Video-Game Based Instruction for Vocabulary Acquisition with English Language Learners: A Bayesian Meta-Analysis

Nigora Abdiyeva[1], Jurabek Abdiyev[2*]

[1]Samarkand State Institute of Foreign Languages. Samarkand, 140104, street Bustonsaroy 93
[2]Physical-technical Institute of NPO "Physics – Sun" of Uzbekistan Academy of Sciences Uzbekistan, Tashkent, ChingizAitmatov street 2B.
Corresponding author: fiztexabdiev@gmail.com(J. Abdiyev)

Abstract: *This meta-analysis reviewed the literature on the efficacy of game-based learning in English as a second language vocabulary acquisition. A systematic search of the literature produced 19 studies that met inclusion criteria. Using Bayesian methods and 20 standardized-mean-difference effect sizes we assessed 1) overall mean effects and between-studies variability, 2) subgroup analyses (grade level, sex, hardware, game type, intervention length, Foreign Services Institute level, allocation, and publication type), and 3) risk of publication bias. The overall effect was moderately large, indicating favorability to the video-game based learning groups. We found evidence of effect-size heterogeneity with a large between-studies standard deviation, along with large Q and I^2 values. Subgroup analyses produced mixed results when partially explaining effect-size variability. Furthermore, all publication bias tests indicated a low risk of publication bias. We found that video-game based learning can make a significant difference in promoting English vocabulary acquisition and that integrating entertainment video games into educational contexts can result in substantial student learning gains.*

Keywords: Vocabulary language acquisition, Game-based learning, English language learners, Bayesian meta-analysis, Digital games, Video games, Second language acquisition.

1. DIGITAL GAMES IN EDUCATION.

Clark et al. (2016) conducted a meta-analysis examining the efficacy of digital games in K-16 educational contexts. They compared digital game vs. non-game treatments and found that digital game treatments outperformed non-game treatments (g = 0.33, 95% confidence interval [0.19, 0.48], K = 57). Additionally, Clark and colleagues found that select moderator variables (e.g., number of game sessions, scaffolding, and camera view) impacted learning outcomes while others did not (e.g., story depth, variety of game actions, and whether the game includes additional non-game instruction). Overall, Clark et al. (2016) demonstrated the efficacy of digital game instruction compared to non-game instruction.

Wouters et al. (2013) conducted a meta-analysis on the efficacy of serious games in comparison to conventional instructional approaches. A serious game is a game that is designed for purposes other than pure entertainment (Michael & Chen, 2005); for example, the existence of a goal, such as education or training, that is an underlying purpose of the game. Wouters et al. (2013) found that while serious games led to greater learning outcomes (d = 0.29, p < .01, K = 77), such games were not significantly more motivating than conventional instruction (d = 0.26, p > .05, K = 31). The lack of enhanced motivation is a curious finding, given that enhanced motivation and engagement are often cited as two primary reasons for the efficacy of digital games in learning environments (Garris et al., 2002; Prensky, 2003; Sung & Hwang, 2013).

Clark et al. (2016) commented on this phenomenon, stating that their own meta-analysis found enhanced intrapersonal learning outcomes, which “not only included motivation but also included intellectual openness, work ethic and conscientiousness, and positive core self-evaluation” (p. 108). Thus, while Clark et al. (2016) did not examine motivation specifically, as was the case with Wouters et al. (2013), they did find enhanced intrapersonal learning outcomes more broadly. These findings may be affected by the fact that Wouters et al. (2013) focused on serious games while Clark et al. (2016) included digital games in general, combing both serious games and commercial-off-the-shelf (COTS) games, which are games designed primarily for entertainment purposes, though they can still present “intellectual challenges and content” (Charsky & Mims, 2008, p. 38). Thus, comparing outcomes between serious and non-serious games may illuminate similarities and differences between these two game types, something the present meta-analysis examines.

As the above meta-analyses illustrate, digital games can lead to positive learning outcomes in various contexts. While meta-analyses on digital games in education are important contributions to the research literature, examining specific fields and content brings its own benefits by creating a tailored focus on the efficacy of digital games in a particular area of study. One context that deserves further examination, and which a meta-analytic review would be a valuable addition to the literature, is the use digital games for teaching English to speakers of other languages (TESOL).

1.1. DIGITAL LEARNING IN SECOND LANGUAGE ACQUISITION.

While research reviews of video games in language learning are limited (Peterson, 2010), more attention has been given to the use of digital technologies in language teaching and learning. Thus, examining this body of research provides background context to the use of digital games in language teaching and learning. Zhao, (2003) conducted one of the first meta-analyses that compared the effectiveness of digital technologies to more traditional analog instruction. Zhao's study (d = 0.81, 95% confidence interval [0.55, 1.07], K = 29) found that language learning outcomes were higher in technology-supported instruction than instruction without technology for studies conducted with college students and adult learners across a variety of language skills, including reading, writing, speaking, listening, grammar, and vocabulary.

Grgurovic, Chapelle, and Shelley (2013) examined the effectiveness of computer assisted language learning and compared groups who received instruction with digital technologies to groups who received non-digital instruction. Their overall mean effect size comparing groups with digital instruction and groups without digital instruction was significantly different from zero (d = 0.235, 95% confidence interval = [0.144, 0.327], K = 49), in favor of the students who received digital instruction.

1.2. DIGITAL GAMES IN SECOND LANGUAGE ACQUISITION.

Peterson (2010) examined the research literature and identified two prominent perspectives on second language acquisition that are also relevant to learning through digital gameplay: psycholinguistic research and sociocultural research. Psycholinguistic research recognizes that people learn languages from being exposed to a language and recognizing patterns of language use (Ellis, Simpson-Vlach, & Maynard, 2008). For example, through language exposure, people begin to notice patterns of how various parts of speech (e.g., nouns, verbs, adjectives, adverbs, prepositions, etc.) are assembled in sentences, and then language learners can begin to use their knowledge of those patterns as they interpret the sentences of others and construct sentences of their own. Given the importance of pattern recognition in psycholinguistic approaches, encouraging students to high frequency and levels of exposure can aid learning (Ellis, 2002).

A sociocultural perspective recognizes, as the name suggests, social and cultural environments and processes as key skills in language learning (Lantolf, 2000). Grounded in the work of Vygotsky (1978), this perspective posits that language learning, like all learning, is mediated by human culture, activities, and artifacts (Lantolf & Thorne, 2007). Thus, it is important for language educators to foster opportunities for language learners to engage in social interactions to help them develop their language skills. Such social interactions and cultural experiences can also facilitate student learning in the zone of proximal development, which is “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers.” (Vygotsky, 1978, p. 86)

While the foci of these two perspectives differs substantially, both perspectives recognize the value of providing meaningful feedback, engaging students in authentic learning experiences, and encouraging collaborative and cooperative activities for students. Furthermore, while some studies utilize one perspective more than the other, many studies cannot be clearly categorized as utilizing one perspective or the other, as studies often have elements of both. For example, a study could include plenty of social interaction amongst participants and include instruction on linguistic patterns as well. Given how it can be difficult to categorize between these perspectives and both offer valuable insights into second language acquisition, this present study does not utilize one perspective or the other. Rather, it seeks to situate the use of digital games in second language vocabulary acquisition in the context of these prominent perspectives.

Peterson (2010) conducted a qualitative review of the literature and found that games supported second language acquisition in a variety of settings for a variety of language skills (e.g., speaking, listening, reading, writing, and vocabulary). Peterson also highlighted a variety of reasons why digital games can support second language acquisition. Among the reasons highlighted were substantial exposure to comprehensible input in the target language (Krashen, 1985), enhanced student motivation (Svensson, 2003), opportunities for authentic interaction in the target language with peers (Bryant, 2006; Peterson, 2006), lowering affective barriers that can negatively impact students’ willingness to interact with their peers (Freiermuth, 2002).

The English language is taught to students around the world to help provide them access to international academic, social, and economic opportunities (Crystal, 2003). As is the case with all languages, different skills need to be learned to communicate in a given language including speaking, listening, reading, writing, grammar, and vocabulary. Researchers in computer assisted language learning and TESOL have investigated the use of digital games to help students learn these language skills.

Researchers have found that digital games can provide engaging environments for social interaction in the target language (Peterson, 2010; Reinders, 2012; Reinders & Wattana, 2015). In terms of listening skills, Suh, Kim, and Kim (2010) found that Korean English as a foreign language (EFL) elementary students who engaged with massive multiplayer online roleplaying games (MMORPGs), along with supplementary materials and activities, improved their listening skills more than the control group who received face-to-face instruction in a pretest-posttest comparison. As for reading and writing in the target language, scholars have also illustrated the potential of MMORPGs, particularly using chat text between players who communicate for a variety of purposes (e.g., friendly banter and strategic planning) while playing online (Peterson, 2012). Vocabulary development is one component of language development that has received more scholarly attention than others, but such research has not yet been studied systematically.

1.3. DIGITAL GAMES AND VOCABULARY LEARNING IN TESOL.

One area of using digital games in TESOL that has received significant attention is vocabulary development (Neville, Shelton, & McInnis, 2009; Smith, Li, Drobisz, Park, & Kim, 2013; Wu & Huang, 2017). This may be because it is easier to utilize and measure digital game-based approaches to vocabulary than other language skills, such as speaking, reading, or grammar. Given the robust literature base in vocabulary acquisition using digital games in TESOL, we examine this literature and demonstrate the benefit of conducting a meta-analytic review of the literature.

A variety of studies have investigated how digital games can facilitate second language (L2) vocabulary learning (Johnson & Valente, 2009; Ranalli, 2008; Turgut & Irgin, 2009). However, not all digital games (and corresponding educational practices) promote learning in the same way. Peterson (2012) found that playing an entertainment-focused MMORPG in conjunction with in-game chat features, participants were exposed to "vocabulary not normally encountered in regular language classes," and participants learned new vocabulary during gameplay (p. 361). Similarly, Turgut and Irgin (2009) found that youth often are exposed to and learn new contextualized vocabulary simply through playing entertainment-focused games in English. These games stand in contrast to the work of other scholars who have examined serious games. For example, Wu and Huang (2017) examined how a mobile vocabulary-focused game can enhance learning for Mandarin Chinese speakers, while Johnson and Valente (2009) illustrate how the Tactical Language and Culture Training System, a serious game primarily used to help U.S. military personnel learn about various languages and cultures around the world, supports vocabulary acquisition.

2. METHODS.

2.1. RESEARCH QUESTIONS.

Overall, there is a wide variety of research examining how games may support English vocabulary development for speakers of other languages. A promising feature of using digital games for vocabulary development is that various researchers have established that vocabulary exposure, use, and acquisition are possible in several contexts through a variety of games. However, a systematic study that investigates the overall efficacy of using digital games to promote vocabulary acquisition for English language learners, as well as determining how digital game features and population characteristics affect vocabulary acquisition, is a necessary addition to the evolving literature. To fulfill this purpose, we answer the following research questions:

1. What are overall quantitative characteristics of video-game based instruction studies?
   a. What is the overall effect of video-game based instruction on EFL vocabulary acquisition compared to non-video game-based instruction?
   b. How heterogeneous are effects from studies on the effectiveness of video-game instruction of EFL vocabulary acquisition?
2. Do select study characteristics (grade level, sex, hardware, game type, intervention length, Foreign Services Institute (FSI) level, allocation, and publication type) moderate the EFL vocabulary acquisition effect?
3. What is the risk of publication bias within the collection of video-game based instruction studies?

2.2. SEARCH PROCESS.

Digital searches used ERIC, PsycINFO, ProQuest, and Web of Science databases to collect relevant studies. Search keywords included all 15 combinations of game terms (digital games, video games, and game-based learning) and language acquisition terms (English language learners, English as a second language, English as a foreign language, computer assisted language learning, and second language acquisition). The initial search located 1126 studies.

As a next step, both authors screened all 1126 studies using four criteria:

1. English was used as a target language
2. Included vocabulary as an outcome (target or secondary)
3. Used digital video games
4. Document was written or available in English

This screening step resulted in 473 studies for further investigation. After deleting 399 duplicate studies (i.e., overlap from different search keyword combinations and databases) there were 74 studies that underwent thorough assessment against inclusion criteria, and then if applicable, full coding.

2.3. INCLUSION CRITERIA.

These following inclusion criteria were used in full study coding:

1. Provided isolation of vocabulary effect: Regardless of study design, other outcomes, and statistical analyses, we were able to extract a treatment effect specific to a vocabulary outcome.

2. Provided isolation of video-game treatment: Within a study design there was at least one comparison of a group that specifically received a video-game treatment and another group that did not receive the video-game treatment.
3. Used intervention-based study design: Studies were experimental or quasi-experimental and not other research designs (e.g., correlational, qualitative).
4. Used unique data set (i.e., not duplicated elsewhere in data set): Data from a study were not duplicated from any other study (e.g., dissertation and related peer-reviewed paper of same or very similar data).
5. Provided sufficient quantitative information to compute effect sizes: Enough quantitative information was provided to compute a standardized-mean-difference effect size (more on this below).

Upon review of the 74 studies, 21 studies satisfied all inclusion criteria to be used in the meta-analysis. From these 21 studies we initially extracted 22 effect sizes – two effects came from one study where one group was used for two comparisons, and thus contributed two effect sizes. However, after further inspection of the 22 effect sizes, one effect size appeared far too large (i.e., standardized mean difference above five), reaching a point of questioning its plausibility. Another study had a discrepancy in terms of reported quantities being standard errors or standard deviations. Upon exclusion, our final data set consisted of 20 effect sizes from 19 individual studies. Fig. 1 provides a flowchart of specific study-level inclusion and exclusion decisions.

