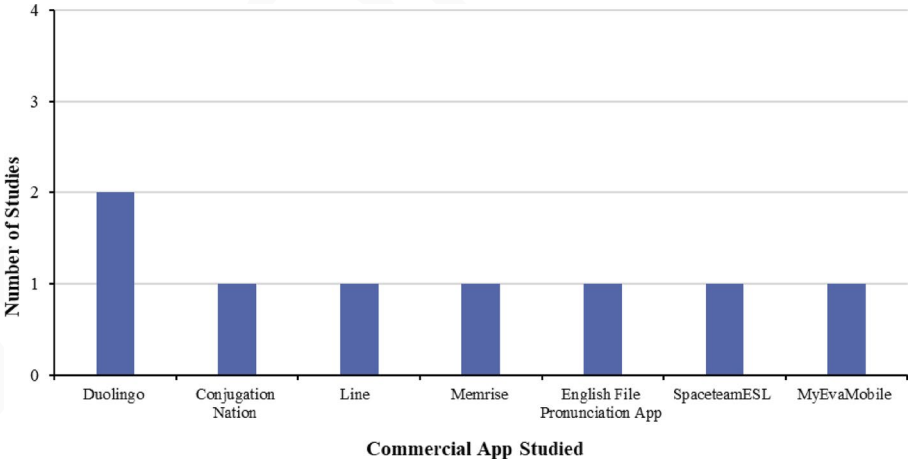

Figure 3. Number of studies per commercial app.

The number of target languages for learning was limited to three with only one French (12.5%), two Spanish (25%), and five English (62.5%) papers represented. The language that apps used to communicate with the users were French, Spanish, English, Persian, and Chinese.

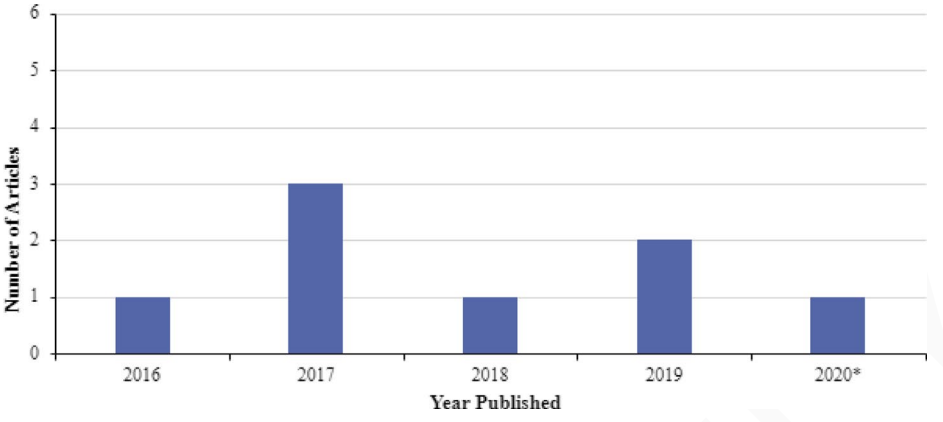

Figure 4. Number of studies published per year.

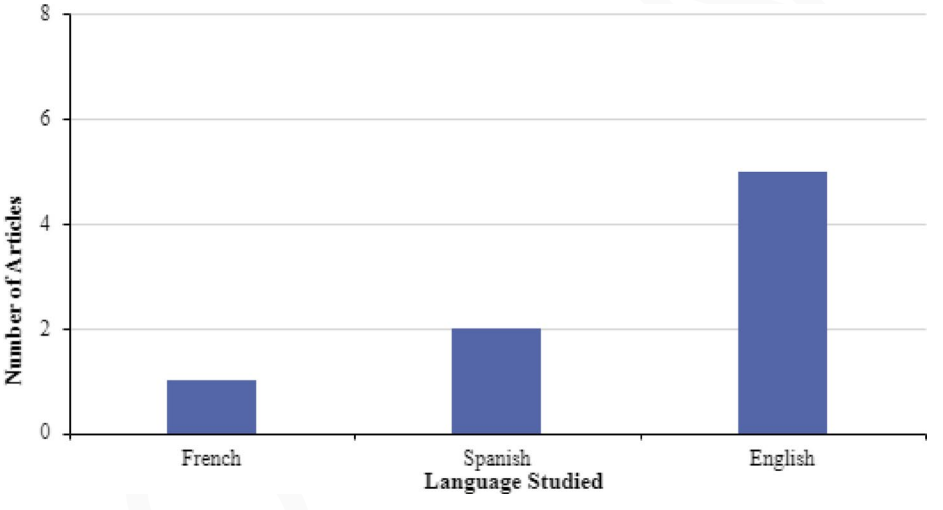

Figure 5. Target language studied.

Countries where studies were completed covered several nations, with seven countries producing the eight studies, two of which occurred in the U.S. (25%). The specific linguistic skills measured varied between studies, with three (37.5%) focusing on vocabulary, one (12.5%) on pronunciation, one (12.5%) on more than two skills, and one each for grammar (12.5%), oral fluency (12.5%), and 'listening and reading' (12.5%). Participants in seven (87.5%) of those studies were college students and app usage was completed solely in class for five (62.5%) of the eight studies. The length of studies varied from two weeks to four consecutive semesters. If the study had followed the four-week minimum used by Burston (2015), the number of articles included would have dropped to six (75%). It appears that mobile phones were the main

