Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2022, Volume 26, Issue 1
pp. 1–25

ARTICLE

Second language (L2) gains through digital game-based language learning (DGBLL): A meta-analysis

Daniel H. Dixon, *Northern Arizona University*
Tülay Dixon, *Northern Arizona University*
Eric Jordan, *Northern Arizona University*

Abstract

Studies on digital game-based language learning (DGBLL) have increased in numbers, creating a pool of studies that can be meta-analyzed to measure the overall effect of digital gaming on second language (L2) development. The current meta-analysis targets digital games that were available to the public at the time of data collection, January of 2020, aggregating their effects on L2 development overall and across a number of moderator variables. These moderator variables include the game developers' intended purpose of the game (educational or entertainment), outcome measures (e.g., vocabulary, overall proficiency), and several game design features such as the type of player interaction (single player, multiplayer, massively multiplayer online), among others. Results indicate that DGBLL has had a small to medium positive effect (Cohen's $d_{weighted}$ *= 0.50) for between-groups designs and a medium effect (*$d_{weighted}$ *= 0.95) for within-group designs. Games designed for entertainment were found to be more effective than those designed for L2 education, although there is some overlap in the 95% confidence intervals of the two groups. The overall findings and those from additional moderator analyses are discussed in light of previous DGBLL findings while offering direction for future research and recommendations for improving the methodological rigor and transparency in DGBLL research.*

Keywords: *Digital Game-based Language Learning, Meta-analysis, Digital Games, L2 Learning Outcomes*

Language(s) Learned in This Study: *English, German, Italian, Japanese, Spanish*

APA Citation: Dixon, D. H., Dixon, T., & Jordan, E. (2022). Second language (L2) gains through digital game-based language learning (DGBLL): A meta-analysis. Language Learning & Technology, 26(1), 1–25. http://hdl.handle.net/10125/73464

Introduction

With the rise in popularity and ubiquitous presence of digital games, Reinhardt and Thorne (2016) argue that "it has become easier to imagine digital games as authentic, consequential, and widely applicable L2 learning resources" (p. 416). More generally, the popularity of games continues to grow year after year with 2020 alone seeing a 20% increase in revenue over 2019, totaling $175 billion USD worldwide (Palandrani, 2021). To reach a wide global audience, game developers have commercial interests in releasing their games in a variety of languages. Consequently, gamers can choose to play a game in a second language (L2), gaining access to potentially hundreds of hours of authentic and engaging L2 input. In addition to the games developed for purely entertainment purposes, digital *edutainment* apps designed for L2 education (e.g., *Duolingo*, *MindSnacks*) have also become popular for self-directed study, allowing anytime and anyplace L2 learning through mobile devices.

In general, findings from digital game-based language learning (DGBLL) research have been largely positive, often praising the meaningful L2 interaction that can take place in these digital worlds (see Dixon & Christison, 2021; Reinhardt, 2019). Qualitative research in this domain has given much attention to L2 interaction in massively multiplayer online games (MMOs) and tends to frame this interaction around well-established second language acquisition theory. For example, on the basis of the sociocultural interpretation of Vygotsky's (1978) zone of proximal development, Peterson (2016) argued that MMOs allow for meaningful L2 interaction through collaboration, socialization, and co-construction of meaning (p. 1184). Support for Peterson's findings comes from Dixon and Christison (2021), who reported that negotiation of meaning (Gass & Varonis, 1994) was triggered by the need for collaboration as L2 gamers produced and received large amounts of meaningful input while they collaboratively navigated the contextually rich virtual world of a popular MMO, *Guild Wars 2* (ArenaNet, 2012).

In the current meta-analysis, the primary aim is to aggregate the effects of digital gaming on L2 learning from results reported in published and unpublished research gathered from an exhaustive search of the literature. The overall effectiveness of digital games is aggregated across several moderator variables that aim to shed light on the extent to which various game designs or mechanics have affected L2 learning outcomes. These variables include (a) whether the game was originally designed for entertainment or educational purposes, (b) the effects of the number of players involved at one time during gameplay, (c) modes of communication used during gameplay, (d) the effect that digital games have on various language skills (e.g., vocabulary, listening, reading), and (e) whether supplementary material was incorporated into the game treatment. In the review of the literature that follows, we discuss previous meta-analyses and the aspects of games used to operationalize our definition of a digital game (detailed in Method). The targeted moderator variables are then discussed in light of published research, providing justification for their inclusion in the current meta-analysis.

Literature Review

Previous DGBLL Meta-Analyses

Previous meta-analyses on DGBLL have contributed much to our understanding of the effectiveness of digital games, particularly on L2 vocabulary. Chen et al. (2018) analyzed 10 primary studies that were published between 2002 and 2014. The studies were coded as using either "adventure" or "non-adventure games" in the treatment. The effect size for game design was significantly larger for adventure games (d = 1.867) than non-adventure games (d = 0.705). Tsai and Tsai (2018) also conducted a meta-analysis measuring the effect of DGBLL on L2 vocabulary. They included 26 studies that were published between 2001 and 2017. They reported a large effect (d = 0.986) for L2 vocabulary gains through DGBLL activities compared to alternative non-gaming activities (p. 351).

While quantitative DGBLL research findings have shown generally favorable L2 learning outcomes, the current literature may have some issues that limit generalizability. One such issue concerns variables associated with the many various designs of digital games, often referred to as *game mechanics*. Reinhardt (2021) cautions, "generalizing the implications from the study of L2 gameplay with one title to other titles, even if they are in the same genre, may be risky. A better approach is to focus on the features of the game that are more universal, that is, the mechanics themselves" (p. 70). These game mechanics are an aspect that needs more attention in the current DGBLL literature, an aspect that the current meta-analysis aims to address by including game design elements and mechanics as moderator variables in the analyses. These moderators were chosen based on concerns raised by DGBLL researchers, concerns that we discuss in the sections that follow.

The Intended Purpose of Digital Games: Entertainment Versus Education

Intuitively, it might seem that games designed for education would be more effective than those designed for entertainment; however, many DGBLL researchers and practitioners have argued the opposite. Reinhardt (2019) notes that the computer-assisted language learning (CALL) community has not given a positive evaluation of L2 educational games in part due to the game industry's low investment in such games. Further limiting the success of educational games is the idea that these games replace 'play' with "repetitive and superficial tasks in which the learning objectives are too obvious" (Reinhardt, 2019, p.

280). That is, L2 educational games simply have not achieved the same degree of player engagement compared to their entertainment counterparts. For example, Loewen et al. (2019) measured the progress of *ab initio* Turkish learners who used *Duolingo* over a 12-week period. After treatment, the researchers assessed participants' progress using the first semester cumulative exam of their university's Turkish program. They reported that only one of the nine participants earned a passing score of 70% or more, which "call into question . . . claims regarding *Duolingo's* efficacy" (p. 308), efficacy that may be affected by pedagogy that relies "primarily on grammar-translation and audiolingual-type activities, as is common in [mobile-assisted language learning] MALL" (p. 308). The researchers highlighted the need to investigate L2 learning games and platforms, like *Duolingo*, because empirical evidence measuring their efficacy is lacking in the current literature. This is a gap that the current meta-analysis aims to address by comparing the effectiveness of entertainment games to educational games through a moderator analysis.

Player Interaction: Single Player, Multiplayer, and Massively Multiplayer Online Games

Digital games, particularly those designed for entertainment, have various designs that allow them to be played alone (single player games), with a small group of players (multiplayer games), or with hundreds of players simultaneously online (MMOs). Sundqvist (2019) hypothesized that "the potential for L2 English learning [could] be greater as…in-game social interaction grows larger" (p. 90). Based on this hypothesis, Sundqvist suggested, MMOs "are more beneficial for learning English than multiplayer games which, in turn, are more beneficial than single player games" (p. 90). In a multiple regression analysis, Sundqvist (2019) used type of game (single player, multiplayer, MMO) and time played as variables to predict L2 vocabulary gains. Results indicated that the number of players was not a robust predictor of vocabulary scores (p. 97). Sundqvist noted the need for more research that investigates "specific interplay" variables such as the number of players involved during gameplay and its effect on L2 learning outcomes. In the current meta-analysis, we aim to explore the relationship between "interplay" variables and language gains by including player interaction as a moderating variable using the single player, multiplayer, and MMO taxonomy.

Game Input, Player Output, and Language Skills

Digital games designed for entertainment often have deeply developed storylines and hundreds of characters that expose the player to massive amounts of both spoken and written input (Dixon, 2021). While this design aspect is often praised by DGBLL practitioners, some have cautioned that such large amounts of input could pose "considerable cognitive demand on learners," preventing the allocation of mental resources to the meaning and grammar of L2 input received during gameplay (Reinders & Wattana, 2012, p. 183). However, Hannibal Jensen (2017) found that "gaming with both spoken and written English is significantly related to vocabulary scores" and reported that vocabulary scores were lower for participants who played games with only spoken or only written input (p. 13). It should be noted that the learners in the study were very young, under 10 years of age, and were reported to be motivated to carry out extracurricular activities, like gaming (p. 16). Therefore, it is unknown if older learners would see similar learning outcomes. Hannibal Jensen suggested that future research "examine exactly what linguistic input/output is being offered" in games to better understand their effects on L2 acquisition (p. 16), which is one of the goals of the current meta-analysis. To this end, we include the mode of input received and the mode of output produced during gameplay, if any, as moderator variables. This analysis uses the following contrast: written only, spoken only, or both written and spoken.