2.4. CODED MODERATORS.

To address our second research question, we coded eight study characteristics (or moderators). In the statistical modeling part of the meta-analysis we assessed each moderator separately to determine if any true effect-size variability could be systematically explained. Because all eight moderators were categorical, moderator analyses were equivalent to subgroup analyses. In instances where moderator information was not provided in a study, authors of those studies were contacted and asked to provide information for moderator analyses. While some authors responded with relevant information, others did not, which is why some entries in Table 1 are designated as “unspecified”. Below are descriptions of the moderator operationalizations. All moderator information for individual studies are provided in Table 1.

2.4.1. GRADE LEVEL.

The grade-level variable partitioned studies with samples of students to Kindergarten – 4th grade (3 studies), 5th – 12th (5 studies), and College (11 studies). We divided the grade levels into these groups to align with common divisions in schooling (i.e., primary-, secondary-, and college-level educational levels). In studies where the age range of a sample was provided in place of grade levels, we converted said range to the appropriate grade level for use in subgroup analyses.

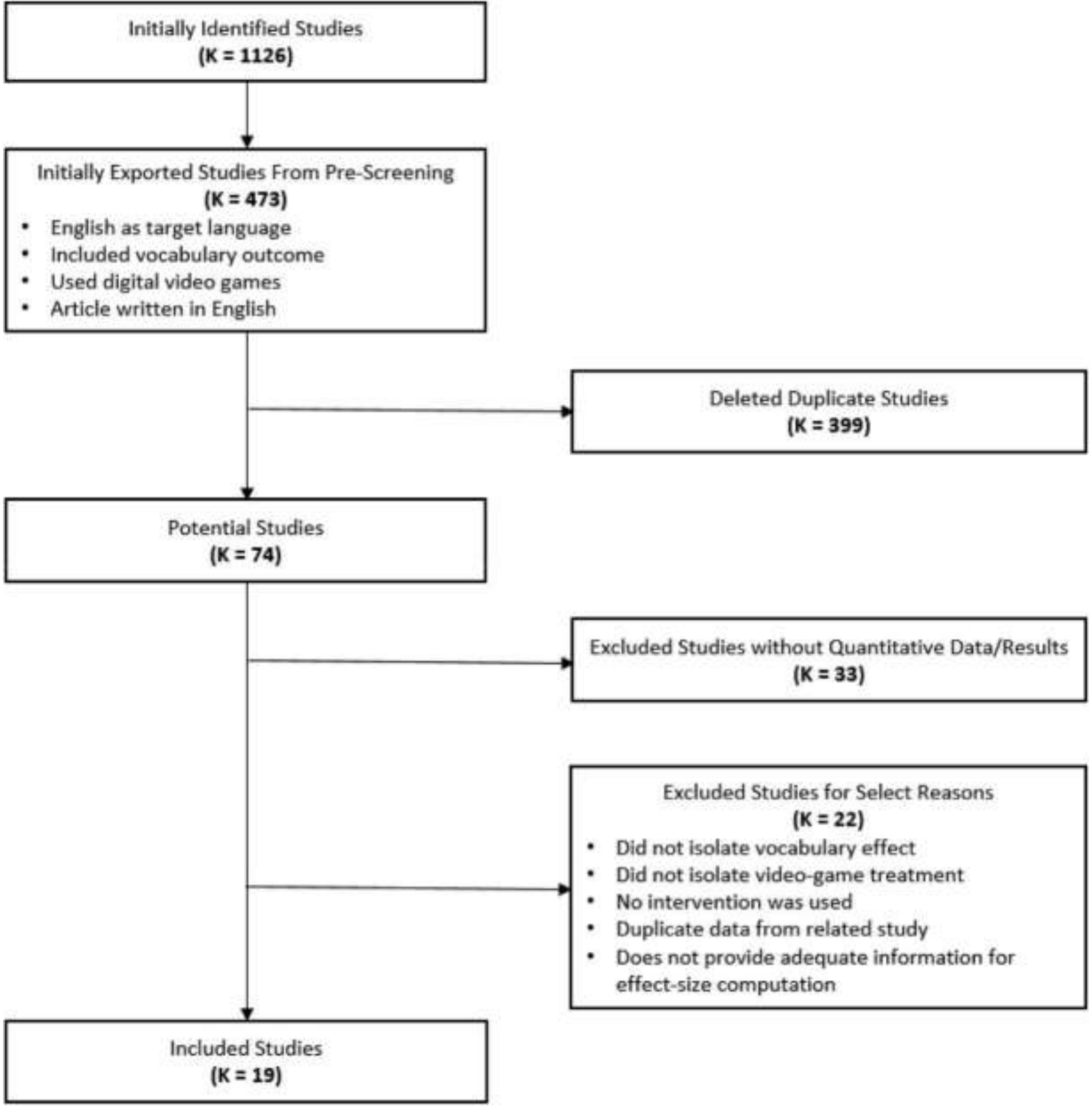
Initially Identified Studies
(K = 1126)
Initially Exported Studies From Pre-Screening
(K = 473)
• English as target language
• Included vocabulary outcome
• Used digital video games
• Article written in English
Deleted Duplicate Studies
(K = 399)
Potential Studies
(K = 74)
Excluded Studies without Quantitative Data/Results
(K = 33)
Excluded Studies for Select Reasons
(K = 22)
• Did not isolate vocabulary effect
• Did not isolate video-game treatment
• No intervention was used
• Duplicate data from related study
• Does not provide adequate information for effect-size computation
Included Studies
(K = 19)

Fig. 1. Flowchart of study inclusion and exclusion decisions. The result is 19 unique studies which produced 20 effect sizes.

2.4.2. HARDWARE.

This variable indicates if game(s) used in a study were played on a computer, video game console (e.g., Playstation or Xbox), mobile device, or if the platform for gaming was unspecified. Hardware was utilized as a moderator because different types of hardware have different affordances and constraints, which may impact how people learn through games on these hardware types. Of the 19 studies, 14 used computers, 2 used a video-game console, and 3 used a mobile device.

2.4.3. GAME TYPE.

The game-type variable indicates if the game(s) used in a study were commercial-off-the-shelf games (i.e., games designed

Table 1
Study characteristics of included primary studies.

		Grade		Hardw	Game	Intervention	FSI	Allocat	Publicati
Study	N	Level	Sex	are	Type	Length	Level	ion	on Type

Aghlara and Tamjid (2011)	40	K – 4th	Female	PC	Serious	Multiple Sessions	Medium	Unspecified	Other
AlShaiji (2015)	60	K – 4th	Female	PC	Serious	Multiple Sessions	High	Random	Journal Article
Alshammari (2013)	24	College	Mixed	PC	Serious	Multiple Sessions	Unspecified	Random	Other
Calvo-Ferrer (2017)	59	College	Mixed	PC	Serious	Multiple Sessions	Low	Random	Journal Article
Cobb and Horst (2011)	50	5th – 12th	Mixed	Console	Serious	Multiple Sessions	Low	Non-random	Journal Article
Ebrahimzadeh (2017)	119	5th – 12th	Male	PC	COTS	Multiple Sessions	Medium	Random	Journal Article
Franciosi (2017)	84	College	Unspecified	PC	Serious	One Session	High	Non-random	Journal Article
Franciosi, Yagi, Tomoshige, and Ye (2016)	162	College	Mixed	PC	Serious	One Session	High	Non-random	Journal Article
Hung (2011)	136	5th – 12th	Mixed	PC	Serious	Multiple Sessions	High	Random	Other
Hung et al. (2015)	30	5th – 12th	Mixed	Mobile	Serious	One Session	High	Random	Journal Article
Jasso (2012)	14	College	Mixed	PC	COTS	One Session	Low	Non-random	Other
Letchumanan, Tan, Paramasivam, Sabariah, and Muthusamy (2015)	70	5th – 12th	Unspecified	PC	Unspecified	Multiple Sessions	Low	Non-random	Journal Article
Salehi (2017)	60	College	Mixed	PC	Serious	Multiple Sessions	Medium	Random	Journal Article
Urun, Aksoy, and Comez (2017)	52	College	Male	Console	COTS	Multiple Sessions	Medium	Non-random	Journal Article
Vahdat and Behbahani (2013)	40	College	Mixed	PC	COTS	Multiple Sessions	Medium	Unspecified	Journal Article
Wu and Huang (2017) I	62	College	Mixed	Mobile	Serious	Multiple Sessions	High	Non-random	Journal Article
Wu and Huang (2017) II	64	College	Mixed	Mobile	Serious	Multiple Sessions	High	Non-random	Journal Article
Yen, Chen, and Huang (2016)	20	College	Mixed	Mobile	Serious	Multiple Sessions	High	Non-random	Journal Article
Yip and Kwan (2006)	100	College	Mixed	PC	Serious	Multiple Sessions	High	Random	Journal Article
Young and Wang (2014)	52	K – 4th	Unspecified	PC	Serious	Multiple Sessions	High	Non-random	Journal Article

primarily for entertainment), serious games (i.e., games designed specifically for non-entertainment purposes, such as education), or if the game type was not specified. Game type was included as a moderator as different game types have diverse purposes and development processes, which may impact learning. Of the 19 studies, 4 were commercial-off-the-shelf games, 14 were serious games, and 1 was not specified.

Related to game type, though not assessed as a moderator because of the excessive heterogeneity and lack of reporting, we did examine the assessment type during experiments. Roughly three-fourths of studies either used researcher-made exams or were un-specified with ample detail. Also, roughly one-fourth of studies used established instruments, such as the Nation and Beglar (2007) Vocabulary Size Test, Ed-Wonderland, or Laufer and Nation (1999) Productive Vocabulary Levels Test.

2.4.5. INTERVENTION LENGTH.

Across studies we found a variety of intervention lengths and intervention length metrics (e.g., days, weeks, number of sessions). As a proxy for intervention length we partitioned studies into two groups: one session or multiple sessions, the same categorization utilized by Clark et al. (2016) in their meta-analysis. There were 15 studies which had more than a single session, typically spanning several weeks with two or more sessions in a given week. The other four studies only used a single intervention session.

2.4.6. LANGUAGE BACKGROUND (FSI LEVEL).

Participants from the studies had a diverse background of primary languages (see data below). Most participants were from Asia (e.g., Taiwan, Hong Kong, Japan, and Malaysia). Six studies included participants from the Middle East (Iran and Saudi Arabia) and Central Europe (Turkey). Three studies included students who primarily spoke Romance languages (e.g., French and Spanish). One study was conducted in the United States and described the participants as hailing from a variety of language backgrounds.

Thus, students in the included studies were from diverse cultural and linguistic backgrounds. Some of the participant's prima ry languages were more closely related to English (e.g., Spanish and French) and thus easier to learn between the languages than other languages that linguistically have less in common with English (e.g., Japanese and Mandarin). This relates to the concept of linguistic distance (Jackson & Kaplan, 2001), which describes how the more commonalities there are between a person's primary language and a target language, "whether due to a genetic relationship or otherwise" the easier it is for speakers to learn a target language (p. 77). Conversely, the less two languages have in common the more difficult it is to learn a target language.

Cysouw (2013) investigated this phenomenon as relates to various languages in comparison to English using data from the FSI based in the U.S. Department of State, which measured the difficulty of learning between English and other languages. Cysouw assigns languages difficulty values using two levels of analysis: a broad level with three groups (I, II, and III, with I being the easiest and III being the most difficult) and more fine-grained levels that ranges from 1 (easiest) to 7 (most difficult). Using Cysouw's scoring system, we demonstrate the FSI levels for participant's primary languages as related to English in Table 2.

As a related note, while the language programs in the studies varied considerably, they generally aim to help students develop vocabulary knowledge that enables them to understand the meaning of vocabulary words and how vocabulary words can be used in general communication, which includes both receptive and productive use.

2.4.7. ALLOCATION.

All studies included in this meta-analysis utilized treatment-control research designs. Across the 19 studies we saw variability among research designs, notably in the method of treatment-control group allocation. This categorical variable considers the treatment allocation as random allocation (8 studies), non-random allocation (9 studies), or unspecified (2 studies).

2.4.8. PUBLICATION TYPE.

This is binary variable which indicates whether an included study is from a peer-reviewed journal or another outlet. Of the 19 included studies, 15 were journal articles, 3 were theses or dissertations, and 1 was a conference proceeding. Some may consider this a form of publication bias assessment. As discussed below, we also included other checks for potential publication bias. Last, all coded data (with moderators) are provided in Table 1.

Table 2
Primary languages from included studies.

Language	Number of Studies	Broad FSI Level	Fine-Grained FSI Level
French	2	I	1
Spanish	1	I	1
Malay	1	I	3
Persian	4	II	4
Turkish	1	II	4
Arabic	1	III	6
Mandarin	7	III	6
Japanese	2	III	7

Note: FSI = Foreign Services Institute; The study that was conducted in the United States with students from a variety of language backgrounds is not included in this table as it was not possible to calculate FSI levels for a group of participants from various primary language backgrounds.