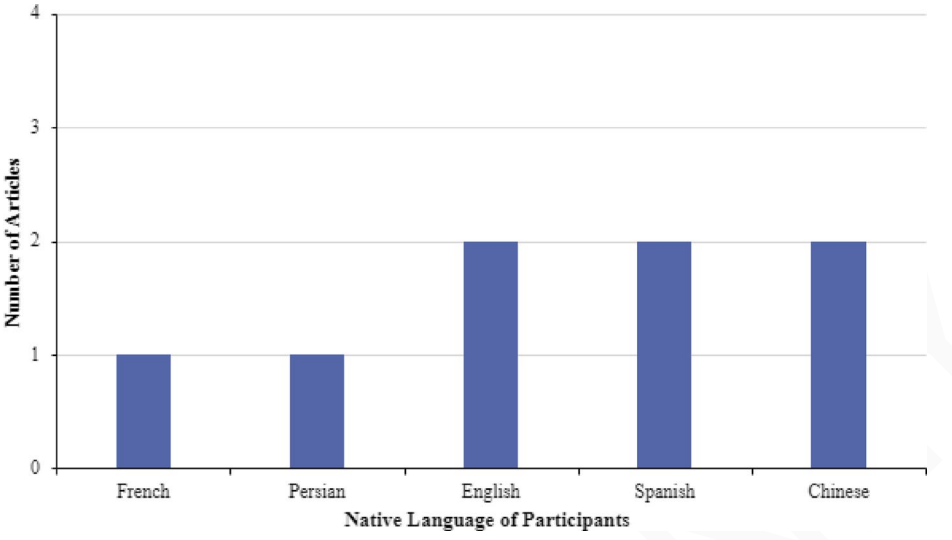

Figure 6. Native language of participants.

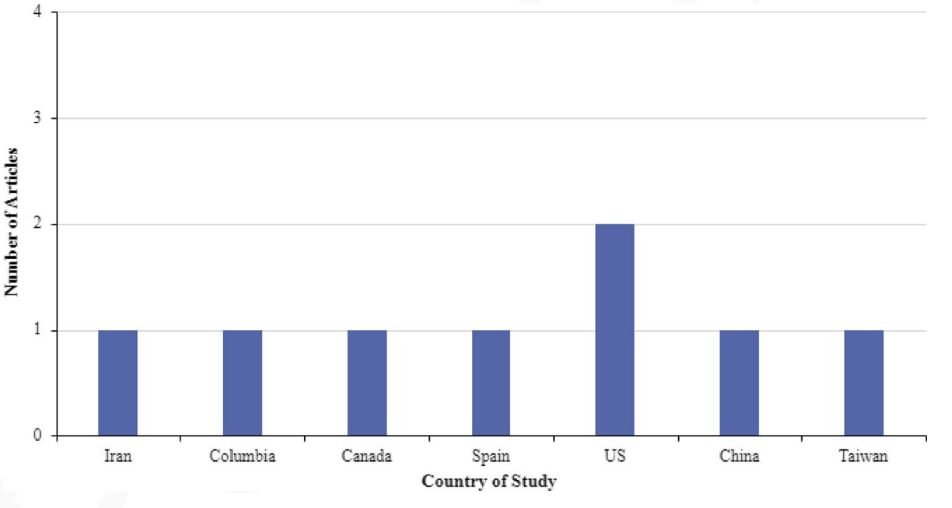

Figure 7. Country where research was carried out.

device used compared to tablets, but this is not certain as some studies did not specify the tool used.

With only eight studies meeting the systematic review's criteria, it is possible to present them individually to help better understand the differences and similarities between them. Their results are in Table 2.

García Botero et al. (2019) Duolingo study divided 52 Columbian college students into two experimental groups and one control group. The control group did not use the app and received traditional classroom instruction only. Both experimental groups were provided with an introduction to the app in addition to traditional classroom instruction, but

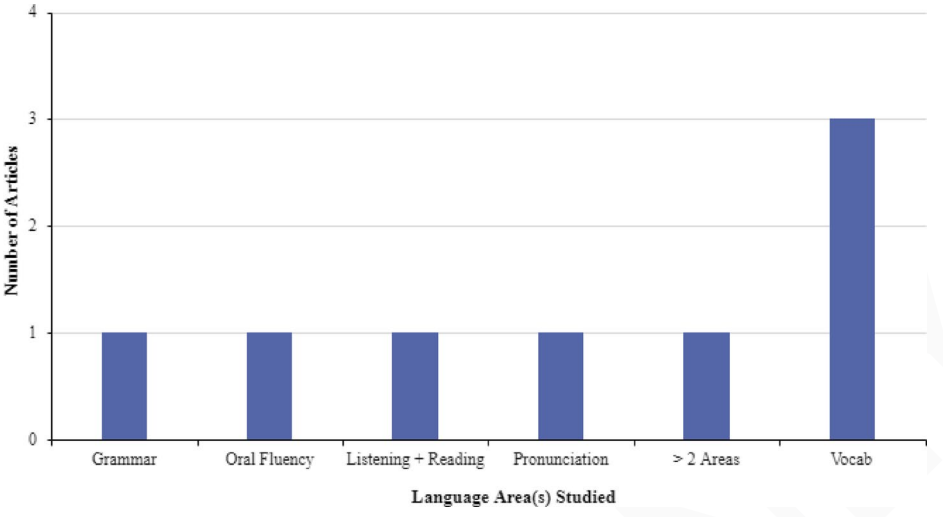

Figure 8. Language area(s) studied.