Teacher-Mediated Supplemental Material

DGBLL researchers have investigated the effects that teacher mediation and supplementary materials can have on the effects of DGBLL. For example, Ranalli (2008) examined whether "*the SIMs* could be rendered pedagogically beneficial to university-level ESL learners by means of supplementary materials designed to meet criteria for CALL task appropriateness" (p. 1). *The SIMs* is a game in which players create characters and attend to those characters' "physical and emotional needs, help them find jobs and resolve domestic and interpersonal problems" in the game's virtual world (Ranalli, 2008, p. 6). Ranalli found greater gains for learners who played *the SIMs* and also received supplementary material versus those who played the game without the supplementary material (p. 10). In order to further investigate the extent to which teacher mediation can affect learning outcomes, we include a moderator analysis contrasting studies that "enhanced" the effect of a digital game with supplementary material versus those that did not. Finally, we also include another set of moderator variables to quantify the extent to which gaming can affect different L2 skills such as vocabulary, grammar, listening, speaking, reading, and writing.

Research Questions

With such a variety of games available today, it is important to investigate how their various designs impact learning outcomes. This focus aligns with Reinhardt's (2021) recommendation for researchers to consider game mechanics rather than simply generalizing about a particular title or genre of games. To be able to consider the designs of games in this meta-analysis, we only included studies that used publicly available digital games in the treatment, a decision that is further detailed and justified in the Method section. A number of moderator variables were also analyzed to shed light on the effects that game designs and mechanics can have on L2 learning. To this end, the following research questions guided the aims of the current study:

1. To what extent does digital gaming affect L2 learning outcomes?
2. To what extent does the effectiveness of digital gaming vary as a function of:
    - the game developers' intended purpose of the game (i.e., education vs. entertainment)
    - player interaction in the game (i.e., single player, multiplayer, MMO)
    - the type of input players receive and the type of output players produce (i.e., written, spoken, or both)
    - outcome measures (e.g., vocabulary, writing, reading)
    - the use of teacher-mediated supplementary material

Method

Defining a Digital Game

Although previous DGBLL meta-analyses have given great attention to transparency and replicability, DGBLL studies in general could use more detail about the design aspects of the targeted games. For instance, one of the inclusion criteria in Tsai and Tsai (2018) stated "a digital vocabulary learning game was *claimed* [emphasis added] to be implemented in the study" (p. 347), but it was unclear whether such claims were thoroughly investigated before their inclusion into the analysis. This is important because definitions of what is considered a digital game or not can vary substantially. For example, one study that was included in Tsai and Tsai's meta-analysis (Yip & Kwan, 2006) used games that were simply digital crossword puzzles and tile-moving games (p. 238). The simple designs of such "games" are quite different from those in popular high-budget entertainment games or edutainment apps like *Duolingo* that incorporate more complex game mechanics like competition, leveling up, and customizable avatars. Thus, deciding which software constitutes the label of digital game is subjective and lacks a widely accepted definition both within and across research domains.

Towards defining a digital game, Franciosi et al. (2016) wrote that digital games intuitively have attributes like "fun" and "engaging," broadly differentiating games from other types of software. They went on to detail a spectrum: At one end are flashcard games that they define as "a rapid sequence of small challenges or tasks and characterized by a high degree of sequenced repetition exercising one or a small set of isolated skills" and point to software like Quizlet as an example (p. 357). At the other end of the spectrum are simulation games that "incorporate a narrative, or a series of interrelated [fictional] events" (p. 358). In the current study, we do not include games at the lower end of the spectrum, such as software that might be better characterized as digital workbook activities or flashcard games. While such software can obviously have benefits for L2 learning, we wanted to focus on games designed with well-established mechanics such as experience points, iterative difficulty, designable avatars, and/or games that tell stories consisting of interrelated *quests* or *missions*. We do, however, include so-called edutainment apps and games (e.g., *Duolingo*) in this study because they often draw on many well-established game mechanics. We felt the need to include such edutainment apps and games because the efficacy of such platforms have not been fully vetted by the research community (see Loewen et al., 2019), which is one of the areas that the current study aims to address. When uncertainty arose concerning the inclusion of a particular game, we drew on our combined decades-long gaming experience to debate its inclusion until unanimous agreement was reached. To decide whether certain software should be labeled as a game, we often acquired the game in question and/or watched online videos of gameplay so we could assess its mechanics or lack thereof. We recognize this decision process required some subjectivity, which we note as a limitation of this study.

We adopt Reinhardt's definitions in the current study to operationalize the entertainment–educational contrast in digital games. For games developed for entertainment purposes, Reinhardt (2019) uses the term *game-enhanced* to refer to "commercial, non-educational, digital games as resources for formal or informal L2 learning" (p. 141). The developers of entertainment games make no claims about their educational value, including their L2 learning potential. Entertainment games (i.e., game-enhanced) are compared to L2 educational games such as *Duolingo*, *Ed-Wonderland*, and *Adventure German*, which are marketed as digital language-learning tools and referred to as game-based teaching and learning tools (Reinhardt, 2019, p. 194).

Literature Search

We conducted an exhaustive literature search to retrieve quantitative studies with a pretest–posttest design (both within-group and between-groups) that explored the effects of digital gaming on L2 learning. We targeted both published and unpublished studies (i.e., theses and dissertations) to mitigate against publication bias. The most common form of publication bias is the increased likelihood of publication for studies that yield significant results, which has been problematic in L2 research (Norris & Ortega, 2000, p. 431). We investigated publication bias within the pool of studies and briefly reported the findings in Appendix C.

To locate studies, we searched 11 databases: Academic Search Premier, Education Abstracts, Education Full Text, ERIC, JSTOR, LLBA, MLA Bibliography, PsychArticles, PsycINFO, ProQuest Dissertations and Theses, and WorldCat. Each database was searched using the following sets of keywords:

> "digital games" OR "video games" OR games OR gaming OR MMORPG OR MMO OR "mobile games" OR "commercial off the shelf" OR COTS OR "digital game-based language learning" OR DGBLL
>
> L2 OR language OR ESL OR EFL
>
> proficiency OR gains OR achievement OR learning OR vocabulary OR grammar OR listening OR speaking OR writing OR reading

In addition to the 11 databases, Google and Google Scholar were also searched. Considering the vast number of results to go through, we ended the search if several pages of results no longer returned relevant studies. In addition, we examined the reference pages of the relevant DGBLL research and the studies that cited earlier meta-analyses.

Study Inclusion and Exclusion Criteria

The eligible studies examined learner gains either with (a) a pre-post within-group design in which learners played a digital game or (b) a between-groups design in which the gains of the learners playing a digital game were evaluated against the gains of a comparison group that received the same instruction without the use of a game. The search yielded 98 studies. Each study was further examined using the set of inclusion and exclusion criteria outlined in Table 1, leaving 26 eligible studies (marked by an asterisk in the References). The eligible studies had 29 samples using a within-group design and 21 using a between-groups design. The search process took place during January of 2020, so studies published after this time were not included.

Two factors motivated the decision to include only studies using publicly available games. First, the current meta-analysis aims to evaluate the various designs (i.e., mechanics) of games and their effect on L2 outcomes, a goal that was possible only if the games were publicly available. Second, we wanted to aggregate the effects of DGBLL from games that are fully developed, meaning that they had already been vigorously tested before being released to the public. Game development is an iterative process in which "beta-testers" are employed to play and test games, identifying bugs and areas of improvement. Once beta-tester feedback is compiled, the game developers make significant changes to their products before releasing a stable version to the public. Even after an initial public release, many games continue to receive updates for years, or even decades like *World of Warcraft* (Blizzard Entertainment, 2004). Thus, by only including publicly available games, we aimed to create boundaries that reflected the advanced development stage of the products that were included in analysis. Further, we aimed to present a list (Appendix B) of all the games employed across the studies included in analyses. The games listed in Appendix B could be found through a simple Internet search and were publicly available as of January 2020 (the time of data collection). To be clear, we are not discounting research involving games in development or those no longer available. Rather, our goal was to focus only on games that readers could reasonably access through standard means for their own research, teaching, or learning purposes.

Table 1

Inclusion and Exclusion Criteria

Included studies feature...	Excluded studies feature...
• A digital game that was publicly available for purchase or download in January of 2020 • An experimental or quasi-experimental pretest–posttest design either within or between groups • Sufficient reporting of means and standard deviations to calculate effect sizes	• No explicit identification of the title of the game used in treatment • Measurements of learner perceptions rather than L2 proficiency gains • Digital exercises included with a textbook as the treatment • Measurements of L1 gains rather than L2 gains

We also note that we included studies published only in English, despite the fact that, collectively, the research team has proficiency in English, Japanese, Turkish, and Portuguese. We found that many of the studies published in these languages were also published in English. Some relevant studies were likely missed due to this decision, which we note as another limitation of the current meta-analysis.