2.5. EFFECT SIZE AND VARIANCE COMPUTATIONS.

As the goal of this meta-analysis was to analyze the treatment effect of video-game instruction on non-video game instruction, the effect size metric for this meta-analysis was the standardized mean difference. While some variation existed between how studies measured vocabulary knowledge, generally speaking, multiple-choice items were the preferred method for assessment. Given this approach, students needed conceptual knowledge in terms of vocabulary meaning and use in context to answer multiple-choice items correctly. Because some studies had relatively small sample sizes (e.g., treatment and control group sample sizes of seven students each), we used a statistically unbiased version of the standardized mean difference proposed by Hedges (1981). The unbiased sample standardized mean difference for the *k*th of $K = 20$ effect sizes was computed as where $\bar{Y}_k^{T}$ and $\bar{Y}^{C}_k$ are respective mean vocabulary acquisition outcomes for the video-game instruction (treatment) and non-video game instruction (control) groups, S_k^{P} is the pooled standard deviation of the two groups, and n_k^{T} and n_k^{C} are respective video-game instruction and non-video game instruction within-study sample sizes. A positive effect-size estimate ($d_k > 0$) is interpreted as a mean difference favoring the video-game instruction group and a negative effect size favoring the non-video game instruction group. Most of the studies did not provide a highly-detailed list of non-video game instructional activities, but rather provided a brief overview of activities that were facilitated in various combinations. The non-video game instructional activities varied from study to study, but many referred to "traditional methods" of vocabulary instruction, which included booklets, worksheets, and listening to instructors and audio clips to expose students to English language vocabulary. Students also read passages in English and engaged in discussions around target English vocabulary.

In 3 of 19 studies all requisite information to compute a standardized mean difference using (1) was not provided. However, using alternative formulas from, among other sources, Borenstein (2009), we were able to compute d_k using t and F statistics from relevant hypothesis tests. In two cases (Alshammari, 2013; Jasso, 2012) we used the *t*-statistic from an independent groups test: where t is the *t*-statistic testing the null hypothesis of no group mean difference. In one case (Franciosi, 2017) we used the F -statistic from a one-way analysis of variance, where F_k is the F-statistic from a one-way analysis of variance. After calculating either (1), (2), or (3), we computed the sample effect-size variances as where all terms have been previously defined.

2.6. EFFECT-SIZE HOMOGENEITY.

When describing our collection of effect sizes, we were interested in the homogeneity of the effects – what is the agreement (or disagreement) of the effectiveness of video-game instruction compared to non-video game instruction on EFL vocabulary acquisition. To do so, we assessed two measures of effect-size homogeneity.

First was the commonly reported Q statistic (e.g., Hedges, 1982),

$$K = \sum v_k^{-1} (d_k - \bar{d}_{\mathrm{IV}})^2, \;\; k=1$$

where $\bar{d}_{\mathrm{IV}}$ is the inverse-variance weighted effect-size mean. Under the null hypothesis of effect-size homogeneity, Q follows an asymptotic chi-square distribution with $K - 1$ degrees of freedom. Larger Q values correspond to more disagreement (i.e., heterogeneity) among effect sizes. We also used the I^2 index (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003) to assess the percentage of effect-size variation that remains after accounting for sampling error,

$$I^2 = \frac{Q-K+1}{Q} \times 100\%.$$

When interpreting I^2, roughly 0%, 25%, 50%, and 75% imply no variation, low variation, moderate variation, and high variation, respectively.

2.7. BAYESIAN META-ANALYSIS.

Frequentist (i.e., non-Bayesian) meta-analysis methods aim to quantitatively synthesize a collection of studies on the same topic or that address the same or similar research questions. In a Bayesian framework the same overarching goal applies. The main difference between the two approaches is the explicit inclusion of prior beliefs or findings when creating models and drawing statistical inferences. As an example, suppose we have a collection of sample effect sizes, **T** (e.g., standardized mean differences, log-odds ratio, correlations), and their respective sample variances, **V**. We are then interested in estimating at least two parameters: the overall mean parameter (i.e., "what is the average effect size?"), μ, and the between-studies standard deviation parameter (i.e., "how variable are the effects?"), τ. Instead of drawing inferences purely on the likelihood of the data, we use a proportional statistical model which combines likelihood information, $p\ (\mathbf{T}|\mu\ ,\ \tau,\ \boldsymbol{\theta}\ ,\ \mathbf{V})$ (where $\boldsymbol{\theta}$ is the vector of true effect sizes), and prior information, p (μ , τ, $\boldsymbol{\theta}$), to compute a posterior distribution, p (μ , τ, $\boldsymbol{\theta}$ $\mathbf{T}$, $\mathbf{V}$). This gives a probability distribution, called the posterior distribution, which permits inferences about our parameters.

Several advantages of Bayesian meta-analysis (BMA) have been presented elsewhere (e.g., Lewis & Nair, 2015; Sutton & Abrams, 2001). All sources of variability are more easily modelled, and in a transparent fashion. Furthermore, as is the case with our set of studies, BMA can be more appropriate when working with a small number of studies than its non-Bayesian counterpart. Last, and what might be one of the most attractive features, because results are described in terms of posterior distributions, we have the capability to make direct statements of probability.

The mechanics of Bayesian estimation are outside the scope of this work. Suffice it to say, all models were conducted in R (R Core Team, 2018) using methods and a DIRECT algorithm described in Röver (2017a). For overall analyses and subgroup analyses (de-scribed below) we assessed qualities of marginal posterior distributions. In all cases we plot the marginal posterior densities (for both the overall effect sizes and between-studies standard deviation) and provide the median, mean, and standard deviation. We also report the 95% Highest Posterior Density Intervals (HPDIs) in all cases. These intervals can be interpreted as the smallest possible credible interval which covers 95% of the marginal posterior distribution. Though similar to the non-Bayesian, often reported confidence interval, HPDIs can make direct probability statements. Last, in all cases we provide Bayes Factor results. These are numerical tests which compare a null model to some other (i.e., non-null) model. In our work, null models are defined as no overall effect (i.e., μ = 0) and no between-studies variability (i.e., τ = 0). From the composition of our Bayes Factors, smaller results provide more evidence against the null model.

2.7.1. OVERALL MODEL.

A main interest of our study was to assess an overall representation of effect sizes. That is, we wanted to assess the overall effectiveness of video-game instruction on vocabulary acquisition for EFL learners. This part of the meta-analysis included looking at an overall weighted mean vocabulary acquisition across studies, as well as an assessment of variability across studies.

For the overall analysis part of our meta-analysis we chose to adopt a random-effects model for several reasons. First, as will be discussed below, results from homogeneity tests indicated likely effect-size discrepancies. The random-effects model allows for a non-zero effect-size variability term, thus allowing us to model effect-size heterogeneity across studies. Second, we aim to generalize our results and inferences outside the set of the 19 collected studies (20 effect sizes) in this meta-analysis, a feature possible with a random-effects model. Third, when estimating overall model quantities, we aim to be more conservative than liberal in terms of estimation. For example, for error in estimation of between-studies variability, we would rather overestimate variability rather than underestimate variability. Fourth, the common formulation of a univariate hierarchical BMA model is inherently a random-effects model.

The random-effects model used in this meta-analysis in hierarchical distributional form is

$$d_k \sim N\,(\delta_k\,,\, v_k)$$
$$\delta_k \sim N\,(\mu\,,\, \tau^2)$$
$$\mu \sim N\,(0,\ 100^2)$$
$$\tau \sim DM\,(s_0),$$

where δ_k represents the true value of the kth effect size, μ population parameter of the overall mean of effects, τ is the between-studies standard deviation parameter, N (•,•) is a Normal distribution with mean and variance parameters, and DM (s_0) is a DuMouchel prior distribution (DuMouchel, 1994). The DuMouchel prior distribution is a log-logistic function of within-study variances (via a harmonic mean of v_k values, or s_0). Several different prior distributions (uniform, square root, Jeffreys) were compared for τ to address sensitivity. Because no major differences were observed, we chose to present results using the DuMouchel prior distribution for τ. Also, because the DuMouchel is considered a proper prior distribution (i.e., mathematical integration of density equals one) we can compute Bayes Factors, providing us with additional parameter information.

Our interest is in both the mean effect sizes (μ) and between-studies standard deviation (τ). We assess the parameter estimates separately via their respective marginal posterior distributions. This is done both graphically (density estimate) and quantitatively using several indices: median, mean, standard deviation, HPDI, and Bayes Factor.

2.7.2. SUBGROUP ANALYSES.

Each of the eight moderators (grade level, sex, hardware, game type, intervention length, FSI, level allocation, and publication type) were categorical variables. As such, we analyzed each moderator separately as individual subgroup analyses. This is to say, for

Table 3
Descriptive statistics of marginal posterior distributions: Effect-size mean.

Model	K	Median	Mean	SD	HPDI: LB	HPDI: UB	Bayes Factor

Overall	20	0.697	0.699	0.111	0.480	0.921	0.004
Grade Level							
Kindergarten – 4th	3	0.569	0.570	0.444	−0.252	1.395	47.511
5th – 12th	5	0.520	0.512	0.201	0.101	0.899	22.132
College	12	0.827	0.831	0.159	0.520	1.152	0.093
Male	2	1.285	1.301	0.614	0.508	2.190	3.529
Female	2	0.887	0.883	0.617	0.053	1.693	11.469
Mixed	13	0.622	0.626	0.138	0.357	1.904	0.349
Unspecified	3	0.443	0.440	0.300	−0.113	0.980	52.035
Hardware							
PC	14	0.689	0.701	0.131	0.444	0.963	0.071
Console	2	0.926	0.936	1.604	−1.407	3.306	37.668
Mobile	4	0.561	0.563	0.236	0.112	1.018	17.160
Game Type							
Commercial/Off-the-Shelf	4	1.370	1.379	0.238	0.936	1.853	0.332
Serious	15	0.544	0.545	0.101	0.345	0.747	0.089
Unspecified	1	–	–	–	–	–	–
Intervention Length							
One Session	4	0.452	0.466	0.229	0.049	0.932	24.811
Multiple Sessions	16	0.743	0.744	0.129	0.489	1.002	0.026
FSI Level							
Low	4	0.511	0.516	0.198	0.141	0.893	11.920
Medium	5	1.069	1.077	0.310	0.467	1.710	3.043
High	10	0.586	0.588	0.147	0.296	0.885	1.720
Unspecified	1	–	–	–	–	–	–
Allocation							
Random	8	0.705	0.702	0.162	0.372	1.021	1.524
Non-Random	10	0.601	0.608	0.161	0.297	0.939	1.734
Unspecified	2	1.186	1.199	1.289	−0.593	3.043	15.663
Publication Type							
Journal Article	16	0.725	0.726	0.133	0.464	0.992	0.048
Other	4	0.512	0.522	0.219	0.112	0.952	14.411

Note: HPDI = Highest Posterior Density Interval; LB = Lower Bound; UB = Upper Bound.

each category within a moderator we modelled individual BMAs. As an example, for the publication type moderator there were two subgroups: journal article or other. For this moderator two BMAs were conducted, one only including the set of studies that were from journals and another only using all other studies. As with the overall model we looked at the mean effect sizes and between-studies standard deviation marginal densities separately.

2.8. PUBLICATION BIAS.

The last set of analyses was completed using non-Bayesian estimation. We performed four checks of publication bias: visual inspection of funnel plot, Trim-and-Fill test (Duval & Tweedie, 2000), Egger's regression test (Egger, Smith, Schneider, & Minder, 1997), and Vevea and Hedges weighted function (Vevea & Hedges, 1995). For evidence against the presence of publication bias we looked for the following for each assessment: Symmetry of effects in a funnel plot, zero or a small number of imputed effect sizes using the Trim-and-Fill method, statistically non-significant slope using Egger's regression test, and a statistically non-significant likelihood ratio test using the Vevea and Hedges method.

Last, Bayesian meta-analysis portions used the bayesmeta package (Röver, 2017b), forest plot and funnel plot were created using the metaphor package (Viechtbauer, 2010), and the Vevea and Hedges weight function used the weightr package (Coburn & Vevea, 2017).

3. RESULTS

3.1. OVERALL ANALYSES.

Quantitative results for overall analyses can be found in Table 3 (mean effect size) and Table 4 (between-studies standard deviation). We also provide a forest plot in Fig. 2. In the forest plot we see that all but one effect-size point estimate (Young & Wang, 2014) are positive, though only 12 of 20 were statistically different from zero. In terms of effect-size estimate precision (indicated in Fig. 2 by either a smaller horizontal line or large black square for higher precision) are similar across studies.

Marginal posterior densities for both parameters (mean effect size and between-studies standard deviation) are shown in Fig. 3. Combining all 20 effect sizes we found that, in terms of the marginal posterior distribution, the mean effect of video-game based

Table 4
Descriptive statistics of marginal posterior distributions: Between-studies standard deviation.

Model	K	Median	Mean	SD	HPDI: LB	HPDI: UB	Bayes Factor
Overall	20	0.383	0.393	0.103	0.202	0.600	< .001
Grade Level							
Kindergarten – 4th	3	0.378	0.496	0.543	0.000	1.314	0.421
5th – 12th	5	0.268	0.304	0.216	0.000	0.686	0.350
College	12	0.413	0.432	0.145	0.176	0.726	0.002
Sex							
Male	2	0.189	0.378	0.998	0.000	1.235	1.062
Female	2	0.171	0.357	0.981	0.000	1.184	1.359
Mixed	13	0.356	0.367	0.140	0.115	0.654	0.026
Unspecified	3	0.194	0.289	0.374	0.000	0.861	1.023
Hardware							
PC	14	0.369	0.383	0.122	0.164	0.632	0.002
Console	2	0.811	1.263	2.356	0.000	3.480	0.064
Mobile	4	0.171	0.242	0.262	0.000	0.705	1.308
Game Type							
Commercial/Off-the-Shelf	4	0.174	0.241	0.255	0.000	0.689	1.224
Serious	15	0.256	0.262	0.108	0.123	0.478	0.123
Unspecified	1	–	–	–	–	–	–
Intervention Length							
One Session	4	0.172	0.242	0.261	0.000	0.702	1.065
Multiple Sessions	16	0.402	0.414	0.121	0.195	0.659	0.001
FSI Level							
Low	4	0.107	0.167	0.200	0.000	0.520	1.899
Medium	5	0.479	0.537	0.330	0.000	1.117	0.106
High	10	0.340	0.357	0.144	0.100	0.656	0.029
Unspecified	1	–	–	–	–	–	–
Allocation							
Random	8	0.315	0.336	0.164	0.000	0.628	0.628
Non-Random	10	0.356	0.372	0.170	0.028	0.699	0.095
Unspecified	2	0.539	0.899	1.938	0.000	2.699	0.231
Publication Type							
Journal Article	16	0.424	0.437	0.122	0.218	0.683	< .001
Other	4	0.127	0.192	0.222	0.000	0.585	1.679

Note: HPDI = Highest Posterior Density Interval; LB = Lower Bound; UB = Upper Bound.