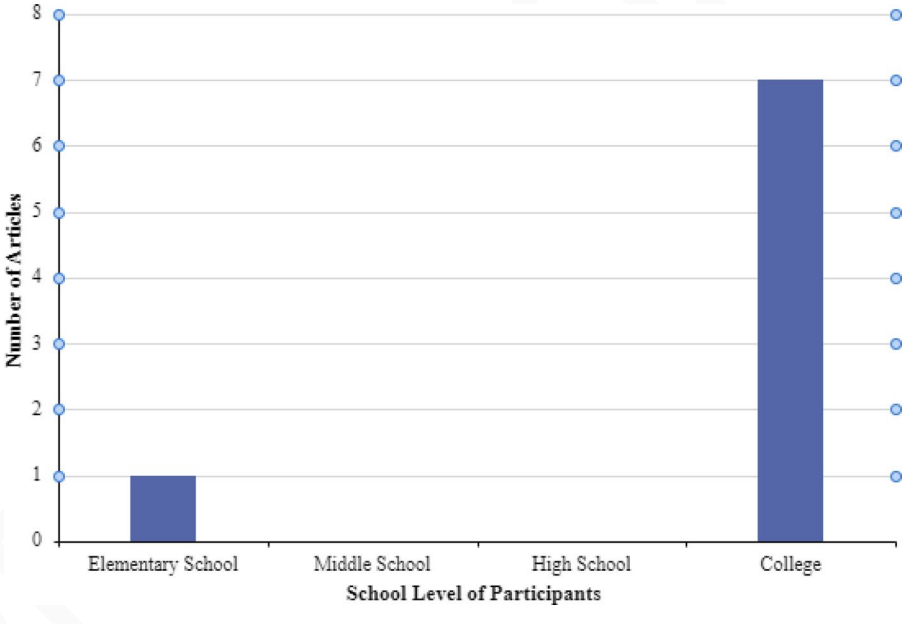

Figure 9. School level of participants.

they varied in that one experimental group, the 'self-regulation and scaffolding group' received instruction about self-regulation strategies. This group was supported for self-regulation through weekly subgoals and scaffolding by the instructor and/or the service desk. After ten weeks, no significant differences were found between the control and experimental groups on listening, reading, or writing, ($p > .05$). Students in the experimental group with self-regulation and scaffolding performed better in writing than the regular experimental group ($p < .05$).

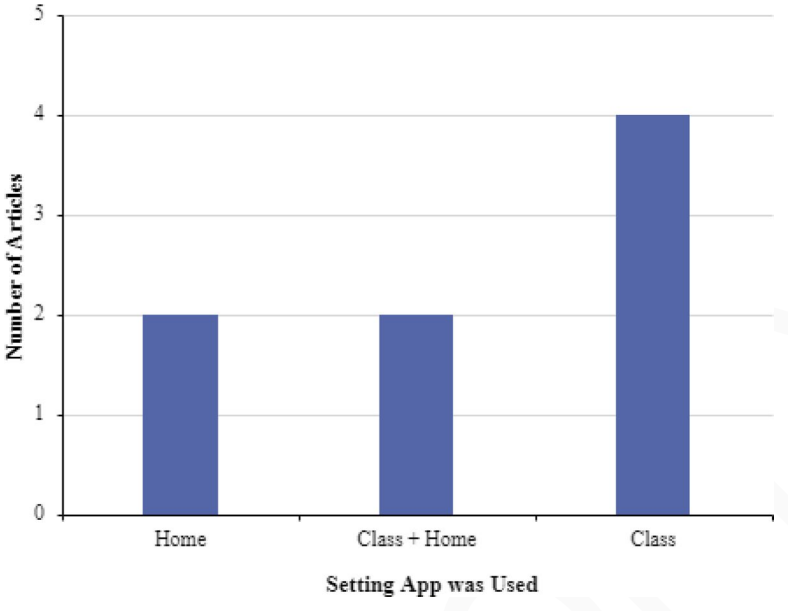

Figure 10. Setting of app use.

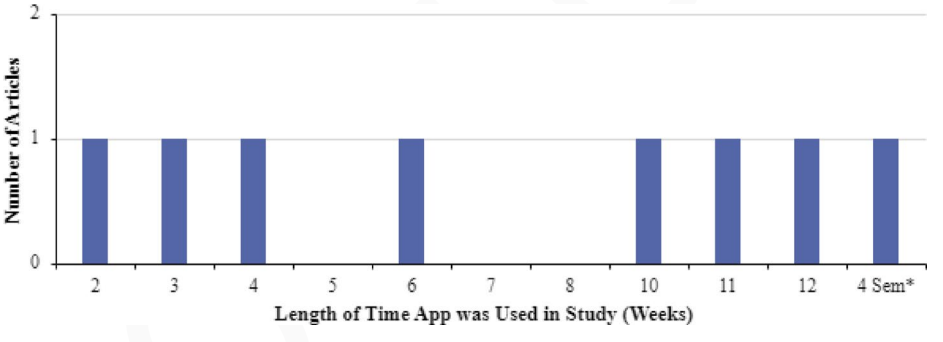

Figure 11. Studies' length of time.

The other Duolingo study (Rachels & Rockinson-Szapkiw, 2018) divided 164 US-based third and fourth grade students into an experimental and control group. For a period of 12 weeks, the experimental group only received Duolingo and the control group received traditional face-to-face teaching methods. Note the curriculum for the control group was adapted to mirror the content of Duolingo. Language knowledge was tested with a researcher-created tool that had vocabulary as its focus with the addition of some grammar-based test items. The results showed there were no statistically significant differences between the experimental and control group on the test ($p > .05$).

Castañeda and Cho (2016) examined the app Conjugation Nation with 80 U.S. university students enrolled in Spanish courses designed to

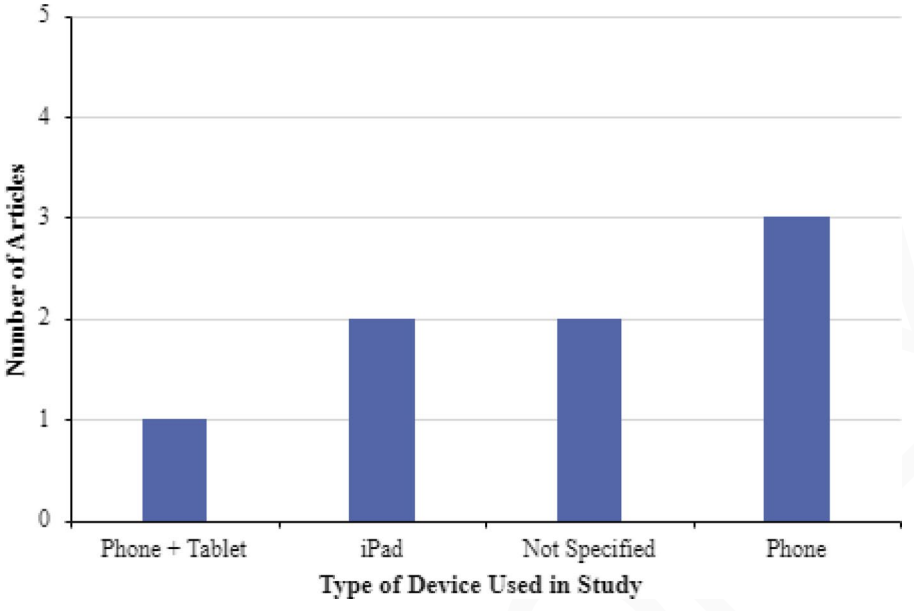

Figure 12. Type of device used in study.