Studies that met the inclusion criteria (n = 26) were coded for several substantive/methodological features:

- the game developers' intended purpose (e.g., designed for entertainment or L2 education)
- player interactivity (e.g., single player, multiplayer, MMO, modes of language input and output such as spoken, written, or both)
- study design (e.g., participant demographics, within-group vs. between-groups design, outcome measures, outcome measurement instrument reliability, effect size reporting)
- whether teacher-mediated supplementary material was provided in addition to the game treatment
- results (e.g., means, n sizes, standard deviations, effect sizes)
- study identification (e.g., year, publication venue, funded vs. nonfunded research, published, unpublished)

We double-coded each article independently and measured interrater reliability. The overall interrater reliability was satisfactory: 92.68 for percentage agreement and 0.95 for Cohen's kappa (see Appendix A). We discussed and resolved all discrepancies, reaching 100% agreement for all reported features in this meta-analysis. The coding scheme is available for download on the IRIS Database.

Analysis and Aggregating Effect Sizes

Within-group and between-groups designs were analyzed separately to allow for a more precise and informative aggregation of effect sizes since within-group designs tend to have greater effect sizes than between-group designs (Plonsky & Oswald, 2014). Cohen's *d* was selected as the effect size measurement, which quantifies "the magnitude of difference between two mean scores…by dividing the difference between two scores by their (pooled) standard deviation" and should be interpreted using domain-specific benchmarks (Loewen & Plonsky, 2016, p. 27). In this meta-analysis, we interpreted the effect sizes using the benchmarks suggested by Plonsky and Oswald (2014) for L2 research:

- For studies with a between-groups design, *d* values around 0.40 as a small effect size, values around 0.70 as a medium effect size, and values around 1.00 as a large effect size
- For studies with a within-group design, *d* values around 0.60 as a small effect size, values around 1.00 as a medium effect size, and values around 1.40 as a large effect size

For some studies, more than one effect size was calculated. For studies employing a within-group design, for each of the dependent variables, an effect size was calculated by comparing the following:

1. the mean scores between the pretest and the posttest
2. the mean scores between the pretest and the delayed posttest (if the study had one)

For between-groups studies, an effect size was calculated by comparing the following:

1. the mean scores between the groups at the posttest
2. the mean scores between the groups at the delayed posttest (if the study had one)
3. the treatment group's mean scores at the pretest level with their scores at the posttest
4. the treatment group's mean scores at the pretest level with their scores at the delayed posttest (if the study had one)

Two effect sizes are reported in each analysis, one unweighted and one weighted by sample size. A larger sample size does not guarantee more representativeness of the target population, but it will likely reduce the sampling error since small samples tend to have greater variation around the sample mean compared to larger samples (Blair & Blair, 2015, p. 95). The weighted effect sizes were calculated by first multiplying the *n* size by its effect size (Cohen's *d*) for each study. Next, the sum of these values ($n * d$) were divided by the number of participants. This calculation results in a weighted effect size in which studies that have larger samples have a greater influence on the overall weighted *d* value.

To examine moderator variables, subgroups were formed containing all studies that measured the moderator variables of interest. However, some contrasts are not reported because at times only one study measured the variable of interest. Plonsky and Zhuang (2019) argue that excluding contrasts with a very low number of studies avoids "presenting unstable moderator effects based on very small numbers of results" (p. 293).

Finally, it is important to note that we used random effects models rather than fixed effects models. Oswald and Plonsky (2010) argue that the random effects model "has stronger conceptual motivation" in second language acquisition research "because rather than assume homogeneity, the RE model tests for it" (p. 96). Thus, we followed their advice in "[avoiding] homogeneity tests (e.g. Q-tests) and [relying] more heavily on a combination of statistics, data visualization, and solid knowledge of the literature" (p. 97) to justify the targeted moderator analyses.

Results

Overall Immediate and Delayed Effects of DGBLL

Table 2 reports estimates of the overall immediate effects of DGBLL. In the third column, the reported effect sizes are the weighted mean values based on each study's sample size. In the fourth column, the unweighted mean effect sizes are reported. The results suggest that DGBLL can have a small to medium positive effect for between-groups designs ($d_{weighted}$ = 0.50) and a medium positive effect for within-group designs ($d_{weighted}$ = 0.95) when interpreted using the benchmarks for L2 research suggested by Plonsky and Oswald (2014).

Table 2

Overall Immediate Results for the Effectiveness of DGBLL

Contrasts	***k***	$M_{d(weighted)}$	$M_{d(unweighted)}$	***SE***	**95% CI**	
					Lower	***Upper***
Between-groups	21	0.50	0.65	0.234	0.19	1.11
Within-group	29	0.95	1.18	0.172	0.85	1.52

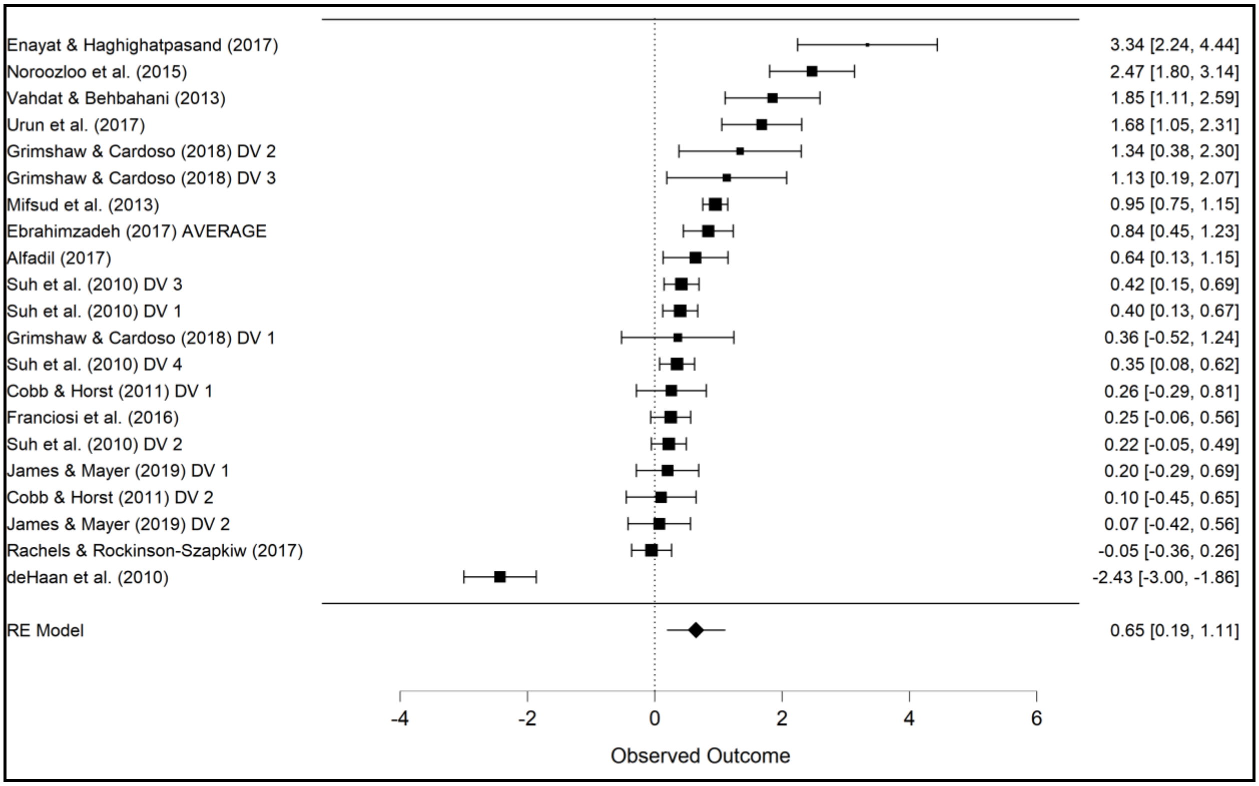
Note. *k* = number of samples; $M_{d(weighted)}$ = mean of effect sizes weighted by sample size; *SE* = standard error; CI = confidence interval. 95% CIs are around the unweighted *d* values.

The forest plots in Figures 1 and 2 illustrate the individual effect size for each study included in this analysis, separated by between-groups (Figure 1) and within-group (Figure 2) designs. These forest plots provide visualizations of each study's unweighted *d* values and their 95% confidence intervals (CIs). These plots were generated using the software suite JASP (JASP Team, 2020). In addition to the upper and lower CI values reported for each study (far-right column), CIs are visualized by the width of the center lines, with wider lines indicating wider CIs. Further, the squares in the middle of each line indicate the size of the standard error in each study. Larger squares indicate less standard error, which typically coincides with larger samples and tighter CIs (Blair & Blair, 2015). The rhombus at the bottom of the figures and the lines stemming from it indicate the unweighted aggregate mean effect size (Cohen's *d*) and the aggregate 95% CIs. The CIs do not cross zero in both between-groups and within-group designs, which indicate statistically significant positive effects from the treatment. These findings suggest that, using Plonsky and Oswald's (2014) L2 research benchmarks, learners can generally expect small to medium positive effects on learning outcomes in DGBLL contexts. It is also worth noting that deHaan et al. (2010) appears to be an outlier (see Figure 1). The reasons for the study's abnormally negative effect may be related to the design of the game used in that particular study, a point we return to in the Discussion section.