This standardized mean difference can be interpreted as large, with the true value of the parameter likely differing from zero. This is supported by the HPDI = [0.480, 0.921], indicating the true standardized mean difference between video-game based instruction and non-video game-based group on vocabulary acquisition lies somewhere between 0.480 and 0.921 with 95% probability, as well as a low Bayes Factor of 0.004.

Paired with our large ˆ result is evidence of effect-size heterogeneity. Our two non-Bayesian assessment of heterogeneity, Q (19) = 60.61, $p < .001$ and $I^2 = 69.27\%$, indicate that there is likely variability among effect sizes. Modeling the between-studies standard deviation (τ) in a Bayesian framework gave a marginal posterior distribution mean of $\hat{\tau}_{\text{Overall}} = 0.393$ (HPDI = [0.202, 0.600]). Provided the HPDI and a Bayes Factor < 0.001 (see Fig. 3 for marginal posterior density) we further believe that all effect sizes are not all in agreement. Why is this case? Is some of this variability systematic and explainable? These two questions are the preface for our next set of results: Use select study characteristics to assess subgroup differences in attempt to explain why we have effect-size variability.

3.2. SUBGROUP ANALYSES.

Below we examine each of the eight moderators (grade level, sex, hardware, game type, intervention length, FSI level, allocation, and publication type). In each case, effect sizes were grouped by levels (two, three, or four) of a moderator. The same analyses and quantities that were described above in the methods section and reported in the overall results section were used here. All quantitative results for the mean effect size and between-studies standard deviation (both now presented within groups) can be found in Tables 3 and 4, respectively. Figs. 4–11 show individual marginal posterior densities for specific moderators.

3.2.1. GRADE LEVEL.

For the grade level moderator we partitioned effects from studies which utilized samples of students that were in Kindergarten – 4th grade ($K_{\text{Kin -4th}} = 3$), 5th – 12th grade ($K_{\text{5th -12th}} = 5$), or College ($K_{\text{College}} = 12$). The mean effects of the marginal posterior distributions for Kindergarten – 4th grade and 5th – 12th grade were similar, with respectively.

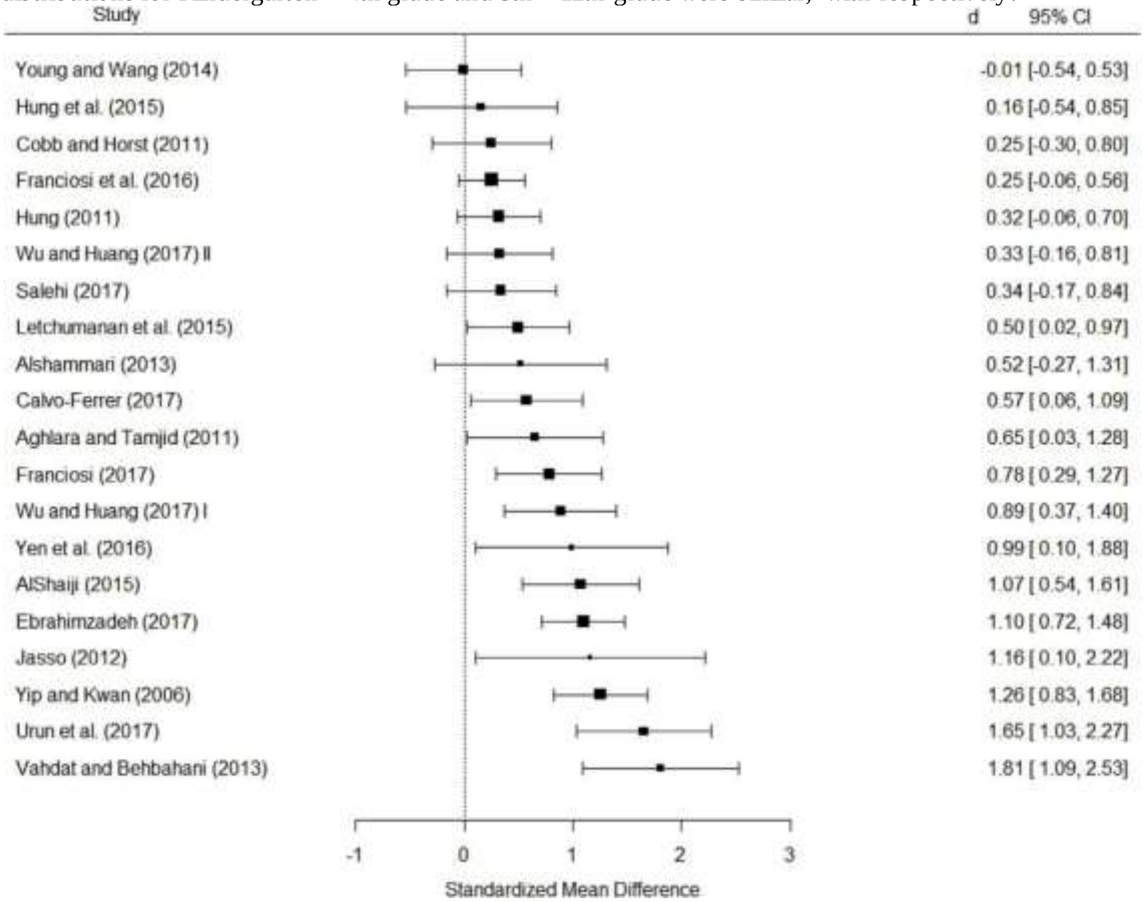

Fig. 2. Forest plot of all included effect sizes. CI = Confidence Interval.

that for the 5th – 12th grade level group. These two differences may be attributable to the low sample size of $K_{\text{Kin -4th}} = 3$. For the College level group, $\hat{d}_{\text{College}} = 0.831$ (HPDI = [0.520, 1.152]), indicating a large effect. It is worth noting that only the College group had a relatively low Bayes Factor (aligning with its HPDI).

Turning to between-studies variability (Table 4), all three grade-level groups displayed signs of within-group variability. Interestingly, the Kindergarten – 4th grade group and the College group were most similar with $\hat{\tau}_{\text{Kin -4th}} = 0.496$ (HPDI = [0.000, 1.314]) and $\hat{\tau}_{\text{College}} = 0.432$ (HPDI = [0.176, 0.726]). However, both the Kindergarten – 4th grade and 5th – 12th grade groups had HPDIs bounded at zero. This is also clearly visible from their marginal posterior densities in Fig. 4, which are clearly non-normal.

3.2.2. HARDWARE.

The hardware moderator partitioned effects from studies which, for the implementation of the video game, used PC ($K_{\text{PC}} = 14$), console ($K_{\text{Console}} = 2$), and mobile ($K_{\text{Mobile}} = 4$). Across all three groups we saw a variety of posterior mean effects (Table 3 has full also includes zero and has the largest Bayes Factor). The PC and mobile groups had similar effect-size estimates of $d_{\text{PC}} = 0.701$ and (a difference of only 0.138) with somewhat overlapping HPDIs. $d_{\text{Mobile}} = 0.563$

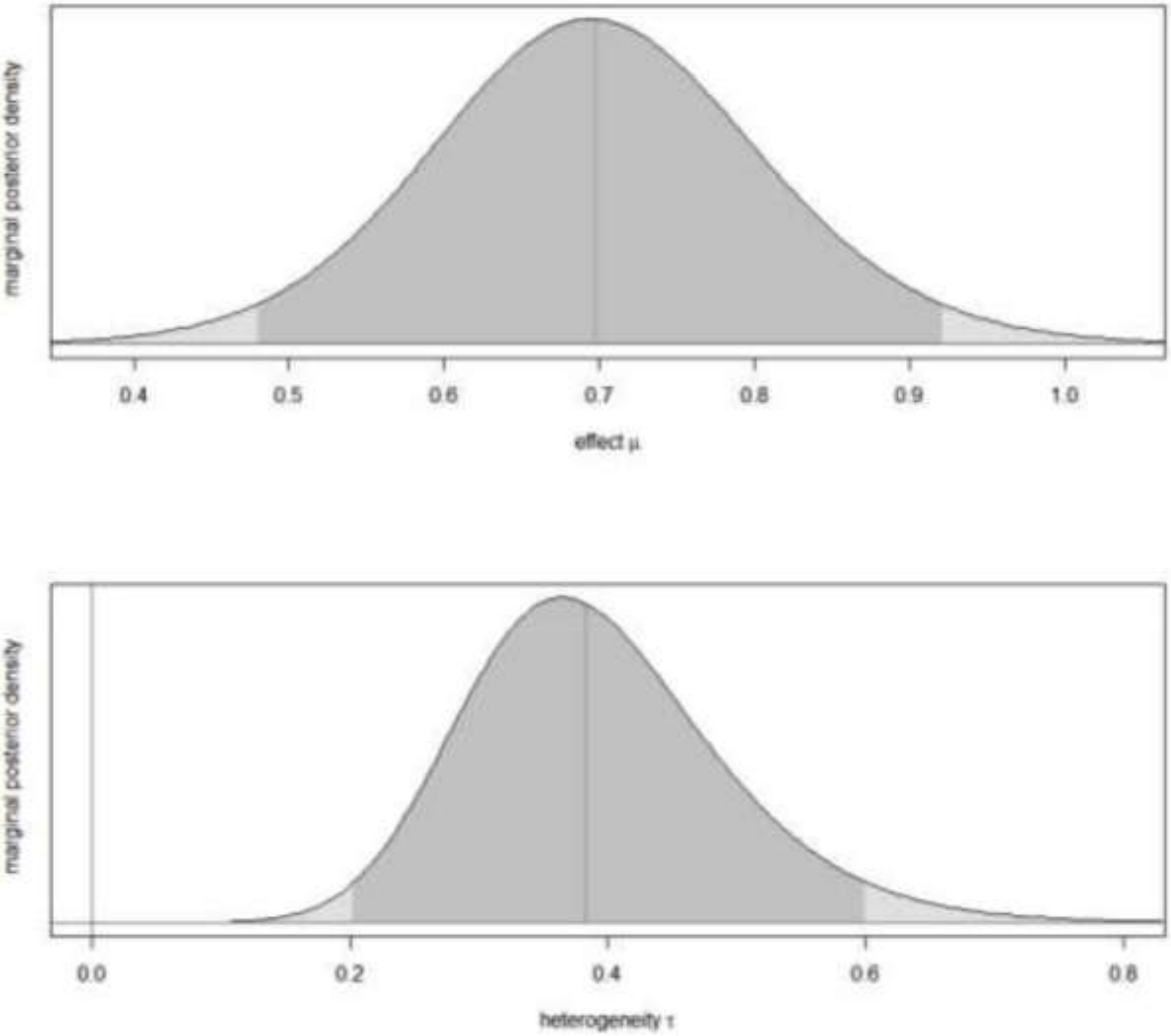

Fig. 3. Overall mean and between-studies standard deviation marginal posterior distributions. The top graphic is the marginal posterior distribution for the overall mean and the bottom graphic is the marginal posterior for the between-studies standard deviation.

Between-studies variability results (Table 4) for the console and mobile groups (i.e., the two smallest sub-groups) were skewed, as shown in Fig. 6. The only $\hat{\tau}$ estimate which appeared to be larger than zero was $\hat{\tau}_{\text{PC}} = 0.383$ (HPDI = [0.164, 0.632]). It is possible that with additional data on the console and mobile groups, we would see different results. For now, the console and mobile groups show large and small degrees of variability, respectively.

3.2.4. GAME TYPE.

As is common distinction in the educational video-gaming literature, we partitioned the game-type moderator by studies which utilized commercial/off-the-shelf games ($K_{\text{COTS}} = 4$), serious games ($K_{\text{Serious}} = 15$), or unspecified ($K_{\text{Game Type–Unspecified}} = 1$). The posterior mean effect (see Table 3) of the commercial/off-the-shelf group was over twice as large as those for serious games: Again focusing on the commercial/off-the-shelf and serious games, within-group between-studies variability (Table 4) was fairly similar for both game types, with $\hat{\tau}_{\text{COTS}} = 0.241$ (HPDI = [0.000, 0.689]) and $\hat{\tau}_{\text{Serious}} = 0.262$ (HPDI = [0.123, 0.478]). The shape of the marginal posterior density (Fig. 7) was very skewed for the commercial/off-the-shelf group compared to that for serious games, possibly an artifact of the small within-group sample size.