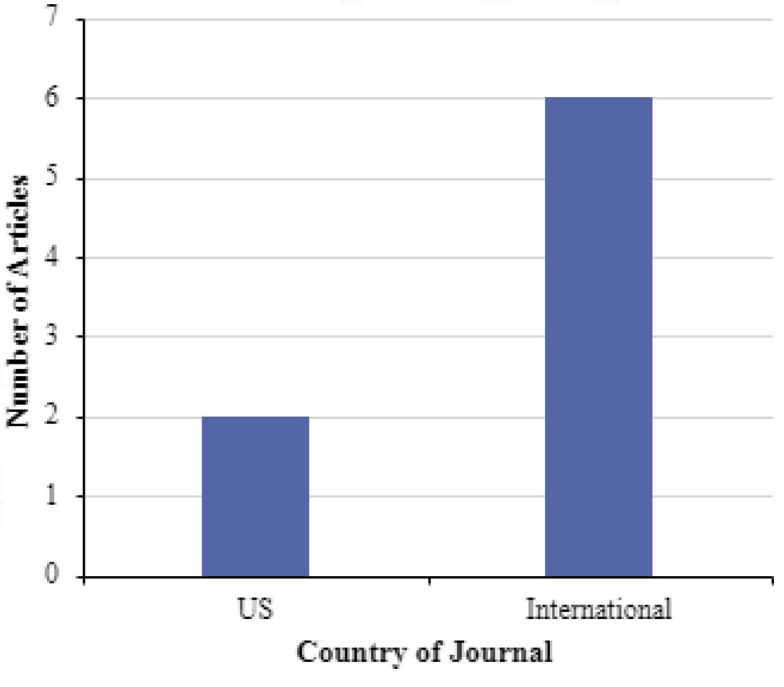

Figure 13. Country of publishing journal.

specifically measure improvement in verb conjugation ability. The app was used by students in pairs or trios during their classes where they also received traditional instruction for four consecutive semesters.

Table 2. Included studies by variables of interest.

Author (Year)	App Name	**Skill(s) Measured**	Country of Study	Language Studied	Outcome
García Botero et al. (2019)	Duolingo	Listening, reading, writing	Columbia	French	No significant differences were found between the control group and experimental groups in any area.
Rachels & Rockinson- Szapkiw (2018)	Duolingo	Vocabulary	U.S.	Spanish	No significant difference between control group and experimental group were found.
Castañeda and Cho (2016)	Conjugation Nation	Grammar	U.S.	Spanish	The group improved skills after traditional teaching and app use.
Ebadi and Ghuchi (2018)	Memrise	Vocabulary	Iran	English	The group using the app performed better than the group using traditional teaching methods.
Fouz-González (2020)	English File Pronunciation	Pronunciation	Spain	English	No significant difference was found between app group and traditional teaching group.
Grimshaw and Cardoso (2018)	Spaceteam EFL	Oral Fluency	Canada (French speaking)	English	No significant difference existed between the groups in either rate of speech or fluency.
Ou-Yang and Wu (2017)	My Eva Mobile	Vocabulary	Taiwan	English	English majors performed better than non-English majors.
Shih (2017)	LINE	Listening Comprehension and Reading Comprehension	Taiwan	English	The group using the app performed better than those using a language laboratory.

Pre- and post-tests allowed participants to act as their own control groups. Results showed statistically significant improvements were made for both present indicative and present subjunctive conjugations. However, the authors were unable to determine whether the app contributed to increased knowledge gains compared to the traditional teaching methods.

Memrise was the app studied by Ebadi and Ghuchi (2018) regarding English vocabulary learning. For four weeks, 40 college students in Iran were divided into a control group that received traditional face-to-face instruction and an experimental group that received the same instruction as the control group plus the use of Memrise outside of class. The participants completed a 40-question pre- and post-test on vocabulary which showed that students using Memrise scored statistically significantly higher than students who just received face-to-face instruction ($p < .05$). Interviews completed with a subsection of participants indicated they used the app between two and three and a half hours per week during their free time.

Fouz-González (2020) studied the English File Pronunciation App's ability to improve the pronunciation skills of 52 Spanish learners studying English at a university in Spain. At the beginning of the study, both an experimental and a control group were formed. Pre- and post- tests were given to both groups in three areas, perception, identification, and production two weeks later. Unfortunately, apart from the control group not using the app for the first two weeks, no information was given regarding whether the control group received any teaching between the pre- and post-test. To provide the same educational opportunities for all, the researchers changed the control group into an experimental group after two weeks, but only for the perception tasks. Results were mixed. No significant difference was found in the discrimination tasks, although there was improvement for both groups. The experimental group, however, improved significantly more than the control group in identification. Note that once the control group became the experimental group, their scores in identification improved by a factor of three compared to when they were not using the app. In the production tasks, the experimental group showed a significant difference in post-test scores for sentence reading and picture description, but not for imitation, although within the imitation task, a significant improvement was seen for the phoneme/ae/. The researchers did not complete a post-test on the new experimental group "to avoid imposing excessive demands on participants" so no results are available.