Table 3 reports the retention of the effects of DGBLL using three contrasts: (a) delayed posttests comparing a treatment group to a control group, (b) delayed versus pretest effects for within-group designs, and (c) delayed versus immediate posttest effects for within-group designs. For the first contrast, comparing the long-term effects between treatment and control groups suggests that long-term retention does not have a clear advantage from language instruction that incorporates digital games over those that do not as the CIs cross zero [-0.62, 1.03]. However, comparing the treatment groups' delayed posttest to pretest scores (i.e., the second contrast in Table 3) results in a large effect ($d_{weighted}$ = 1.36), indicating instruction using digital games led to a high degree of retention. Still, it appears that there are some losses between immediate and delayed posttests as evidenced by the $d_{weighted}$ value of -0.69 and the CIs [-0.73, 0.25] seen in the third contrast.

Figure 1

Forest Plot of the Overall Immediate Between-Groups Effects

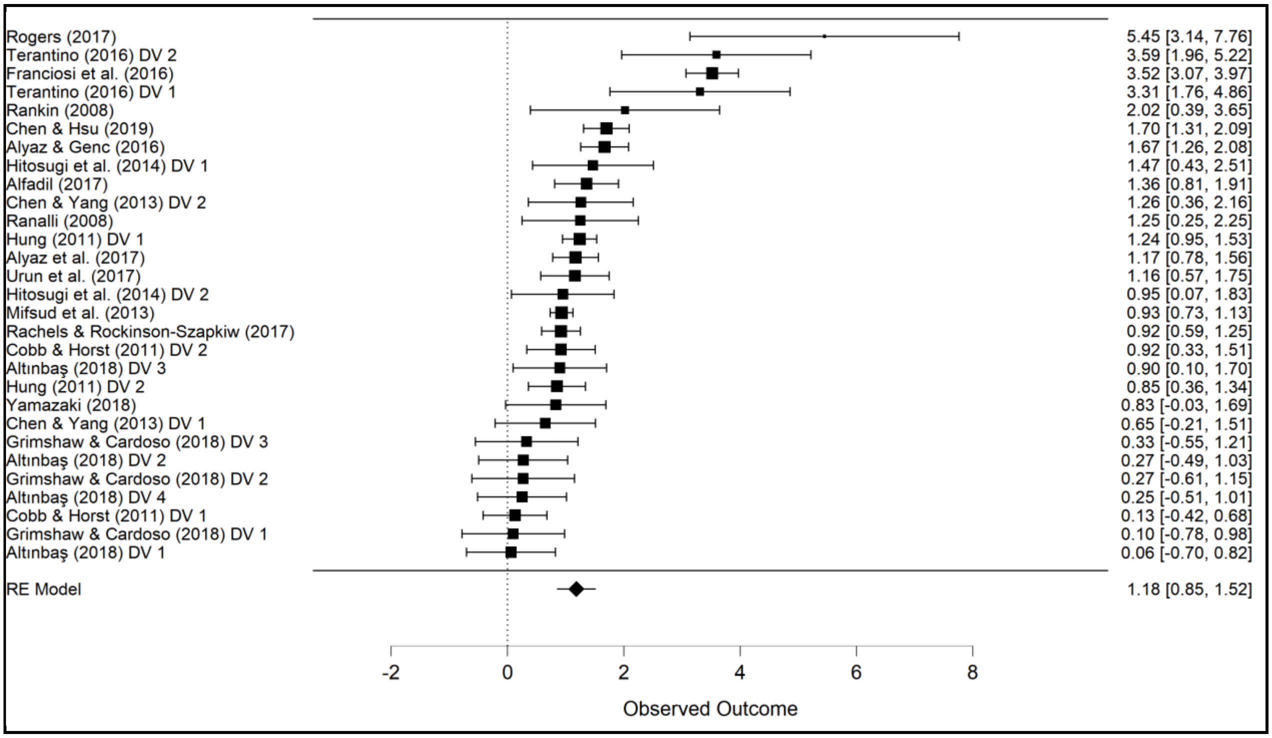


Note. DV = Dependent Variable; Since some studies measured multiple dependent varialbes, each unique DV was separated in the analyses.

RE = Random Effects; Random effects models were used over fixed effects models as described in the Method section above.

Figure 2

Forest Plot of the Overall Immediate Within-Group Effect



Note. DV = dependent variable; Since some studies measured multiple dependent varialbes, each unique DV was separated in the analyses.

RE = Random effects; Random effects models were used over fixed effects models as described in the Method section above.

Table 3

Retention of the Instructional Effect

Contrasts	***k***	***M*$_{d(weighted)}$**	***M*$_{d(unweighted)}$**	***SE***	**95% CI**	
					Lower	***Upper***
Delayed: Treatment vs. control	8	0.09	0.21	0.421	-0.62	1.03
Treatment: Delayed vs. pretest	8	1.36	1.03	0.279	0.48	1.57
Treatment: Delayed vs. immediate	8	-0.69	-0.24	0.251	-0.73	0.25

Note. *k* = number of samples; *M*$_{d(weighted)}$ = mean of effect sizes weighted by sample size; *SE* = standard error; CI = confidence interval. 95% CIs are around the unweighted *d* values.

Moderator Analyses

Table 4 and Table 5 report the results of the moderator analyses for between-groups and within-group designs, respectively. For the contrast regarding the game developers' intended purpose, results indicate that games designed for entertainment purposes were more effective for L2 learning than those designed for education. However, the upper and lower limits of the CIs in columns 6 and 7 are much wider for entertainment games due to the greater variance (i.e., standard error) in the aggregated L2 outcomes, especially for between-groups designs as seen in Table 4. The results appear to support claims regarding the greater efficacy of entertainment games compared to those designed for L2 education, but it should be noted that the overlapping 95% CIs for both between-groups and within-group effects suggest that more data are needed before definitive conclusions can be drawn. For added transparency, Figures 3, 4, 5, and 6 display forest plots that report the unweighted mean *d* values of each study, their standard errors, and 95% CIs. These plots can be interpreted in the same way that was previously described for Figures 1 and 2.

Table 4

Between-Groups Moderator Effects

Contrasts	***k***	$M_{d(weighted)}$	$M_{d(unweighted)}$	***SE***	**95% CI**	
					Lower	***Upper***
Game purpose						
Entertainment	7	0.66	1.12	0.704	-0.26	2.50
Education	14	0.45	0.40	0.094	0.21	0.58
Players						
Multiplayer	4	0.87	0.87	0.161	0.55	1.18
Single player	13	0.55	0.69	0.382	-0.06	1.43
Dependent variables						
Vocabulary	10	0.59	0.87	0.492	-0.09	1.84
Other	11	0.46	0.42	0.112	0.20	0.63
Game input						
Written	8	0.43	0.53	0.244	0.05	1.00
Written and spoken	13	0.58	0.73	0.359	0.02	1.43
Player output						
None	8	0.73	0.96	0.600	-0.22	2.14
Spoken	4	1.29	1.17	0.298	0.59	1.76
Written and spoken	7	0.27	0.26	0.069	0.13	0.40
Teacher mediation						
Yes	6	0.91	1.28	0.546	0.21	2.35
No	17	0.40	0.37	0.209	-0.04	0.78

Note. *k* = number of samples; $M_{d(weighted)}$ = mean of effect sizes weighted by sample size; *SE* = standard error; CI = confidence interval. 95% CIs are around the unweighted *d* values.

Table 5

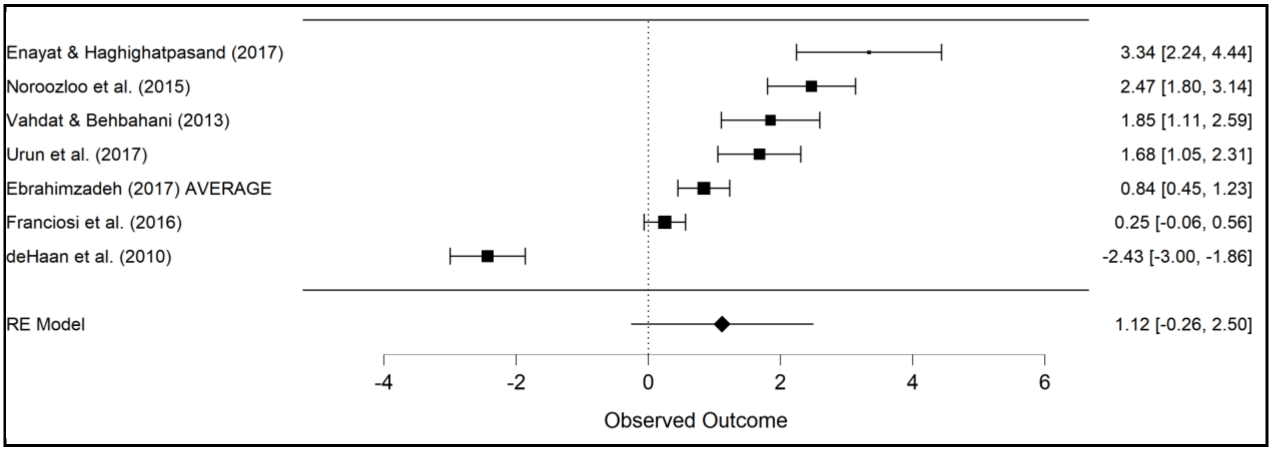
Within-Group Moderator Effects

Contrasts	***k***	***M**$_{d(weighted)}$*	***M**$_{d(unweighted)}$*	***SE***	**95% CI**	
					Lower	***Upper***
Game purpose						
Entertainment	12	2.03	1.31	0.368	0.59	2.04
Education	17	1.13	1.10	0.150	0.80	1.39
Players						
MMO	9	0.59	0.93	0.272	0.40	1.47
Single player	17	1.45	1.44	0.217	1.01	1.86
Dependent variables						
Vocabulary	19	1.64	1.46	0.210	1.05	1.87
Listening	2	1.84	2.15	1.342	-0.48	4.78
Other	8	0.79	0.53	0.159	0.22	0.98
Game input						
Written	5	2.27	1.35	0.600	0.17	2.52
Written and spoken	20	1.15	1.09	0.139	0.82	1.37
Player output						
None	14	1.60	1.61	0.244	1.13	2.09
Written	4	0.98	0.82	0.237	0.35	1.28
Spoken	4	0.66	0.54	0.278	-0.01	1.08
Written and spoken	7	0.94	1.12	0.523	0.10	2.15
Teacher mediation						
Yes	7	2.09	1.63	0.354	0.94	2.32
No	22	1.05	0.99	0.162	0.67	1.30