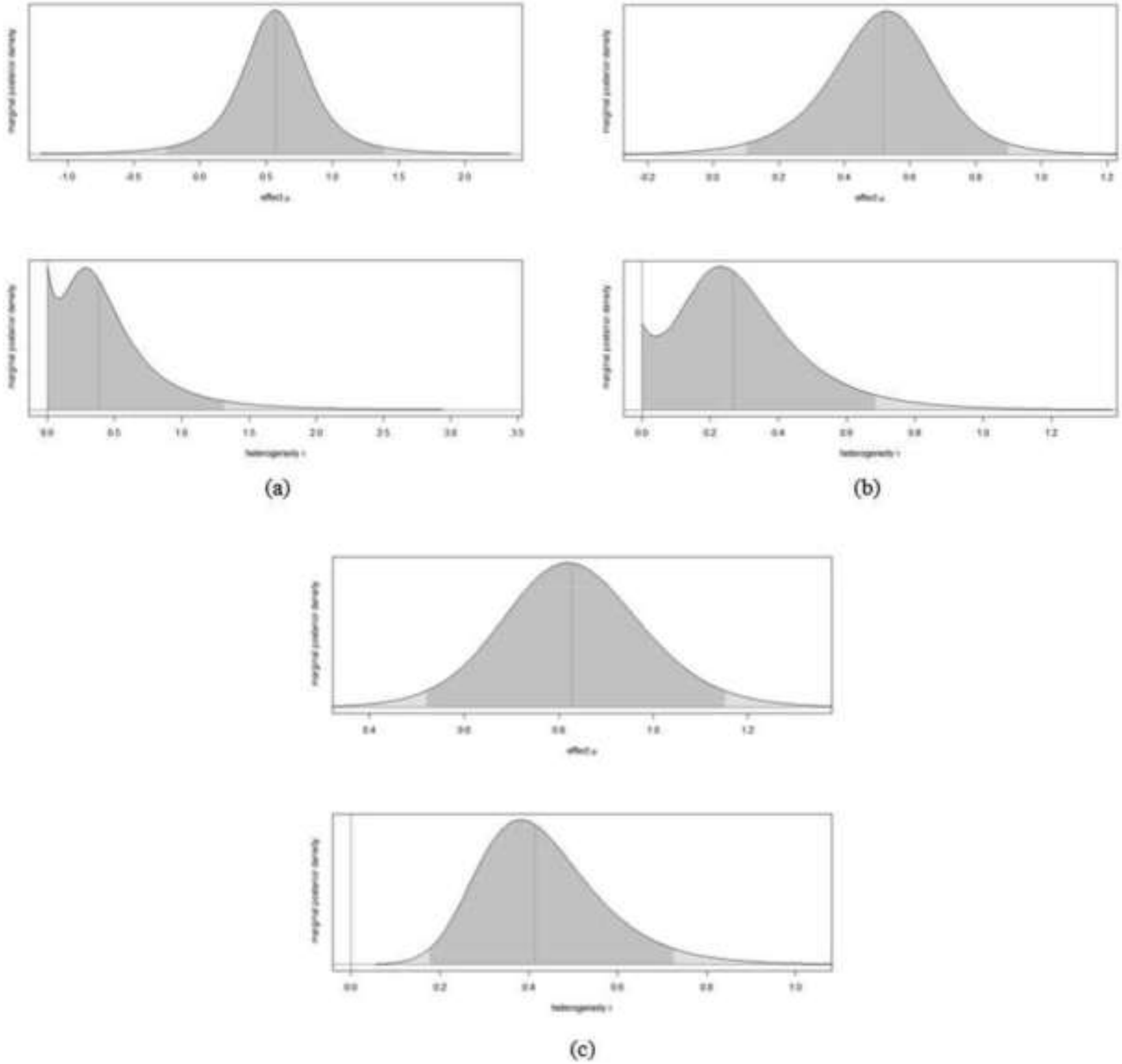

Fig. 4. Mean and between-studies standard deviation marginal posterior distributions for Grade Level moderator. In each cell (a, b, c) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) Kindergarten – 4th, (b) 5th – 12th, and (c) College.

3.2.5. INTERVENTION LENGTH.

All studies were partitioned into two groups depending on the length of the intervention: one session ($K_{\text{One Session}} = 4$) or multiple sessions ($K_{\text{Multiple Sessions}} = 16$). Referring to Table 3 for mean effect-size results and Table 4 for between-studies standard deviation results, we see a somewhat large difference of effects depending on the length of intervention. Namely, for those effects from studies

which used interventions of only one session, $\hat{d}_{\text{One Session}} = 0.466$ (HPDI = [0.049, 0.932]), and for effects from studies which used two

to believe that a longer intervention length tends to provide a stronger effect, or higher EFL vocabulary acquisition.

As expected, variability of effects in the multiple sessions group, $\hat{\tau}_{\text{Multiple Sessions}} = 0.414$ (HPDI = [0.195, 0.659]), was larger than that for the one session group, $\hat{\tau}_{\text{One Session}} = 0.242$ (HPDI = [0.000, 0.702]). With the multiple sessions group there was naturally a larger range of intervention length scenarios (e.g., days, weeks, multiple sessions within a week). With such a variety (both in scenarios and study reporting) this is a more challenging form of variability to capture. The shape of the marginal posterior densities (Fig. 8) show stark differences in shape, with that for the one session group very right skewed.

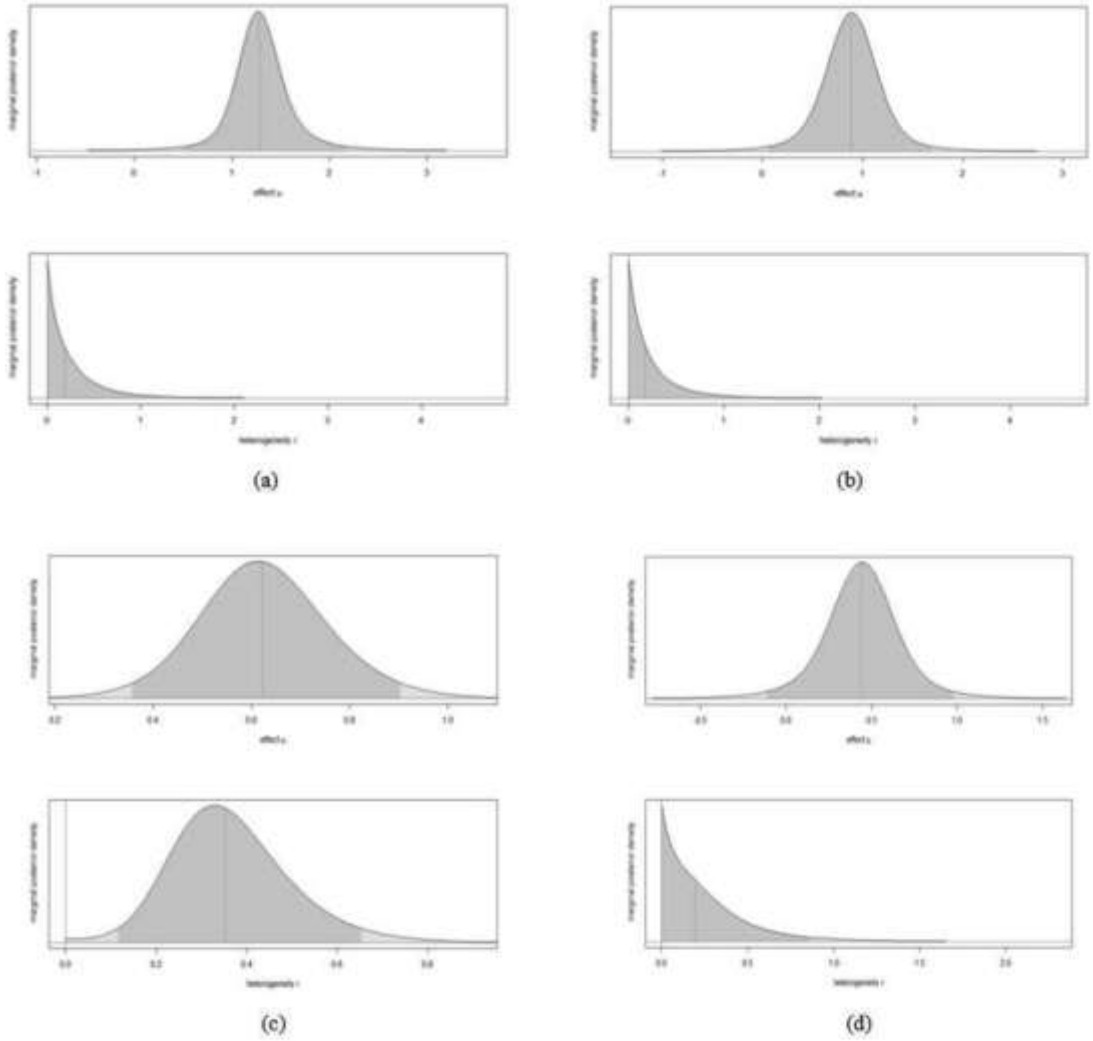

Fig. 5. Mean and between-studies standard deviation marginal posterior distributions for Sex moderator. In each cell (a, b, c, d) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) Male, (b) Female, (c) Mixed, and (d) Unspecified.

3.2.6. FSI LEVEL.

For the FSI Level moderator, studies were partitioned into four groups: Low FSI ($K_{\text{One Session}} = 4$), Medium FSI ($K_{\text{FSI –Medium}} =$ 5), High FSI ($K_{\text{FSI – High}} = 10$), or Mixed FSI ($K_{\text{FSI –Unspecified}} = 1$). For the mixed FSI level group, it was designated as "mixed" because this study had students from a variety of primary language backgrounds, and thus, there was not a corresponding FSI level that could be assigned to the group. Given that there was only one study that had students from a variety of primary language backgrounds, this study was not included in the FSI moderator analyses. Across the three groups (low, medium, and high FSI level), estimated posterior (HPDI = [0.000, 0.520]), $\hat{\tau}_{\text{FSI –Medium}} = 0.537$ (HPDI = [0.000, 1.117]), and $\hat{\tau}_{\text{FSI –High}} = 0.357$ (HPDI = [0.100, 0.656]). As such, there is not an immediately visible trend of between-studies variability increasing or decreasing as a function of FSI level. Last, as shown in Fig. 9, the marginal posterior densities were dissimilar across groups, ranging from highly skewed, moderately skewed, and bell-shaped for low FSI, medium FSI, and high FSI groups, respectively.C.G. Thompson and S. von Gillern *Educational Research Review 30 (2020) 100332*

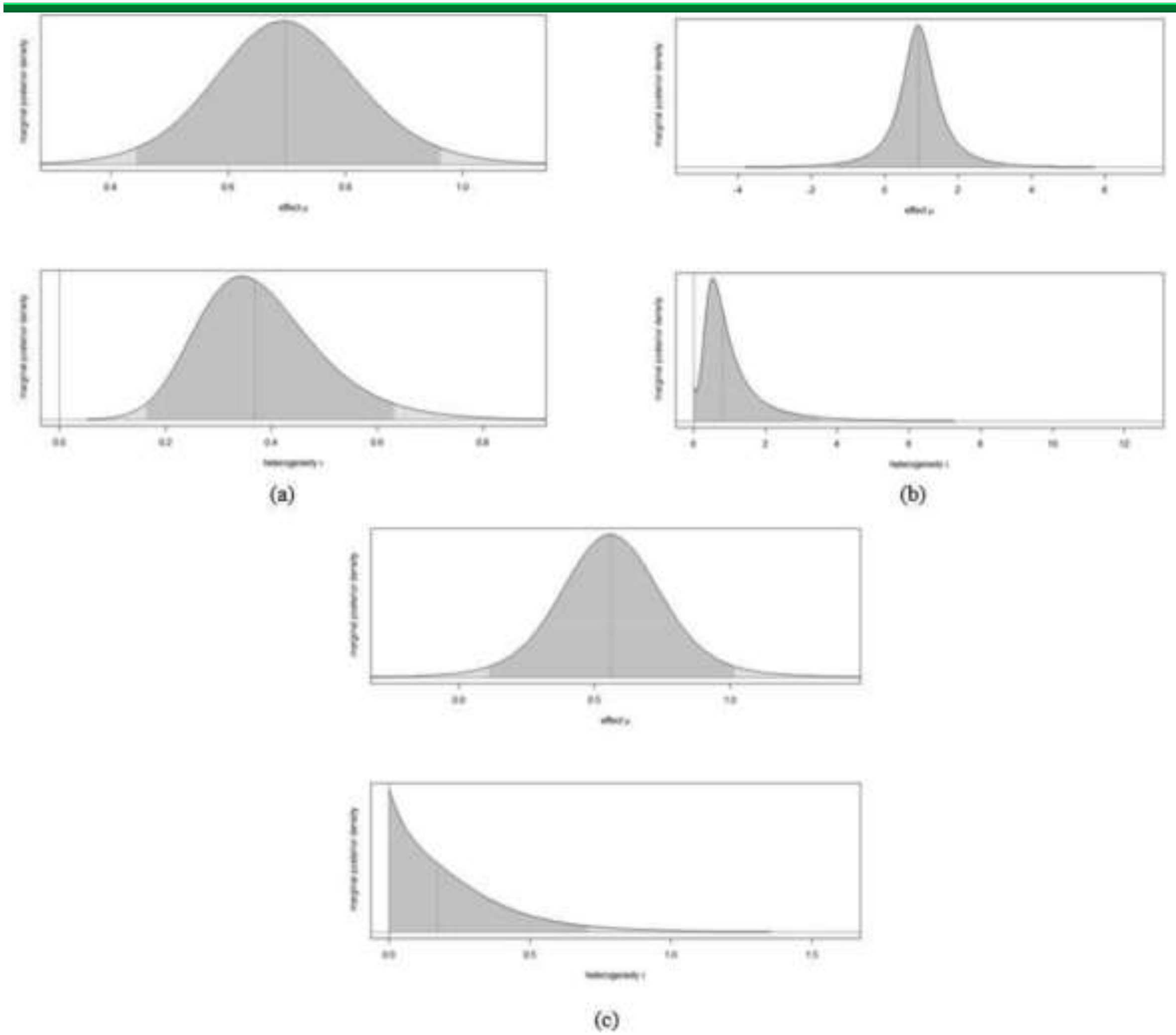

Fig. 6. Mean and between-studies standard deviation marginal posterior distributions for Hardware moderator. In each cell (a, b, c) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) PC, (b) Console, and (c) Mobile.