Grimshaw and Cardoso (2018) asked whether the Spaceteam ESL app could improve oral fluency as measured by syllables produced per minute and judge's ratings. Twenty French speaking Canadian university students were divided into two groups, a control group that was taught via

traditional face-to-face methods, and an experimental group which started each class with 15 minutes of app use over six weeks. Participants using the app worked in teams and were required to provide clear information to each other during the game. Pre- and post-tests showed no significant difference between the groups in either rate of speech or fluency as measured through judges' ratings, although the experimental group did score statistically significantly higher on their delayed post-test compared to their pre-test ($p < .05$).

Shih (2017) studied the effectiveness of the LINE app on English listening comprehension and reading comprehension with 72 university juniors in Taiwan. The control group received traditional English instruction and participated in a language lab over the course of ten weeks while the experimental group received direct instruction and used the mobile app. Pre- and post-tests showed that the experimental group performed significantly better than the control group.

Finally, Ou-Yang and Wu's study (2017) was different from the others described above in that it did not compare groups with and without an app. Instead, it used the same app, MyEva Mobile, to compare different groups on their vocabulary learning. The two groups consisted of 55 English majors and 53 non-English majors at a university in Taiwan. Results showed that while both groups improved after using the app for two weeks, the English majors improved significantly more than the non-English majors.

Discussion

The goal of this study was to address our five research questions. A discussion of each research question follows in addition to a more general discussion of efficacy, methodology, limitations, and recommendations.

The first research question asked how many scientifically reputable studies exist on FLL app efficacy. The most important finding linking the current study to the efficacy studies of Sung et al. (2015) and Burston (2015) was discovering of the extreme paucity of rigorous studies completed on the efficacy of mobile apps for both FLL and MALL in general. Burston (2015) laments that "statistically reliable measures of learning outcomes are few and far between" in MALL (p. 16) while Sung et al. (2015) states that there is a "dearth of review research into the effectiveness of MALL and mobile-device-assisted teaching" (p. 8). As FLL apps are a subcategory of MALL, these statements are even more representative of app efficacy. The fact that only eight studies met this review's criteria has important implications for what we know about FLL app efficacy which will be explored later in the discussion.

The second question investigated what general trends exist in the research. Overall, the publication of FLL app efficacy studies began with only one study in 2016 with no real growth to 2020. Studies were spread internationally with seven countries accounting for the publication of the eight included articles. The amount of time the studies covered ranged from two weeks to four semesters, thereby precluding any information about using apps for long-term language learning.

While Sung et al.'s review (2015) found that elementary school-age students were the most commonly studied group in MALL, the current study of FLL apps found college students to be more frequently represented, which was closer to Burston's (2015) MALL review which found that the most common age group of participants was 25-49 years old. Burston and Sung et al.'s studies were both in the field of MALL and represented the same approximate timeframe. However, Sung et al. included 44 studies which comprised of peer-review journal articles, unpublished conference presentations, and doctoral dissertations, whereas Burston's study used the MALL implementation database as a source, also allowing for publications such as conference proceedings, but with stringent inclusion criteria. The current study differed from these in that it limited the area of inquiry from MALL to FLL mobile apps and required studies to be published in peer-review journals. Given the small number of studies meeting the criteria, it may be more meaningful to consider the field of MALL when gathering information on specific age groups that have been studied. Note that Sung et al. had college students listed as the second most common group of participants studied after elementary students with percentages of 26.7 and 40 respectively.

The third research question focused on which languages were taught and to what linguistic audiences. Native speakers of five different languages, French, Persian, English, Spanish and Chinese, were participants in the studies. Concurring with Sung et al. (2015) findings that English as a target language was the most studied, five of the eight studies focused on English learning with the other languages being Spanish and French. This is representative of English being the most studied foreign language worldwide (Serrato & Rodriguez, 2020). However, it is concerning that even for English, only five studies met our criteria. The problem lies herein; if it takes several studies to show that a particular mobile app is effective in teaching certain areas of language, this may only apply to learners who are at a certain level. Add to this the fact that each app needs to be tested for efficacy in several areas of language such as pronunciation, grammar, and listening comprehension. Next, imagine that research shows that a company has an app that is effective in teaching English to native Spanish speakers. It will not automatically follow that another version of this app, such as one teaching English

to Japanese speakers or teaching Russian to English speakers will be as effective. The amount of research required to evaluate the efficacy of a wide variety of mobile apps in FLL is nearly non-existent compared to what is needed.

Although we could have considered changing our inclusionary criteria to allow for the inclusion of a wider selection of studies regarding FLL app efficacy, we decided to keep our original inclusionary criteria for two reasons. First, Burston maintains high demands for rigorous research, although we did omit two criteria deemed unnecessary. Second, the nature of systematic reviews requires selection criteria to be finalized and possibly published before the study is carried out, ensuring that methods are not changed after the fact. Moher et al. (2015) stated that "Ideally, systematic reviews are based on pre-defined eligibility criteria and conducted according to a pre-defined methodological approach" (p. 1).

Regarding the fourth research question of what specific aspects of linguistic knowledge and skills were measured, the current study concluded, similar to Dehghanzadeh et al. (2019), Heil et al. (2016), Sung et al. (2015), Burston (2015), and Duman et al. (2015), that vocabulary was the most represented language area studied in FLL, both by apps and overall MALL. Out of eight studies meeting the criteria for inclusion, three had vocabulary as its only measure, two of which measured improvement in English vocabulary for participants speaking Arabic (Ebadi & Ghuchi, 2018) and Chinese (Ou-Yang & Wu, 2017), and one which examined the acquisition of Spanish vocabulary for participants in the U.S. While most FLL mobile apps and MALL programs were not limited to teaching vocabulary, results showed that this area was the most often measured. Although vocabulary is certainly necessary to learn a language, it is but one area of language. Perhaps the reason for this focus on vocabulary learning is the relatively easy possibility of measuring the acquisition of a vocabulary item in a binary fashion within an app, such as measuring whether a word was correctly chosen from a group or produced from memory. It is certainly more difficult to develop a study that measures the correctness of pronunciation, for example, as measures are somewhat subjective and progress is made in increments, not in a binary fashion of being correct or incorrect. Grammar is also a relatively difficult area to measure, given the high number of grammatical components that exist and also interact with each other. Although a specific exercise such as forming the past tense of a given verb could be assessed with a binary scoring system, getting an overall measure of grammatical skill is a complex task. Similarly, skills such as oral fluency, listening comprehension, and reading ability are likely to be considered as more daunting to measure. It is also

possible that the explicit nature of vocabulary learning provides app users with the feeling of greater face validity and perceived efficacy.