Note. *k* = number of samples; *M*$_{d(weighted)}$ = mean of effect sizes weighted by sample size; *SE* = standard error; CI = confidence interval. 95% CIs are around the unweighted *d* values; MMO = massively multiplayer online games.

Figure 3

Forest Plot of Between-Groups Effects Among Games Designed for Entertainment

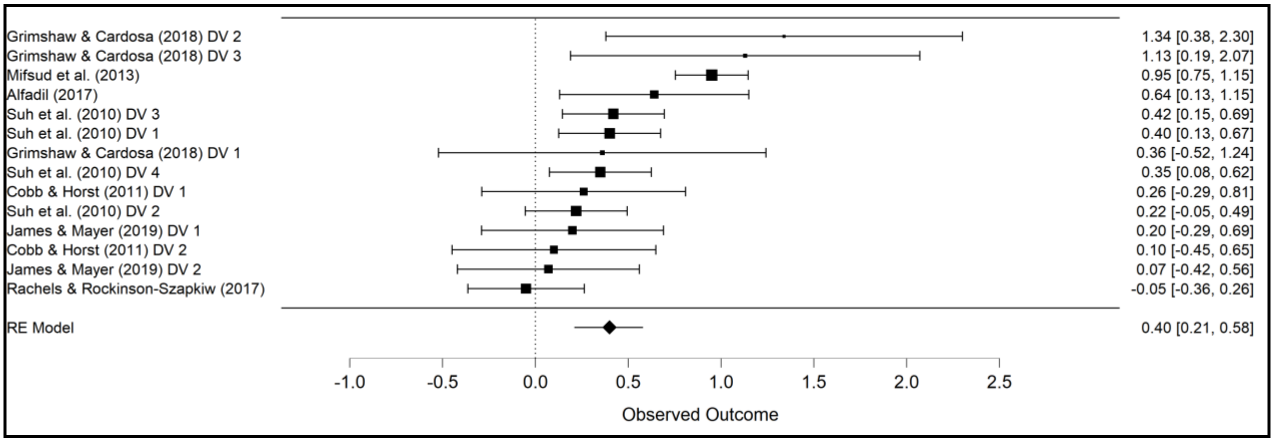


Note. DV = Dependent Variable; Since some studies measured multiple dependent varialbes, each unique DV was separated in the analyses.

RE = Random Effects; Random effects models were used over fixed effects models as described in the Method section above.

Figure 4

Forest Plot of Between-Groups Effects Among Games Designed for Education

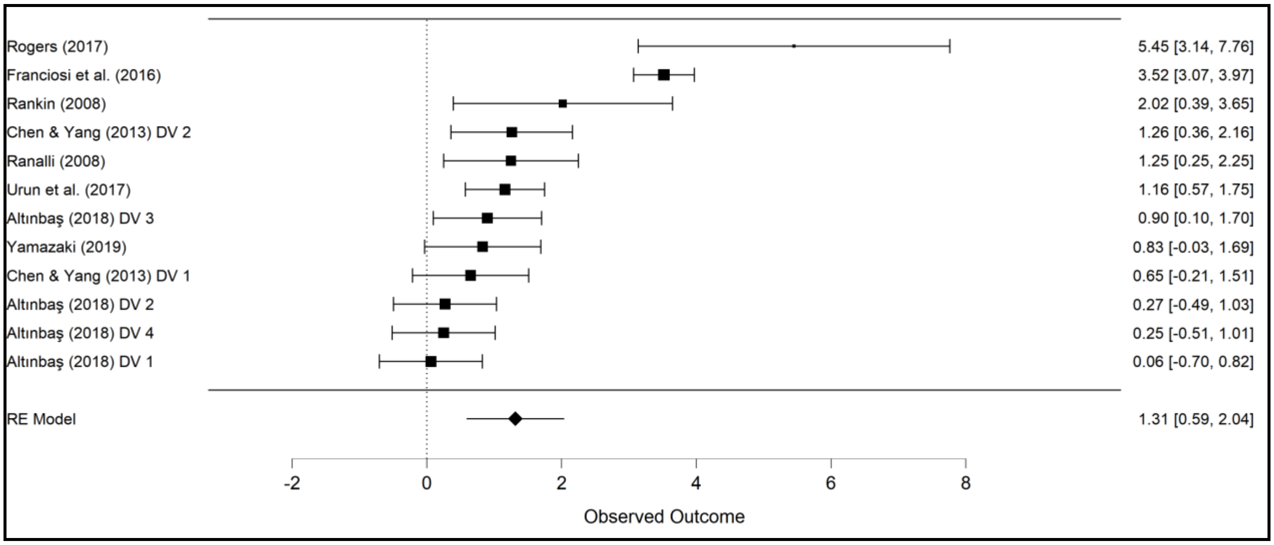


Note. DV = Dependent Variable; Since some studies measured multiple dependent varialbes, each unique DV was separated in the analyses.

RE = Random Effects; Random effects models were used over fixed effects models as described in the Method section above.

Figure 5

Forest Plot of Within-Group Effects Among Games Designed for Entertainment

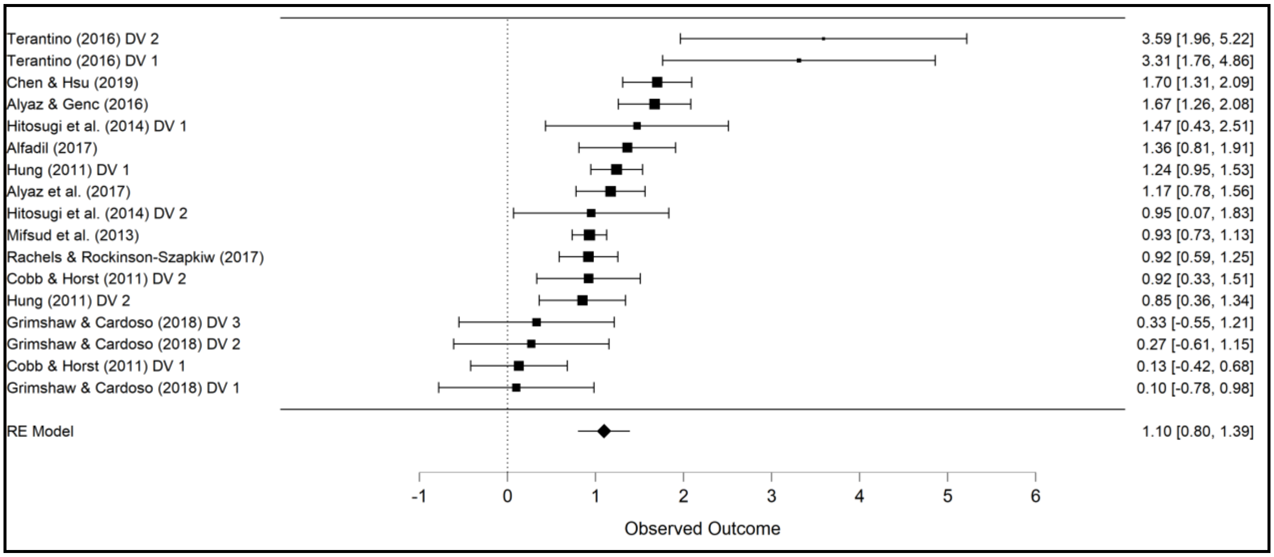


Note. DV = Dependent Variable; Since some studies measured multiple dependent varialbes, each unique DV was separated in the analyses.

RE = Random Effects; Random effects models were used over fixed effects models as described in the Method section above.

Figure 6

Forest Plot of Within-Group Effects Among Games Designed for L2 Education



Note. DV = Dependent Variable; Since some studies measured multiple dependent varialbes, each unique DV was separated in the analyses.

RE = Random Effects; Random effects models were used over fixed effects models as described in the Method section above.