3.2.7. ALLOCATION.

Because all included studies in this meta-analysis were experiments that included at least one video-gaming group and one non-video gaming group, we partitioned the allocation moderator by studies which utilized random allocation ($K_{\text{Random}} = 8$), non-random allocation ($K_{\text{Non -Random}} = 10$), or unspecified ($K_{\text{Allocation -Unspecified}} = 2$). The posterior mean effect (Table 3) for the random allocation group was only a small degree larger than that for the non-random allocation group, $\hat{d}_{\text{Random}} = 0.702$ (HPDI = [0.297, 0.939]) and (HPDI = [0.297, 0.939]). This suggests a conservative effect-size magnitude for quasi-experimental studies $d_{\text{Non -Random}} = 0.608$ compared to true experimental studies. However, the HPDI for the non-random allocation group is fully nested within the HPDI for the random allocation group, suggesting the difference between the two means of the marginal posteriors is not meaningful.

Similar to the mean effect results described above, within-group between-studies variability (see Table 4) for the random allocation and non-random allocation groups was similar with $\hat{\tau}_{\text{Random}} = 0.336$ (HPDI = [0.000, 0.328]) and $\hat{\tau}_{\text{Non -Random}} = 0.372$ (HPDI = [0.028, 0.699]). The shapes of the marginal posterior densities (Fig. 10) were also comparable between the two groups with a slight right-skew.

3.2.8. PUBLICATION TYPE.

The last moderator was the type of publication, which was coded as either a journal article ($K_{\text{Journal}} = 16$) or other ($K_{\text{Other}} = 4$). As one might expect, we saw a somewhat larger posterior mean effect for the journal article group, (HPDI = [0.464,
$d_{\text{Journal}} = 0.726$

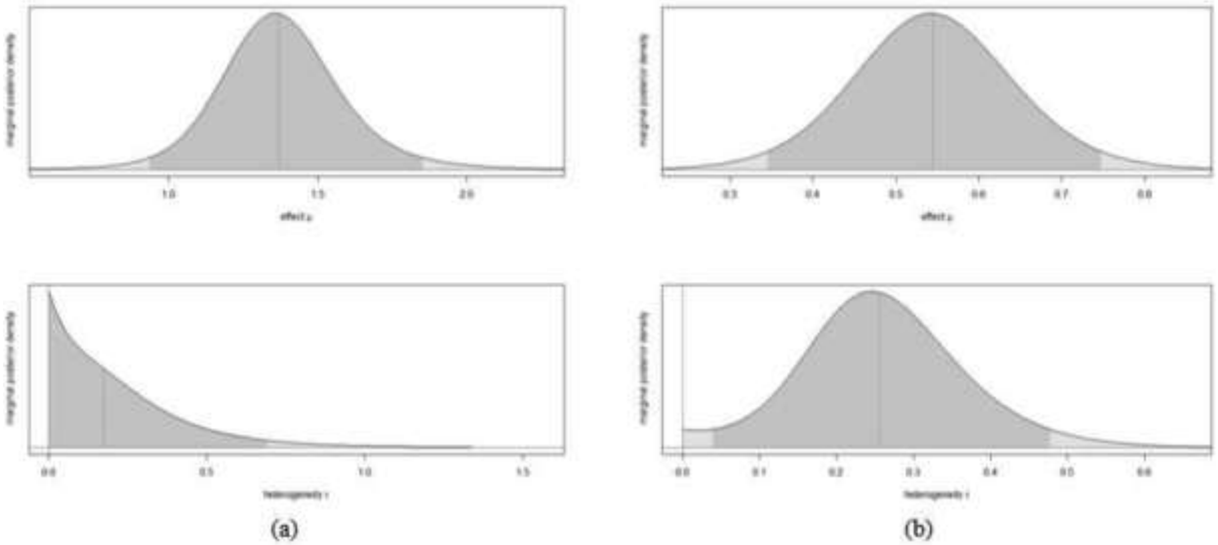

Fig. 7. Mean and between-studies standard deviation marginal posterior distributions for Game Type moderator. In each cell (a, b) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) Commercial/Off-the Shelf and (b) Serious. With only one study not specifying Game Type, no moderator analyses were performed on the unspecified category.

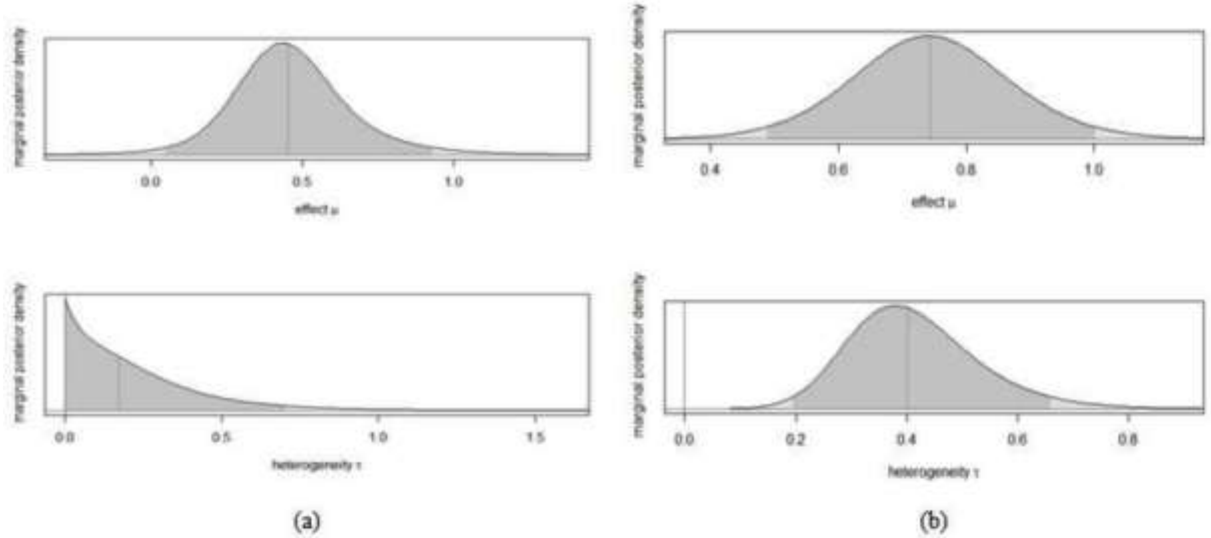

Fig. 8. Mean and between-studies standard deviation marginal posterior distributions for Intervention Length moderator. In each cell (a, b) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) One Session and (b) Multiple Sessions.

Between "published" and "unpublished" studies, but further examination of this contrast is provided in the next section.

In terms of between-studies variability, at first glance it may seem that the effects in journal article group, $\hat{\tau}_{\text{Journal}} = 0.437$ (HPDI = [0.218, 0.683]), were more variable than effects in the other-type group, $\hat{\tau}_{\text{Other}} = 0.192$ (HPDI = [0.000, 0.585]). However, further inspection of the marginal posterior densities (Fig. 11) shows drastically differing shapes – the journal article group appears approximately normally distributed while the other-type group is right skewed.

3.3. PUBLICATION BIAS.

We assessed the risk of publication bias using four methods: visual inspection of funnel plot, Trim-and-Fill test, Egger's regression test, and Vevea and Hedges weighted function. All assessments pointed to a minimal risk of publication bias. Looking at the funnel plot (Fig. 12), there is some visible asymmetry but not to the degree that we would classify as substantial. The Trim-and-Fill result

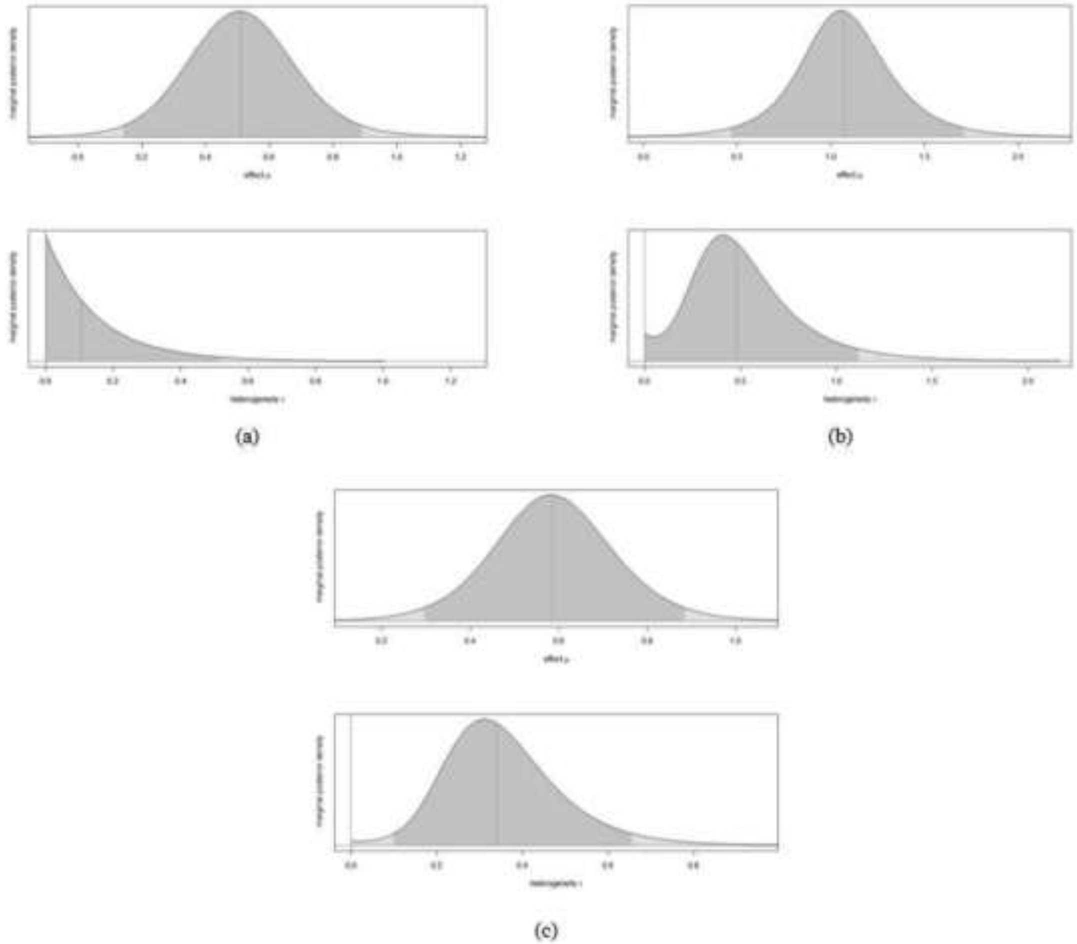

Fig. 9. Mean and between-studies standard deviation marginal posterior distributions for Foreign Services Institute (FSI) level moderator. In each cell (a, b, c) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) Low FSI Level, (b) Medium FSI Level, and (c) High FSI Level. With only one study not having a FSI Level, no moderator analyses were performed on the unspecified category.

It was zero imputed effects (right-side or left-side imputation), hence the lack of extra imputed effects in the funnel plot. Egger's regression test was not statistically significant (Z = 1.25, p = .21), nor was the Vevea and Hedges likelihood ratio test (χ^2 (1) = 0.03, p = .86).

4. DISCUSSION.

4.1. RQ 1 - OVERALL ANALYSIS.

Overall, this Bayesian meta-analysis demonstrates that using digital games to help English learners develop vocabulary can be an effective approach for English language instruction (ˆ , HPDI = [0.480, 0.921],). This finding complements the d_{Overall} = 0.699 K = 20 meta analytic findings of both Zhao (2003) and Grgurovic, Chapelle, and Shelley (2013), both of whom investigated the efficacy of digital vs. non-digital instructional activities amongst language learners and found that digital instruction led to greater learning outcomes that non-digital instruction. These reviews, however, did not focus on digital game vs non-game instructional approaches.

Our study's focus on comparing digital game-based instructional activities to non-digital game-based instructional activities, thus, makes a unique contribution to the field. However, further research is needed to develop a fuller understanding and comparing the effects of digital game-based approaches, digital non-game-based approaches, and different forms of non-digital instructions. There are a variety of permutations possible in these efforts, and more research is needed to provide greater clarity about what approaches are most effective in different contexts.