Although few apps are limited to vocabulary learning, Heil et al. (2016) reported in their study of a selection of 50 mobile apps available in Google Play or the Apple iTunes App Store, that 42 out of the 50 emphasized teaching vocabulary items as isolated units without context. This is tempered by the fact that 23 of the 50 were also found to emphasize vocabulary in context, meaning that only 19 of the 50 were limited to isolated vocabulary items. The researchers asserted that this was still problematic as language is a communicative tool that requires learners to know how to meaningfully use the words they have learned.

One possible way of addressing this in future research is to find measures of full language competency including various areas of language knowledge and skills which could be used in any FLL app, thereby providing a basis for comparison. Suggested guidelines for foreign language assessment have been proposed by groups such as the European Association for Language Testing and Assessment (EALTA, 2006), the Japanese Language Testing Association (JLTA, 2002), and the International Language Testing Association (ILTA, 2007). Evaluating specific guidelines in terms of mobile apps is beyond the realm of this discussion but we encourage the FLL community to further explore this area.

The fifth and final research question asked what the efficacy outcomes of these studies were. Considering the very small number of studies examined in this review, an overall analysis of the studies' outcomes does not lead to a simple interpretation. First, we examined the efficacy of the two Duolingo studies to try to make a direct comparison. However, each study focused on different areas of language (listening, reading, and writing versus vocabulary), tested different age groups (college students versus third and fourth graders) with different native languages (Spanish versus English) while attempting to learn different languages (French versus Spanish). The studies also greatly differed in research design with one analyzing whether self-regulation and scaffolding training used within the app was more effective compared to the app without self-regulation training. In contrast, in the other study, the control group received traditional teaching while the experimental group only used the app for foreign language learning. García Botero et al. (2019) concluded that app use in addition to traditional teaching was not useful in terms of learning measures unless combined with self-regulation and scaffolding activities. However, Rachels and Rockinson-Szapkiw (2018) study indicated that Duolingo was equally effective as a traditional curriculum. Caution must be used when interpreting the results. Their study limits its evaluation of Duolingo to vocabulary although the app teaches other aspects of language too. This limits the study's claims to

one of many language areas. Note that the face-to-face curriculum of the control group was based on the content of Duolingo, thereby asking the question of whether the Duolingo vocabulary content could be taught better through traditional instruction or through the app. In this case, the answer appears to be the latter. Furthermore, the study uses a convenience sample as opposed to a randomized sample, making it impossible to say that the results are generalizable to other groups. Strong differences in study design also made it difficult to compare the only two studies that exist for the same app.

It should be noted that while the ACT Framework promoted measurement of efficacy in the skills the instructional tool was designed to improve (Mattern, 2019), commercial apps were not designed by the researchers and generally do not come with information about what specific areas of language-learning they target. This means that the failure of an app to improve learning outcomes in a specific area such as vocabulary in no way means that the app is not useful in other areas of language-learning.

The range of methodologies used continued to grow with the number of studies examined. While the LINE, Memrise, and Spaceteam EFL apps were relatively similar in that they compared app use to traditional teaching practices, others had very different methods and goals. For example, students using Conjugation Nation all received face-to face instruction along with app use, and acted as their own controls, but without the possibility of attributing success to a particular part of the instruction. The study of MyEva Mobile differed from all the others by comparing different groups using the same app, in this case English majors and non-English majors, to see if one was more successful than the other, thereby telling us relatively little about whether the app was successful relative to traditional teaching. Furthermore, it is possible that the English-majors had more English-language input before and during the experiment, leading to this group having more exposure to the vocabulary items which in turn led to higher vocabulary scores (Peters, 2018). It should be noted that in several studies, the researchers gave participants instructions on how to use the apps, five out of eight studies were used only during class and some were used with students working in groups. This is certainly a different context compared to that of an average consumer who may download an app on a tablet device and then use it alone at times of their choice.

The overall results of efficacy studies will be important because as more research is carried out in this area, consumers will be well served by scientific information about the efficacy of FLL apps. Studying a foreign language is time consuming, therefore, app users may be interested in knowing which apps have been shown to be effective. Also,

very specific efficacy information could be useful for individuals who want to focus on a specific area of language such as vocabulary instead of working on more global areas. For instance, a focus on listening as a first skill is recommended by Gangaiamaran and Pasupathi (2017) who make the case that this is the first skill that infants use during the acquisition of their native language. They also correctly note that listening to a foreign language after achieving a certain level of proficiency is even more difficult due to the natural instinct to focus attention not only on the sounds of the language, but also on the meaning they carry. This leads to the interesting question of whether in some specialized cases, apps focusing on a single skill such as listening may be pedagogically desirable, leading to ensuing app use combining the teaching of several linguistic skills. As the acquisition of several linguistic skills is necessary to create fluency in a language, we recommend interdisciplinary collaborations between app designers, linguists, and academic researchers to develop more efficacious FLL apps and improve FLL app quality for consumers.

One more group that will be informed by these results are researchers in MALL. Hopefully, the realization of the extremely limited number of studies that met the criteria of this systematic review will spur the field to design studies that are scientifically rigorous and to publish them in peer-reviewed journals. Database searches turn out very high numbers of publications on FLL apps, but the majority of them, although many are likely to be of high quality, are published in other venues, lacking the gold standard that these journals have to offer.