Player interaction may have had a moderating effect, as single player games had greater effects than MMO games in within-group comparisons (Table 5). Conversely, multiplayer games saw greater effects when compared to single player games for between-groups comparisons (Table 4). It should be noted that among between-groups study samples, only four samples made use of multiplayer games, and there were insufficient studies using MMO games so they could not be included in the analysis. For within-group comparisons, there were insufficient multiplayer (non-MMO) samples to aggregate these contrasts. The results limit inferences being made regarding the effects of player interaction on L2 learning, which supports Sundqvist's (2019) call for more research on the effects of player interaction in DGBLL contexts.

Further, the moderator analysis of dependent variables is severely limited by the lack of diversity in the studies sampled. The most common dependent variable in both between-groups (Table 4) and within-group (Table 5) comparisons was, by far, vocabulary acquisition, which produced medium to large positive effects. Other measures of L2 learning such as fluency, grammar, listening, speaking, reading, and writing skills were addressed by individual studies but not with enough frequency to allow for comparisons.

Language input from the game does not appear to produce greatly differing effects. There was little difference between games that included purely textual input and games with both text and speech (Table 4), and this contrast had much overlap in the 95% CIs. Within-group effect estimates (Table 5) indicated that games with only written input had a higher $d_{weighted}$ value of 2.27 when compared to games that had both written and spoken input ($d_{weighted}$ = 1.15). Games that required players to produce L2 output saw some inconsistency. For between-groups, games requiring only spoken output had the highest $d_{weighted}$ value of 1.29; however, there were only four samples in this category. For within-group studies, games requiring no output had the highest weighted $d_{weighted}$ value of 1.60 compared to those that required either spoken, written, or both types of output.

As for teacher-mediated supplementary material, the moderator analysis revealed that the gains were about twice as great when supplementary material was given to learners in addition to the game treatment for both types of research designs. However, overlapping CIs suggest more research is needed before definitive conclusions can be drawn. Nevertheless, findings do support those of Ranalli (2008) in that developing supplementary materials appear to be well worth the effort for teachers.

Discussion

As of January 2020, the body of research corroborates the conclusion that instructional treatments employing digital games can effectively support language learning, particularly vocabulary acquisition. Overall, DGBLL compares favorably to instruction that does not use digital games as seen by the aggregated effect sizes reported in Table 2. However, those advantages decrease over time as evidenced by the small delayed-posttest effects aggregated by the retention of the instructional effect reported in Table 3.

Moderator analyses suggest that games designed for entertainment may offer advantages over those designed for L2 education, although entertainment games produce much wider CIs (see Tables 4 and 5). The larger effect for entertainment games may be explained by the concerns raised about educational games lacking meaningful and engaging L2 interaction (Reinhardt, 2019; Thorne et al., 2012). At the same time, the effect sizes associated with entertainment games varied widely, which may be related to the differences in the designs of the games. For example, the most influential outlier (deHaan et al., 2010) had a very large negative effect (d = -2.43). They reported that participants watching an entertainment game, *PaRappa the Rapper 2* (NanaOn-Sha & Sony Interactive Entertainment, 2001), learned more L2 vocabulary than those actually playing the game. Certainly, the L2 learning potential of a game is greatly affected by whether or not attention to the game's language input is required. To this point, deHaan et al. (2010) noted that "interactivity hindered the language acquisition process" and that "English was not crucial for gameplay [which] may have contributed to the results" (p. 85). In the case of between-groups comparisons, if the deHaan et al. (2010) data were removed, the $d_{weighted}$ value for entertainment game treatments would rise to 1.19 from 0.66, the mean d would rise to 1.67 from 1.12, the standard error would be reduced to 0.443, and the CIs would be tighter without crossing zero (95% CIs: [0.80, 2.54]). This rise would be statistically significant, providing stronger support to the claim that entertainment games can be better suited for L2 learning. Such findings illustrate the importance of considering the various designs and mechanics within digital games for DGBLL contexts.

Outside of the designs of the games, supplementary material appears to have a clear advantage over simply having learners play a game without any teacher mediation. Effect size differences were twice as high for designs that included supplementary material versus those that did not. The vocabulary moderator analysis in the current study compares favorably to past meta-analyses discussed in the Literature Review section. Chen et al. (2018) reported an overall effect size of 0.78 (a medium effect size), and Tsai and Tsai (2018) reported a d value of 0.99 (a large effect size) on L2 vocabulary acquisition in DGBLL contexts for between-groups research designs. In the current study, gaming was found to have a medium effect for between-groups designs (d = 0.87), although the weighted effect size is slightly smaller at 0.59 (a small to medium effect size). In short, based on the findings of the current and previous meta-analyses, DGBLL practitioners can expect a medium to large positive effect on L2 vocabulary. Despite the complementary findings with respect to L2 vocabulary, it is interesting to note that there was very little overlap with studies included in the current meta-analysis and the previous two: only three of 26 studies in Tsai and Tsai (2018) were found to meet our inclusion criteria, while only two studies overlapped with Chen et al. (2018). This small overlap demonstrates the large difference in study eligibility criteria in meta-analyses generally and, more specifically, the various operational definitions that digital games have had across DGBLL studies.

The type of player interaction (single player, multiplayer, MMO) saw mixed results in our analysis, which does not fully support earlier hypotheses that the L2 learning potential of games increases with a greater number of players interacting at one time (Sundqvist, 2013). The between-groups analysis lends some support to this hypothesis with multiplayer games showing a medium to large effect ($d_{weighted}$ = 0.87) on L2 learning, which was greater than the medium effect of single player games ($d_{weighted}$ = 0.55). In contrast, the moderator analysis for within-group studies found single player games to be far more effective than MMOs with a $d_{weighted}$ value of 1.45 for single player and a much smaller $d_{weighted}$ value of 0.59 for MMOs, which does not support Sundqvist's hypothesis. One reason for the greater effectiveness of single player games compared to MMOs may be due to the increased input and cognitive demands typical of MMOs, stemming from the need to communicate with other players. This increased cognitive demand, as Reinders and Wattana (2012) note, may prevent the allocation of mental resources and hinder L2 acquisition. Further, MMOs tend to have communication take place in real time, whereas single player games tend to allow time to be paused, giving players more time to process the input. This time-control mechanic has been "recognized as useful" for L2 learners (Reinhardt, 2019, p. 121), although the current literature is limited in drawing conclusive inferences which demonstrates the need for more quantitative research investigating the effects of player interaction as well as time-controlling mechanics on L2 learning outcomes.

Related to the processing circumstances of games, our moderator analysis of input and output also had mixed results. For within-group designs, games requiring no output from participants showed the greatest positive effect ($d_{weighted}$ = 1.60) on L2 learning outcomes. Games requiring only written output were less effective but still had a small to medium positive effect ($d_{weighted}$ = 0.98), while games requiring only spoken output had the smallest effect ($d_{weighted}$ = 0.66). Although these results seem to suggest that games requiring no output are more effective, this moderator analysis also found that games requiring both spoken and written output outperform games requiring only spoken or only written output. Studies using between-groups designs further complicate definitive interpretations of these results. For between-groups, games requiring only spoken output had the greatest positive effect ($d_{weighted}$ = 1.29), followed by those requiring no output ($d_{weighted}$ = 0.73), while those requiring both spoken and written were found to be the least effective ($d_{weighted}$ = 0.27). There were not enough samples in this contrast to aggregate effects from games requiring only written output. As for input, results from the analyses were mixed as well. Between-groups designs saw games with both written and spoken input as more effective than those with only written input, but within-group designs saw the opposite results (see Tables 4 and 5).

Direction for Future DGBLL Research

Given that few studies measured L2 gains outside of vocabulary knowledge, it is unsurprising that differences in language input and player output showed little difference in the analyses. It is feasible that certain aspects of language learning, such as oral skills, would be greatly benefitted by games featuring oral input and output; however, such effects were not observed due in part to L2 vocabulary knowledge standing as the dependent variable of choice in DGBLL research. Although some studies have investigated fluency and core language skills, there is a need for more studies measuring learning outcomes outside of vocabulary to determine other skill-specific effects from DGBLL.

Finally, we have several recommendations for adding transparency and interpretability in DGBLL research. We found that only around 27% of the studies reported dependent variable reliability and about 20% of the studies discussed assumptions related to their statistical analyses. Just over a quarter of studies reported an effect size, and those that did report effect sizes tended to offer little or no interpretation. Addressing these issues would add to the methodological rigor of DGBLL research. In addition, we want to strongly encourage DGBLL researchers to provide detailed descriptions, or at least a list of the games used in their studies (see Appendix B). Unfortunately, some studies did not report the title of the game used and, thus, could not be included in our analyses. Furthermore, many studies were not included in our analysis because the studies measured effects from games that were not publicly available at the time of data collection. We again want to reiterate that we are not in any way discouraging researchers from publishing results from games that are still in development as the beta-testing process is a requirement for releasing a stable product; however, it would be beneficial if these games, in their current state, were made available at the time of publication to allow for independent assessments.

An additional area that could benefit from more transparency is the length of treatment. Originally, we intended to code for length of treatment as a moderator variable, but the reporting of length was inconsistent across studies. The vague reporting practices ran the gamut from "two weeks" to "five sessions," making it difficult to convert the treatment length to a single unit of measurement, such as the number of treatment hours.