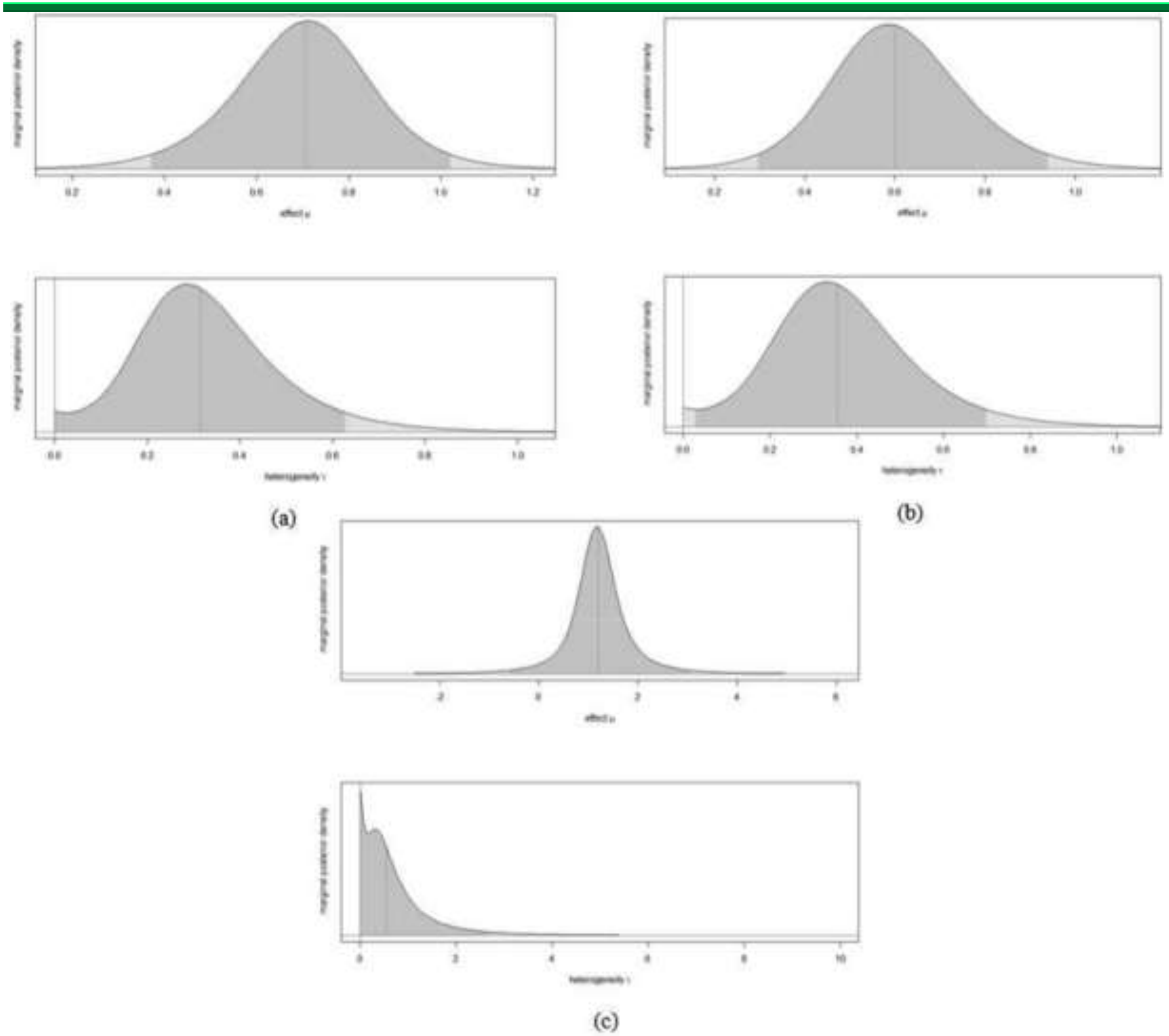

Fig. 10. Mean and between-studies standard deviation marginal posterior distributions for Allocation moderator. In each cell (a, b, c) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) Random, (b) Non-Random, and (c) Unspecified.

Furthermore, given that this study found digital game instructional approaches outperformed non-digital game approaches, this suggests that educators can integrate digital games to promote TESOL vocabulary development. However, as is the case with any instructional approach, careful planning is needed to enhance the likelihood of student success and learning. Thus, TESOL educators should thoughtfully consider educational resources available, goals for learning, and approaches to integrating digital games into their classroom in order to support student achievement.

4.2. RQ 2 – SUBGROUP ANALYSES.

The subgroup analyses results indicate moderate-to-large effect sizes over a variety of moderator variables. Interestingly, there were more college-level studies (K = 12) than studies in primary and secondary settings (K = 3 and K = 5, respectively). Furthermore, college-level participants showed greater gains than their younger counterparts. This may be influenced by college-level students being more motivated to develop their English skills as academic, social, and economic opportunities that accompany English as a global language (Crystal, 2003) are more relevant to their immediate future. Intrapersonal learning outcomes, including motivation, can be enhanced through game-based instruction when compared to non-game-based instruction (Clark et al., 2016). Thus, further research may illuminate if and how intrapersonal learning outcomes vary by age group. While the reason for college students outperforming their younger peers with digital game vocabulary instruction is not clear at this time, further research is needed to better understand digital games in primary and secondary settings in order to further understand younger students'

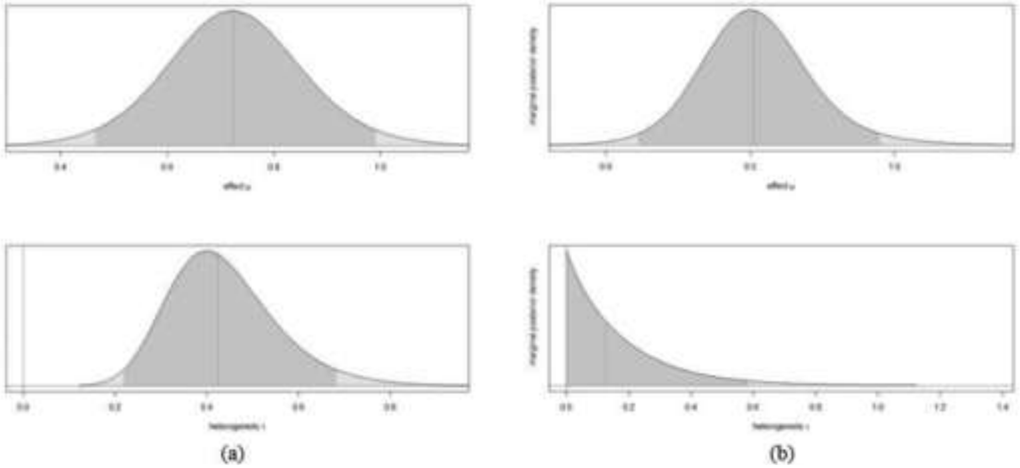

Fig. 11. Mean and between-studies standard deviation marginal posterior distributions for Publication Type moderator. In each cell (a, b) the top graphic is the marginal posterior distribution for the mean and the bottom graphic is the marginal posterior for the between-studies standard deviation. Categories for results are (a) Journal Article and (b) Other.

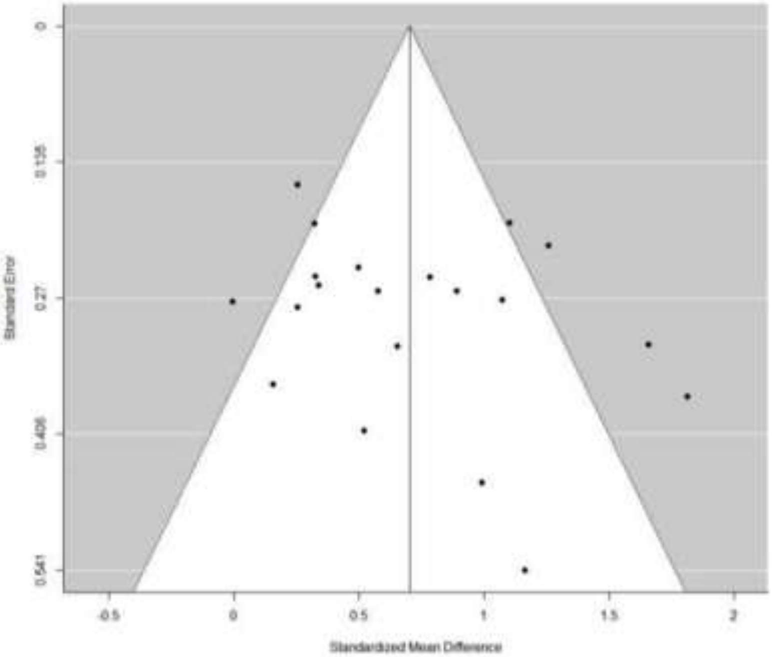

Fig. 12. Funnel plot of effect sizes with 95% confidence interval boundaries.

Learning experiences with games (less research has been conducted in the early years of education than at the college level). Hardware type is another area that deserves scholarly attention. While previous meta-analyses have examined a variety of moderator variables and game factors (Clark et al., 2016; Wouters et al., 2013), they did not explicitly examine hardware type as a moderator variable. Hardware is an important factor in gameplay as different types of hardware (e.g., computers, consoles, and mobile devices) have unique limitations and affordances, which influence human-game interactions and learning. While the results of this study showed a larger effect size for console-based game experiences than PC experiences and mobile experiences, only two console-based studies were included in the meta-analysis. This small sample size restricts implications of the console studies larger effect size. Nonetheless, as different types of hardware have different affordances and limitations, this meta-analysis demonstrates that each of the three main types of modern hardware for gaming can support English vocabulary learning. Furthermore, given the rapid increase of access to mobile technologies in recent years and their capabilities for facilitating learning (Briz-Ponce, Pereira, Carvalho, Juanes-Méndez, & García-Peñalvo, 2017; Wu & Huang, 2017), further research on digital games via mobile devices, comparing how different hardware types may affect English vocabulary learning, is necessary. Also, given that moderator analyses indicated different types of hardware can support TESOL vocabulary learning with moderate to large effect sizes with each type of hardware, educators should be aware that each hardware type outperformed non-game instructional approaches. Last, given the efficacy across hardware types, educators should consider which type(s) of hardware they have available and develop thoughtful plans for integration based upon hardware availability, course learning goals, and student preferences.

Similar to how different hardware can affect gameplay experiences, so can game type (e.g., commercial-off-the-shelf games vs. serious games). As these different game types were designed for different purposes, they can influence overall game experiences (Charsky & Mims, 2008; Michael & Chen, 2005). Neither Clark et al. (2016) nor Wouters et al. (2013) distinguished between entertainment games and serious games in their meta-analyses. The present study's inclusion of COTS and serious games as a game-type moderator variable represents a unique contribution to the literature. This study revealed a large effect size for COTS, but the number of studies that used a COTS game (K = 4) was limited, making it difficult to draw solid conclusions about the efficacy of COTS games. Nonetheless, results indicate that both English vocabulary development using COTS and serious games appears more effective than non-game treatments.

A variety of COTS video games have been incorporated into TESOL environments for various purposes. Given that limited COTS games met inclusion criteria for this meta-analysis, further research is needed to examine COTS and their effectiveness in TESOL settings. One genre that has seen significant attention is MMORPGs. These games support large networks of players who often join together in the online world, engaging in quests and communicating in the process. Scholars have investigated how MMORPGs can facilitate reading, writing, and general social interaction (Peterson, 2010; Reinders & Wattana, 2015). Other scholars have integrated entertainment video games into TESOL environments and found that students can develop English vocabulary through various entertainment games ranging from a real-time strategy game, in which players gather resources, develop bases, and battle other players (Ebrahimzadeh, 2017) to an adventure style murder mystery game, in which the player solves puzzles and collects clues (Vahdat & Behbahani, 2013). While these represent valuable findings, overall, there is limited research on COTS games in TESOL environments. Given that the present study demonstrated that COTS games led to greater vocabulary learning outcomes than serious games, further research is needed on COTS games in TESOL contexts, which have largely received less attention than serious games, likely because of COTS games' lack of explicit educational focus (Van Eck, 2008).

Given that COTS games (K = 4) outperformed serious games (K = 15) in the moderator analysis, educators should consider integrating COTS games into their TESOL vocabulary instruction. An important caveat of this, however, is that educators need to consider the types of vocabulary likely to be covered in the game and utilized by students and determine if these types of vocabulary align with learning goals for the students. Many games encourage social interaction between players, which can occur through speaking, listening, reading, and writing, which could promote general social communication skills and its accompanying vocabulary. However, some games, particularly those that lack social interaction, may have a narrow focus of vocabulary related to gameplay that does not align with the educational goals of the teachers and students. For example, if students play a non-social fantasy game, they may learn vocabulary related to wizards and mythical creatures that is not likely highly valuable for them in their general academic achievement. Educators, thus, must carefully consider how digital games may (or may not) support students in accomplishing their learning goals as relates to TESOL vocabulary development and structure their activities accordingly.

4.3. RQ 3 – PUBLICATION BIAS.

All four assessments of publication bias (visual inspection of funnel plot, Trim-and-Fill test, Egger's regression test, and Vevea and Hedges weighted function) pointed to a minimal risk of publication bias.

4.4. LIMITATIONS.

While this study illuminates digital games as an effective medium for promoting English vocabulary acquisition to speakers of other languages, a few limitations are worth noting. This meta-analysis restricted its collection of studies to those which met all inclusion criteria. A study being excluded from a meta-analysis does not mean that the study is less valuable than included studies, rather it may not contain the necessary information for methodological calculations and thus cannot be included. Also, this study focused specifically on English as a target language. Alternatively, we could have included all second-language learning vocabulary contexts with a variety of primary and target languages, concentrating on English, which is a frequent target language for learners around the world (Crystal, 2003). This may have reduced potentially confounding variables related to vocabulary development, such as script and pragmatics, that can occur across different languages. Last, in many cases, although we had relevant information to code moderators, the limited number of 19 studies posed a challenge for reliable subgroup estimation in some cases.

Last, studies included in this meta-analysis represent a wide variety of countries and communities and include diverse participants. English language exposure varies within communities and between participants. While it would be interesting to account for such differences in English language exposure between studies and participants and code for them in a moderator analysis, the studies did not provide enough information about participants' level of English language exposure in their daily lives. Thus, we were unable to account for such differences in our analysis.