Limitations

Limitations of this systematic review must be acknowledged. First, no review of this type can claim to be completely exhaustive since search terms and databases must be limited by some constraints, some of which are due to be imperfect. We also acknowledge that several efficacy studies exist that are not listed here. For example, a 2012 final white paper on the efficacy of Duolingo by Vesselinov and Grego (2016) reported findings from an experiment of 196 adults studying Spanish for eight weeks. Results claimed that a person with no knowledge of Spanish would be able to reach a degree of knowledge of the language equivalent to a college semester after an average of 34 hours of study with the app. However, this article was not published in a peer-reviewed academic journal. Second, as noted by Haidich (2010), a meta-analysis requires that studies use the same designs, have similar participants, and have the same outcomes. In our case, since studies had different outcomes, used different participants, different mobile apps, a

meta-analysis would not be possible. Therefore, we completed a systematic review and were unable to compare effect sizes between studies.

Recommendations

Based on this study, there are multiple recommendations. The *first recommendation has two parts*. First, FLL app designers should be encouraged to build apps that focus on communicative ability as a whole. The strong focus on vocabulary teaching and measurement has been shown repeatedly. Out of the eight studies discussed here, three of them focused on vocabulary teaching. Dehghanzadeh et al. (2019) research, which examined 22 FLL games, found that 17 were focused on vocabulary. Vocabulary was also the most commonly studied area in Sung et al. (2015) MALL efficacy meta-analysis and Burston's (2015) review of MALL learning outcomes. Heil et al. (2016) reviewed the 50 most popular FLL apps, to find that 42 taught vocabulary in isolation. This review also showed vocabulary to be the most tested area. This suggests that many commercial FLL apps focus on vocabulary, often independently of other language areas. The thesis that vocabulary should not be taught in isolation was put forward not only in pedagogical terms but in psycholinguistic terms by Bolgün and McCaw (2019) who evaluated language technology in light of our understanding of the neuroscience of memory, particularly the declarative and procedural memory systems. They argue that "language technology caters predominantly to the declarative memory system" involved in explicit memorized knowledge, whereas it should "cater to the procedural memory" system involved in implicit grammatical processes like grammar and sentence structure (Bolgün & McCaw, 2019). The idea is that explicit knowledge of vocabulary does not equate to an implicit understanding of its use conversationally.

The second part of this recommendation is closely related to the first. Tools measuring the overall communicative competence of foreign language learners must be developed and used both within and across studies to compare app efficacy in meaningful ways. Just as standardized IQ tests provide general measures of intelligence, standardized language tests would allow researchers to more easily compare results across future studies. It is important to note, however, that just because studies may choose to examine a single aspect of language, such as vocabulary, does not mean the app is limited to that aspect. Developing measures of overall communicative efficacy would be an important step forward in expanding the field of FLL efficacy.

Our next recommendation is for FLL researchers to measure learning outcomes in a scientific, rigorous manner. Perhaps the most important finding from this systematic review is the realization that despite the

large number of returned results when searching for studies about FLL apps, only a tiny percentage of these measured efficacy and met the criteria of this study, which were more relaxed than those put forward by Burston (2015). This means that the number of high-quality scientific studies measuring the effect that commercial FLL apps have on learning outcomes is shockingly small. As stated above, this needs to change for the good of FLL students, app consumers, app developers, and FLL instructors. Given the large number of commercially available FLL apps, there is much room for further exploration into their efficacy. While it is certainly valuable to examine these apps in other contexts such as exploring user enjoyment, motivation, and usage trends, it is vital to the field that the efficacy of MALL tools, in this case apps, is measured in systematic ways that allow studies to be compared.

We also recommend that studies be developed that compare several different experimental groups, each using a different FLL app only, to control groups using clearly defined traditional instruction. Measures would be taken globally or by testing different areas of language including pronunciation, listening comprehension, vocabulary, and grammar. This would be of use to consumers who have decided to use an app but need assistance in deciding which one(s) to choose. Also useful, particularly for FLL educators, would be studies where different apps are combined with identical traditional methods. Furthermore, comparisons of app efficacy when instructions are given versus when the user is left to their own devices would be helpful. This would allow FLL instructors using apps as part of their teaching to know whether instruction in app use is to be recommended. In summary, there is an opportunity in the field to continue researching the efficacy of mobile applications so educators, students, and consumers can be better informed when making decisions about applications designed to teach foreign languages.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Chrystal Sapphire Dragonflame http://orcid.org/0000-0001-6688-5807
Amanda A. Olsen http://orcid.org/0000-0002-7707-7271

References

Babbel. (2020, December 21). *About us.* https://about.babbel.com/en/about-us/

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bolgün, M. A., & McCaw, T. (2019). Toward a neuroscience-informed evaluation of language technology. *Computer Assisted Language Learning*, *32*(3), 294–321. https://doi.org/10.1080/09588221.2018.1516675

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

Castaneda, D. A., & Cho, M. H. (2016). Use of a game-like application on a mobile device to improve accuracy in conjugating Spanish verbs. *Computer Assisted Language Learning*, *29*(7), 1195–1204. https://doi.org/10.1080/09588221.2016.1197950

Chan, J. (2019, January 31). Top language apps worldwide for 2018 by downloads. *Sensortower*. https://sensortower.com/blog/top-language-apps-2018-ww

Clearinghouse, W. W. (2012). What works clearinghouse. *Internet site*. http://ies.ed.gov/ncee/wwc.

Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2019). Using gamification to support learning English as a second language: A systematic review. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2019.1648298 Q7

Doabler, C. T., Clarke, B., Kosty, D., Turtura, J. E., Firestone, A. R., Smolkowski, K., … Maddox, S. A. (2019). Efficacy of a first-grade mathematics intervention on measurement and data analysis. *Exceptional Children*, *86*(1), 77–94. https://doi.org/10.1177/0014402919857993

Duman, G., Orhon, G., & Gedik, N. (2015). Research trends in mobile assisted language learning from 2000 to 2012. *ReCALL*, *27*(2), 197–216. https://doi.org/10.1017/S0958344014000287

Duolingo. (2020, December 21). *Free language courses for English speakers*. https://www.duolingo.com/courses

Dynarski, M. (2015). Using research to improve education under the Every Student Succeeds Act. *Evidence Speaks Reports*, *1*(8), 1–5.