Conclusion

In sum, results lend strong initial support to the claim that digital games can be largely effective for L2 acquisition despite some of the moderator analyses having inconclusive results. Games that were designed for entertainment were found to be generally more effective than those developed specifically for L2 education. Therefore, it seems that L2 educational games could benefit from incorporating more engaging and authentic language interaction, potentially increasing their products' effectiveness. Single player games were generally found to be more effective than MMOs but less effective than multiplayer (non-MMO) games. This finding does not fully support previous hypotheses that more players involved in gameplay at one time have greater potential for L2 learning (Sundqvist, 2019). Finally, DGBLL studies incorporating supplementary learning material had clear advantages over those that simply had learners playing a game without teacher mediation; the extent of the differences (see Tables 4 and 5) suggests that the time that it takes to develop such material is likely well worth the effort.

Given that our initial search resulted in 98 studies examining the effect of DGBLL on L2 gains, it seems that L2 gaming will continue to be a popular domain of research, especially given that the game industry at large is expected to continue to see massive growth (Palandrani, 2021). Our analyses revealed a number of exciting opportunities for future DGBLL research, which can add greater precision to our collective understanding of the extent to which various digital games and their many designs and mechanics affect L2 acquisition. Future research could interpret these quantitative findings against the backdrop of qualitative case studies, an endeavor that we did not undertake in the current meta-analysis but strongly encourage future research to explore.

References

*** denotes studies included in the meta-analysis.**

*Alfadil, M. M. (2017). *Virtual reality game classroom implementation: Teacher perspectives and student learning outcomes* [Doctoral dissertation, University of Northern Colorado]. Scholarship & Creative Works @ Digital UNC. https://digscholarship.unco.edu/cgi/viewcontent.cgi?article=1406&context=dissertations

*Altınbaş, M. E. (2018). *The use of multiplayer online computer games in developing EFL skills* [Master's thesis, Middle East Technical University). OpenMETU. https://open.metu.edu.tr/handle/11511/27368

*Alyaz, Y., & Genc, Z. B. (2016). Digital game-based language learning foreign language teacher education. *Turkish Online Journal of Distance Education*, *17*(4), 130–146.

*Alyaz, Y., Spaniel-Weise, D., & Gursoy, E. (2017). A study on using serious games in teaching German as a foreign language. *Journal of Education and Learning*, *6*(3), 250–264. http://doi.org/10.5539/jel.v6n3p250

ArenaNet. (2012). *Guild Wars 2* [Digital game]. ArenaNet, NCSOFT.

Blair, E., & Blair, J. (2015). *Applied survey sampling*. Sage.

Blizzard Entertainment. (2004). *World of Warcraft* [Digital game]. Blizzard Entertainment.

*Chen, H.-J. H., & Hsu, H.-L. (2019). The impact of a serious game on vocabulary and content learning. *Computer Assisted Language Learning*, *33*(7), 811–832. https://doi.org/10.1080/09588221.2019.1593197

*Chen, H.-J. H., & Yang, T.-Y. C. (2013). The impact of adventure video games on foreign language learning and the perceptions of learners. *Interactive Learning Environments*, *21*(2), 129–141. https://doi.org/10.1080/10494820.2012.705851

Chen, M.-H., Tseng, W.-T., & Hsiao, T.-Y. (2018). The effectiveness of digital game-based vocabulary learning: A framework-based view of meta-analysis. *British Journal of Educational Technology*, *49*(1), 69–77. https://doi.org/10.1111/bjet.12526

*Cobb, T., & Horst, M. (2011). Does Word Coach coach words? *CALICO Journal*, *28*(3), 639–661. https://doi.org/10.11139/cj.28.3.639-661

*deHaan, J., Reed, W. M., & Kuwanda, K. (2010). The effect of interactivity with a music video game on second language vocabulary recall. *Language Learning & Technology*, *14*(2), 74–94. http://doi.org/10125/44215

Dixon, D. H. (2021). The linguistic environments of digital games: A discriminant analysis of language use in game mechanics. *CALICO Journal, 39*(2). https://doi.org/10.1558/cj.20860

Dixon, D. H., & Christison, M. A. (2021). L2 gamers' use of learning and communication strategies in massively multiplayer online games (MMOs): An analysis of L2 interaction in virtual online environments. In K. Kelch, P. Byun, S. Safavi, & S. Cervantes (Eds.), *CALL Theory Applications for Online TESOL Education* (pp. 296–321). IGI Global. https://doi.org/10.4018/978-1-7998-6609-1.ch013

*Ebrahimzadeh, M. (2017). Readers, players, and watchers: EFL students' vocabulary acquisition through digital video games. *English Language Teaching*, *10*(2), 1–18. http://doi.org/10.5539/elt.v10n2p1

*Enayat, M. J., & Haghighatpasand, M. (2017). Exploiting adventure video games for second language vocabulary recall: A mixed-methods study. *Innovation in Language Learning and Teaching*, *13*(1), 61–75. https://doi.org/10.1080/17501229.2017.1359276

*Franciosi, S., Yagi, J., Tomoshige, Y., & Ye, S. (2016). The effect of a simple simulation game on long-term vocabulary retention. *CALICO Journal*, *33*(3), 355–379. http://doi.org/10.1558/cj.v33i2.26063

Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, *16*(3), 283–302. https://doi.org/10.1017/S0272263100013097

*Grimshaw, J., & Cardoso, W. (2018). Activate space rats! Fluency development in a mobile game-assisted environment. *Language Learning & Technology*, *22*(3), 159–175. https://doi.org/10125/44662

*Hannibal Jensen, S. (2017). Gaming as an English language learning resource among young children in Denmark. *CALICO Journal*, *34*(1), 1–19. http://doi.org/10.1558/cj.29519

*Hitosugi, C. I., Schmidt, M., & Hayashi, K. (2014). Digital game-based learning (DGBL) in the L2 classroom: The impact of the UN's off-the-shelf videogame, Food Force, on learner affect and vocabulary retention. *CALICO Journal*, *31*(1), 19–39. http://doi.org/10.11139/cj.31.1.19-39

*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* [Doctoral dissertation, Teachers College, Columbia University]. ProQuest Dissertations Publishing. https://www.learntechlib.org/p/123523/

*James, K. K., & Mayer, R. E. (2019). Learning a second language by playing a game. *Applied Cognitive Psychology*, *33*(4), 669–674. https://doi.org/10.1002/acp.3492

JASP Team. (2021). *JASP* (Version 0.13.1) [Computer software]. https://jasp-stats.org/

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

Loewen, S., & Plonsky, L. (2016). *An A-Z of applied linguistics research method*s. Palgrave.

*Mifsud, C. L., Vella, R., & Camilleri, L. (2013). Attitudes towards and effects of the use of video games in classroom learning with specific reference to literacy attainment. *Research in Education*, *90*(1), 32–52. http://doi.org/10.7227/RIE.90.1.3

NanaOn-Sha & Sony Interactive Entertainment. (2001). *PaRappa the Rapper 2* [Digital game]. Sony Computer Entertainment.

*Noroozloo, N., Ahmadi, S. D., & Gholami Mehrdad, A. (2015). The effect of using a digital computer game (SIMS) on children's incidental English vocabulary learning. *Cumhuriyet Science Journal (CSJ)*, *36*(3), 1991–2000. http://dergipark.org.tr/en/pub/cumuscij/issue/45132/564512

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. https://doi.org/10.1017/S0267190510000115

Palandrani, P. (2021, March 9). Video games & Esports: Building on 2020's rapid growth. *Global X*. https://www.globalxetfs.com/video-games-esports-building-on-2020s-rapid-growth

Peterson, M. (2016). The use of massively multiplayer online role-playing games in CALL: An analysis of research. *Computer Assisted Language Learning*, *29*(7), 1181–1194. https://doi.org/10.1080/09588221.2016.1197949

Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Plonsky, L., & Zhuang, J. (2019). A meta-analysis of L2 pragmatics instruction. In N. Taguchi (Ed.), The *Routledge handbook of second language acquisition and pragmatics* (pp. 287–307). Routledge.

*Rachels, J. R., & Rockinson-Szapkiw, A. J. (2017). 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

*Ranalli, J. (2008). Learning English with *The Sims*: Exploiting authentic computer simulation games for L2 learning. *Computer Assisted Language Learning, 21*(5), 441–455. https://doi.org/10.1080/09588220802447859

Rankin, Y. A. (2008). Design and evaluation of massive multiplayer online role playing games that facilitate second language acquisition* [Doctoral dissertation, Northwestern University]. https://www.proquest.com/openview/9f48d15795ac033f86ef9312ee7b0f14/1?pq-origsite=gscholar&cbl=18750

Reinders, H., & Wattana, S. (2012). Talk to me! Games and students' willingness to communicate. In H. Reinders (Ed.), *Digital games in language learning and teaching* (pp. 156–188). Palgrave.

Reinhardt, J. (2019). *Gameful second and foreign language teaching and learning: Theory, research, and practice*. Palgrave Macmillan.

Reinhardt, J. (2021). Not all MMOs are created equal: A design-informed approach to the study of L2 learning in multiplayer online games. In M. Peterson, K. Yamazaki, & M. Thomas (Eds.), *The state of play: Digital games and language learning* (pp. 69–88). Bloomsbury.