5. CONCLUSIONS.

This study adds to the existing literature that digital games can facilitate learning in a variety of contexts (Clark et al., 2016; Peterson, 2010; Wouters et al., 2013), including helping speakers of other languages develop their English vocabulary knowledge (Hung, Young, & Lin, 2015; Ranalli, 2008; Wu & Huang, 2017). As the literature demonstrates, at least some of the hype of learning through digital games is justified, but further research is needed to better understand how such processes unfold and affect students of different backgrounds. Two areas that would specifically benefit from further research as highlighted by this study are the use of commercial-off-the-shelf games and examining how hardware may impact learning through digital games in the area of second language development and vocabulary acquisition. These areas, among others, would help illuminate how both hardware and software influence learning in digital game environments.

Declaration of competing interest

We declare that there is no conflict of interest with respect to the research, authorship, and/or publication of this work. No external funding was received or is linked to this work.

REFERENCES

Aghlara, L., & Tamjid, N. H. (2011). The effect of digital games on Iranian children's vocabulary retention in foreign language acquisition. Procedia – Social and Behavioral Sciences, 29, 552–560.

AlShaiji, O. A. (2015). Video games promote Saudi children's English vocabulary retention. Education, 136, 123–132.

Alshammari, A. N. (2013). A quantitative study of the impact of immersive game-based learning on enhancing vocabulary instruction and acquisition for English language learners (Unpublished master's thesis)IL: Western Illinois University.

Borenstein, M. (2009). Effect sizes for continuous data. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.). The handbook of research synthesis and meta-analysis (pp.
221–235). New York: Russell Sage Foundation.

Briz-Ponce, L., Pereira, A., Carvalho, L., Juanes-Méndez, J. A., & García-Peñalvo, F. J. (2017). Learning with mobile technologies–Students' behavior. Computers in Human Behavior, 72, 612–620.

Bryant, T. (2006). Using World of Warcraft and other MMORPGs to foster a targeted, social, and cooperative approach toward language learning. Retrieved September, 28, 2019, from https://web.archive.org/web/20061013063948/http://www.academiccommons.org/commons/essay/bryant-MMORPGs-for-SLA.

Calvo-Ferrer, J. R. (2017). Educational games as stand-alone learning tools and their motivational effect on L2 vocabulary acquisition and perceived learning gains.
British Journal of Educational Technology, 48, 264–278.

Carr, J. M. (2012). Does math achievement h'app'en when iPads and game-based learning are incorporated into fifth-grade mathematics instruction? Journal of Information Technology Education: Research, 11, 269–286.

Charsky, D., & Mims, C. (2008). Integrating commercial off-the-shelf video games into school curriculums. Tech Trends, 52, 38–44.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. Review of Educational Research, 86, 79–122.

Cobb, T., & Horst, M. (2011). Does word coach coach work? CALICO Journal, 28, 639–661.

Coburn, K. M., & Vevea, J. L. (2017). weightr: Estimating weight-function models for publication bias. R package version 1.1.2 https://CRAN.R-project.org/package= weightr.

Crystal, D. (2003). English as a global language (2nd ed.). Cambridge: Cambridge University Press.

Cysouw, M. (2013). Predicting language learning difficulty. In L. Borin, & A. Saxena (Eds.). Approaches to measuring linguistic differences (pp. 57–82). Berlin: De Gruyter Mouton.

DuMouchel, W. (1994, September). Hierarchical Bayes linear models for meta-analysis (tech. rep. No. 27. National Institute of Statistical Sciences.

Duval, S., & Tweedie, R. (2000). A nonparametric 'Trim and Fill' method of assessing publication bias in meta-analysis. Journal of the American Statistical Association, 95, 89–98.

Ebrahimzadeh, M. (2017). Readers, players, and watchers: EFL students' vocabulary acquisition through digital video games. English Language Teaching, 10, 1–18. Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. British Medical Journal, 315, 629–634. Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. Studies in Second Language Acquisition, 24, 143–188.

Ellis, N. C., Simpson-Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. TESOL Quarterly, 42, 375–396.

Esposito, N. (2005). A short and simple definition of what a videogame is. Proceedings of the DiGRA 2005 conference: Changing views - Worlds in play.

Franciosi, S. J. (2017). The effect of computer game-based learning on FL vocabulary transferability. Educational Technology & Society, 20, 123–133.

Franciosi, S. J., Yagi, J., Tomoshige, Y., & Ye, S. (2016). The effect of a simple simulation game on long-term vocabulary retention. CALICO Journal, 33, 355–379.

Freirmuth, M. R. (2002). Connecting with computer science students by building bridges. Simulation & Gaming, 33, 299–315.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. Simulation & Gaming, 33, 441–467.

Gee, J. P. (2007). What video games have to teach us about learning and literacy (2nd ed.). New York, NY: Macmillan.

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. ReCALL, 25(2), 165–198.

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. Computers in Human Behavior, 54, 170–179.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107–128.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490–499.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. Statistics in Medicine, 21, 1539–1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. British Medical Journal, 327, 557–560.

Hung, K.-H. (2011). The design and development of an education-designed massively multiplayer online role-playing game (EDD MMORPG) for young Taiwanese Mandarin-speaking learners learning English vocabulary words (Unpublished doctoral dissertation). Teachers CollegeNY: Columbia University.

Hung, C. M., Huang, I., & Hwang, G. J. (2014). Effects of digital game-based learning on students' self-efficacy, motivation, anxiety, and achievements in learning mathematics. Journal of Computers in Education, 1, 151–166.

Hung, H.-C., Young, S. S.-C., & Lin, C.-P. (2015). No student left behind: A collaborative and competitive game-based learning environment to reduce the achievement gap of EFL students in Taiwan. Technology, Pedagogy and Education, 24, 35–49.

Hwang, G. J., Chiu, L. Y., & Chen, C. H. (2015). A contextual game-based learning approach to improving students' inquiry-based learning performance in social studies courses. Computers & Education, 81, 13–25.

Jackson, F. H., & Kaplan, M. A. (2001). Lessons learned from fifty years of theory and practice in government language teaching. In J. E. Alatis, & A.-H. Tan (Eds.).
Language in our time (pp. 71–87). Washington, DC: Georgetown University Press.

Jasso, P. R. (2012). A non-academic computer video game: Its effect on vocabulary acquisition in the EFL classroom (unpublished master's thesis)NY: Syracuse University.

Johnson, W. L., & Valente, A. (2009). Tactical language and culture training systems: Using AI to teach foreign languages and cultures. AI Magazine, 30, 72–83.

Krashen, S. D. (1985). The input hypothesis: Issues and implications. New York: Longman.

Lantolf, J. P. (2000). Introducing sociocultural theory. In J. P. Lantolf (Ed.). Sociocultural theory and second language learning (pp. 1–26). Oxford, England: Oxford University Press.

Lantolf, J. P., & Thorne, S. L. (2007). Sociocultural theory and second language learning. In B. VanPatten, & J. Williams (Eds.). Theories in second language acquisition:
An introduction (pp. 197–220). Mahwah, NJ: Lawrence Erlbaum.

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. Language Test, 16, 33–51.

Letchumanan, K., Tan, B. H., Paramasivam, S., Sabariah, M. R., & Muthusamy, P. (2015). Incidental learning of vocabulary through computer-based and paper-based games by secondary school ESL learners. Social Sciences & Humanities, 23, 725–740.

Lewis, M. G., & Nair, N. S. (2015). Review of applications of Bayesian meta-analysis in systematic reviews. Global Journal of Medicine and Public Health, 4, 1–9. Meluso, A., Zheng, M., Spires, H. A., & Lester, J. (2012). Enhancing 5th graders' science content knowledge and self-efficacy through game-based learning. Computers &
Education, 59, 497–504.

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. Computers & Education, 70, 29–40.

Michael, D. R., & Chen, S. L. (2005). Serious games: Games that educate, train, and inform. Boston, MA: Cengage Learning.

Nation, P., & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31, 9–13.

Neville, D. O., Shelton, B. E., & McInnis, B. (2009). Cybertext redux: Using digital game-based learning to teach L2 vocabulary, reading, and culture. Computer Assisted Language Learning, 22, 409–424.

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation.
Computers & Education, 52, 1–12.

Peterson, M. (2006). Learner interaction management in an avatar and chat-based virtual world. Computer Assisted Language Learning, 19, 79–103.

Peterson, M. (2010). Massively multiplayer online role-playing games as arenas for second language learning. Computer Assisted Language Learning, 23, 429–439.

Peterson, M. (2012). Learner interaction in a massively multiplayer online role playing game (MMORPG): A sociocultural discourse analysis. ReCALL, 24, 361–380.

Prensky, M. (2003). Digital game-based learning. Computers in Entertainment (CIE), 1, 1–4.

Prensky, M. (2005). Computer games and learning: Digital game-based learning. In J. Raessens, & J. Goldstein (Eds.). Handbook of computer game studies (pp. 97–122).
Cambridge, MA: MIT Press.

R Core Team (2018). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: http://www.R-project.org, version 3.4.4.

Ranalli, J. (2008). Learning English with the sims: Exploiting authentic computer simulation games for L2 learning. Computer Assisted Language Learning, 21, 441–455.

Reinders, H. (Ed.). (2012). Digital games in language learning and teaching. Basingstoke, England: Palgrave Macmillan.

Reinders, H., & Wattana, S. (2015). Affect and willingness to communicate in digital game-based learning. ReCALL, 27, 38–57.

Ronimus, M., Kujala, J., Tolvanen, A., & Lyytinen, H. (2014). Children's engagement during digital game-based learning of reading: The effects of time, rewards, and challenge. Computers & Education, 71, 237–246.

Röver, C. (2017a). Bayesian random-effects meta-analysis using the bayesmeta R package. arXiv preprint 1711.08683. URL http://www.arxiv.org/abs/1711.08683. Röver, C. (2017b). bayesmeta: Bayesian random-effects meta-analysis. R package version 1.5. Retrieved from https://CRAN.R-project.org/package=bayesmeta. Salehi, H. (2017). Effects of using instructional video games on teaching English vocabulary to Iranian pre-intermediate EFL learners. International Journal of Learning
and Change, 9, 111–130.

Smith, G. G., Li, M., Drobisz, J., Park, H. R., Kim, D., & Smith, S. D. (2013). Play games or study? Computer games in eBooks to learn English vocabulary. Computers & Education, 69, 274–286.

Squire, K. (2011). Video games and learning: Teaching and participatory culture in the digital age. New York, NY: Teachers College Press.

Suh, S., Kim, S. W., & Kim, N. J. (2010). Effectiveness of MMORPG-based instruction in elementary English education in Korea. Journal of Computer Assisted Learning, 26, 370–378.

Sung, H. Y., & Hwang, G. J. (2013). A collaborative game-based learning approach to improving students' learning performance in science courses. Computers & Education, 63, 43–51.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. Statistical Methods in Medical Research, 10, 277–303.

Svensson, P. (2003). Virtual worlds as arenas for language learning. In U. Felix (Ed.). Language learning online: Towards best practice (pp. 123–142). Amsterdam: Swets & Zeitlinger.

Tokac, U., Novak, E., & Thompson, C. (2019). Effects of game-based learning on students' mathematics achievement: A meta-analysis. Journal of Computer Assisted Learning, 35(3), 407–420.

Turgut, Y., & Irgin, P. (2009). Young learners' language learning via computer games. Procedia -Social and Behavioral Sciences, 1, 760–764.

Urun, M. F., Aksoy, H., & Comez, R. (2017). Supporting foreign language vocabulary learning through Kinect-based gaming. International Journal of Game-Based Learning, 7, 20–35.

Vahdat, S., & Behbahani, A. R. (2013). The effect of video games on Iranian EFL learners' vocabulary learning. The Reading Matrix, 13, 61–71.

Van Eck, R. (2008). COTS in the classroom: A teachers guide to integrating commercial off-the-shelf (COTS) games. In R. Ferdig (Ed.). Handbook of research on effective
electronic gaming in education. Hershey, PA: Idea Group.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. Psychometrika, 60, 419–435.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36, 1–48.

von Gillern, S. (2016). The gamer response and decision framework: A tool for understanding video gameplay experiences. Simulation & Gaming, 47(5), 666–683.

von Gillern, S., & Alawad, Z. (2016). Games and game-based learning in instructional design. International Journal of Technologies in Learning, 23(4), 1–7.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. Journal of Educational Psychology, 105, 249–265.

Wu, T.-T., & Huang, Y.-M. (2017). A mobile game-based English vocabulary practice system based on portfolio analysis. Educational Technology & Society, 20, 265–277. Yen, L., Chen, C.-M., & Huang, H.-B. (2016). Effects of mobile game-based English vocabulary learning APP on learners' perceptions and learning performance: A case study of Taiwanese EFL learners. In R. M. Idrus, & N. Zainuddin (Eds.). Proceedings of the 11th international conference on e-learning (pp. 255–262). Academic Conferences International Limited.

Yip, F. W. M., & Kwan, A. C. M. (2006). Online vocabulary games as a tool for teaching and learning English vocabulary. Educational Media International, 43, 233–249. Young, S., Wang, Shwu-Ching, & Yi-Hsuan (2014). The game embedded CALL systemto facilitate English vocabulary acquisition and pronunciation. Journal of Educational Technology & Society, 17, 239–251.

Zhao, Y. (2003). Recent developments in technology and language learning: A literature review and meta-analysis. CALICO Journal, 21, 7–27.