Ebadi, S., & Ghuchi, K. D. (2018). Investigating the effects of blended learning approach on vocabulary enhancement from EFL learners' perspectives. *Journal on English Language Teaching*, *8*(2), 57–68. https://doi.org/10.26634/jelt.8.2.13981

Fouz-González, J. (2020). Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation app. *Language Learning & Technology*, *24*(1), 62–85. http://hdl.handle.net/10125/44709

Gangaiamaran, R., & Pasupathi, M. (2017). Review on use of mobile apps for language learning. *International Journal of Applied Engineering Research*, *12*(21), 11242–11251. https://doi.org/10125/44709

García Botero, G., Botero Restrepo, M. A., Zhu, C., & Questier, F. (2019). Complementing in-class language learning with voluntary out-of-class MALL. Does training in self-regulation and scaffolding make a difference? *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2019.1650780 Q8

Goldman, S. R., Greenleaf, C., Yukhymenko-Lescroart, M., Brown, W., Ko, M. L. M., Emig, J. M., … Britt, M. A. (2019). Explanatory modeling in science through text-based investigation: Testing the efficacy of the Project READI intervention approach. *American Educational Research Journal*, *56*(4), 1148–1216. https://doi.org/10.3102/0002831219831041

Grimshaw, J., & Cardoso, W. (2018). Activate space rats! Fluency development in a mobile game-assisted environment. *Language Learning & Technology*, *22*(3), 159–175. https://core.ac.uk/download/pdf/211326392.pdf

Haidich, A. B. (2010). Meta-analysis in medical research. *Hippokratia*, *14*(Suppl 1), 29–37. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049418/

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/10.4995/eurocall.2016.6402

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798.

Mattern, K. (2019). *ACT's efficacy framework: The intersection of learning, measurement, and navigation. issue brief*. ACT, Inc.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1–9. https://doi.org/10.1186/2046-4053-4-1

Q9

Olsen, A., Dragonflame, C., & Tommerdahl, J. (Under Review). *A systematic review*.

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

Pandey, M., Litoriya, R., & Pandey, P. (2019). Perception-based classification of mobile apps: A critical review. In A. K. Luhach, K. B. G. Hawari, I. C. Mihai, P. Hsiung, & R. B. Mishra (Eds.), *Smart computational strategies: Theoretical and practical aspects* (pp. 121–133). Springer. https://doi.org/10.1007/978-981-13-6295-8_11

Peters, E. (2018). The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. *Itl - International Journal of Applied Linguistics*, *169*(1), 142–168. https://doi.org/10.1075/itl.00010.pet

Purpura, J. E. (2013). Assessing grammar. *The Companion to Language Assessment*, *1*, 100–124. https://doi.org/10.1002/9781118411360.wbcla147

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*(1-2), 72–89. https://doi.org/10.1080/09588221.2017.1382536

Rovio Entertainment Corporation. (2020, March 2). *About - Rovio*. https://www.rovio.com/about/

Serrato, D. I., & Rodriguez, B. C. P. (2020). Academic e-tandems as a strategy for English language learning in a Mexican university. *Open Praxis*, *12*(3), 417–424. https://doi.org/10.5944/openpraxis.12.3.1099

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & Group, P.-P. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, *349* https://doi.org/10.1136/bmj.g7647

Q10

Shih, R. C. (2017). The effect of English for Specific Purposes (ESP) learning-language lab versus mobile-assisted learning. *International Journal of Distance Education Technologies (IJDET)*, *15*(3), 15–30. https://doi.org/10.4018/IJDET.2017070102

Sriganesh, K., Shanthanna, H., & Busse, J. W. (2016). A brief overview of systematic reviews and meta-analyses. *Indian Journal of Anaesthesia*, *60*(9), 689–694. https://doi.org/10.4103/0019-5049.190628

Sung, Y. T., Chang, K. E., & Yang, J. M. (2015). How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review*, *16*, 68–84. https://doi.org/10.1016/j.edurev.2015.09.001

Toste, J. R., Capin, P., Williams, K. J., Cho, E., & Vaughn, S. (2019). Replication of an experimental study investigating the efficacy of a multisyllabic word reading inter-

vention with and without motivational beliefs training for struggling readers. *Journal of Learning Disabilities*, *52*(1), 45–58. https://doi.org/10.1177/0022219418775114

Toto, G. A., & Limone, P. (2019). Contemporary trends in studies on mobile learning of foreign languages: A meta-analysis. *International Journal of Engineering Education*, *1*(2), 85–90. https://doi.org/10.14710/ijee.1.2.85-90

Vaughn, S., Martinez, L. R., Williams, K. J., Miciak, J., Fall, A. M., & Roberts, G. (2019). Efficacy of a high school extensive reading intervention for English learners with reading difficulties. *Journal of Educational Psychology*, *111*(3), 373. https://doi.org/10.1037/edu0000289

Vesselinov, R., & Grego, J. (2016). *The Babbel efficacy study: Final report* [White paper]. Babbel. https://press.babbel.com/fr_CA/releases/downloads/Babbel-Efficacy-Study.pdf

Wanzek, J., Vaughn, S., Roberts, G., & Fletcher, J. M. (2011). Efficacy of a reading intervention for middle school students Identified with Learning Disabilities. *Exceptional Children*, *78*(1), 73–87. https://doi.org/10.1177/001440291107800105

Yu, Z. (2019). A systematic review on mobile technology-assisted English learning. *International Journal of e-Collaboration*, *15*(4), 71–88. https://doi.org/10.4018/IJeC.2019100105