Reinhardt, J. & Thorne, S. (2016). Metaphors for digital games and language learning. In F. Farr & L. Murray (Eds.), *Routledge handbook of language learning and technology* (pp. 415–430). Routledge.

*Rogers, S. A. (2017). *A MMORPG with language learning strategic activities to improve English grammar, listening, reading, and vocabulary* (UMI No. 10265484) [Doctoral dissertation]. ProQuest Dissertations and Theses. https://www.academia.edu/35805995/A_Massively_Multiplayer_Online_Role_playing_Game_with_Language_Learning_Strategic_Activities_to_Improve_English_Grammar_Listening_Reading_and_Vocabulary

Sackett, D. L. (1979). Bias in analytical research. *Journal of Chronic Diseases*, *32*(1–2), 51–63. https://doi.org/10.1016/0021-9681(79)90012-2

*Suh, S., Kim, S. W., & Kim, N. J. (2010). Effectiveness of MMORPG-based instruction in elementary English education. *Journal of Computer Assisted Learning*, *26*(5), 370–378. https://doi.org/10.1111/j.1365-2729.2010.00353.x

Sundqvist, P. (2013). The SSI model: Categorization of digital games in EFL studies. *European Journal of Applied Linguistics and TEFL*, *2*(1), 89–104.

Sundqvist, P. (2019). Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning & Technology*, *23*(1), 87–113. https://doi.org/10125/44674

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

Thorne, S. L., Fischer, I., & Lu, X. (2012). The semiotic ecology and linguistic complexity of an online game world. *ReCALL*, *24*(3), 279–301. https://doi.org/10.1017/S0958344012000158

Tsai, Y., & Tsai, C. (2018). Digital game-based second-language vocabulary learning and conditions of research designs: A meta-analysis study. *Computers & Education*, *125*, 345–357. https://doi.org/10.1016/j.compedu.2018.06.020

*Urun, M. F., Aksoy, H., & Comez, R. (2017). Supporting foreign language vocabulary learning through Kinect-based gaming. *International Journal of Game-Based Learning*, *7*(1), 20–35. https://doi.org/10.4018/IJGBL.2017010102

*Vahdat, S., & Behbahani, A. R. (2013). The effect of video games on Iranian EFL learners' vocabulary learning. *The Reading Matrix*, *13*(1), 61–71. https://www.researchgate.net/publication/301340668_The_effect_of_video_games_on_Iranian_EFL_learners'_vocabulary_learning

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

*Yamazaki, K. (2018). Computer-assisted learning of communication (CALC): A case study of Japanese learning in a 3D virtual world. *ReCALL*, *30*(2), 214–231. https://doi.org/10.1017/S0958344017000350

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*(3), 233–249. https://doi.org/10.1080/09523980600641445

Appendix A. Interrater Reliability

Feature coded	**Percentage agreement**	**Cohen's kappa**
Intended purpose of the game (entertainment vs. educational)	100	1
Player interaction (MMO, multi, single)	96.2	0.913
Input received	84.6	0.57
Output produced	88.5	0.817
Dependent variable (e.g., vocabulary, writing, listening)	93	0.877
Study design (within-group vs. between-groups)	100	1
Control group pretest *N*	93.3	0.996
Control group pretest *M*	85.7	0.994
Control group pretest *SD*	84.6	0.992
Control group posttest *N*	92	0.999
Control group posttest *M*	95.8	1
Control group posttest *SD*	95.8	0.998
Control group delayed posttest *N*	85.7	0.984
Control group delayed posttest *M*	100	1
Control group delayed posttest *SD*	85.7	0.941
Treatment or within-group pretest *N*	93.8	0.854
Treatment or within-group pretest *M*	93.5	1
Treatment or within-group pretest *SD*	93.5	1
Treatment or within-group posttest *N*	95.3	1
Treatment or within-group posttest *M*	95.2	0.995
Treatment or within-group posttest *SD*	97.6	1
Treatment or within-group delayed posttest *N*	88.9	0.99
Treatment or within-group delayed posttest *M*	88.9	0.973
Treatment or within-group delayed posttest *SD*	88.9	0.971
Instrument reliability reporting	95	0.89
Effect size reporting	90.9	0.792
Average	92.68	0.95

Appendix B. Games Used in Studies Included in the Meta-Analysis

Title of game	Studies
3rd World Farmer	Franciosi et al. (2016)
Adventure German – A Mysterious Mission (2013)	Alyaz & Genc (2016)
Adventure German: The Mystery of the Nebra Sky Disc	Alyaz et al. (2017)
BONE	Chen & Yang (2013)
The ClueFinders Reading Adventures: The Mystery of the Missing Amulet	Mifsud et al. (2013)
Counter Strike; League of Legends; Player Unknown's Battlegrounds; RuneScape; Seafight	Altınbaş (2018)
Duolingo	Rachels & Rockinson-Szapkiw (2017); James & Mayer (2019)
Ed-Wonderland Version 2.0	Hung (2011)
Everquest II	Rankin (2008); Rogers (2017)
Food Force	Hitosugi et al. (2014)
House of Language	Alfadil (2017)
Meet-Me	Yamazaki (2018)
Nori School MMO	Suh et al. (2010)
PaRappa the Rappa 2	deHaan et al. (2010)
Playing History (Chapter "Slave Trade")	Chen & Hsu (2019)
Runaway: A Road Adventure	Vahdat & Behbahani (2013)
The Secret of Monkey Island - Special Edition	Enayat & Haghighatpasand (2017)
The Sims	Noroozloo et al. (2015); Ranalli (2008)
Spaceteam ESL	Grimshaw & Cardoso (2018)
Spanish Smash, LinguPinguin, Busuu Kids, Bilingual Child Bubbles, Bilingual Child	Terantino (2016)
Tom Clancy's Ghost Recon: Future Soldier	Urun et al. (2017)
Warcraft III: The Frozen Throne	Ebrahimzadeh (2017)
Word Coach	Cobb & Horst (2011)

Appendix C. Publication Bias

To draw accurate inferences about the effects observed, the possibility of publication bias should be addressed. Despite the five unpublished studies included in the analysis, it is still possible that studies with smaller and/or non-statistically significant effects will be underrepresented in the available literature (i.e., availability bias). Figures C1 and C2 display scatterplots of study effect sizes (x axis) and corresponding sample sizes (y axis), with the estimated overall effect represented by a red dashed line. Given the assumption that these studies are measuring comparable constructs with comparable effects, these plots should resemble a triangular funnel plot: studies with larger sample sizes should converge on the true effect, whereas studies with smaller sample sizes should show greater variance. The general positive skew of the between-groups effects (Figure C1) suggests the most common form of publication bias. That is, studies which observed nonsignificant effects may not have reached publication (Norris &

Ortega, 2000, p. 431). The within-group effects (Figure C2) also appear positively skewed, though to a lesser degree. This is unsurprising given observations that within-group effects may be comparatively less prone to publication bias (Sackett, 1979, p. 59). Participants often see measurable improvement from pretest to posttest regardless of treatment; hence, the results are more often deemed significant and less likely to be suppressed.

Figure C1

Between-Groups Effects and Sample Sizes Scatterplot

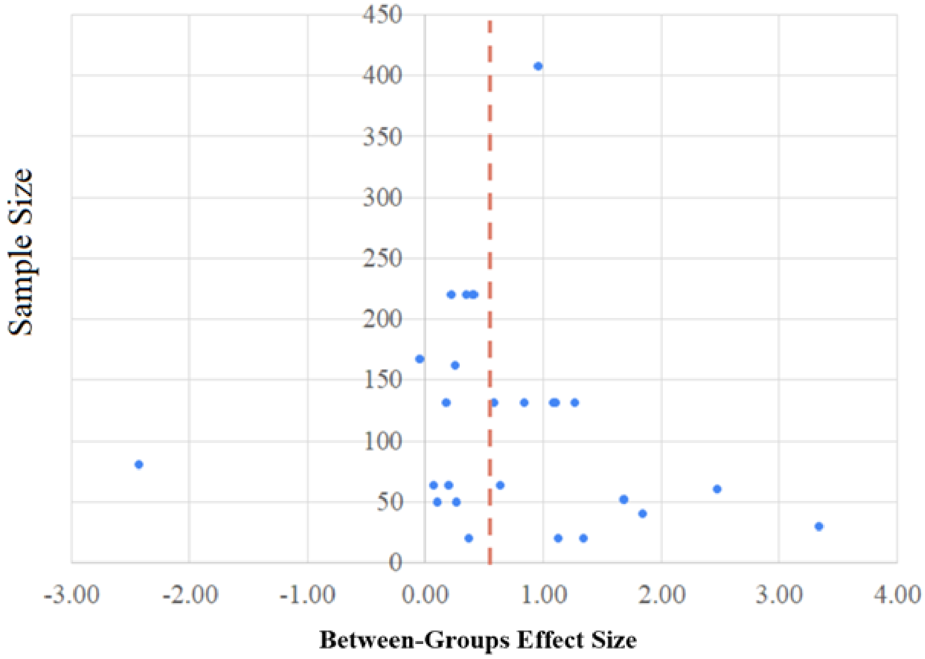
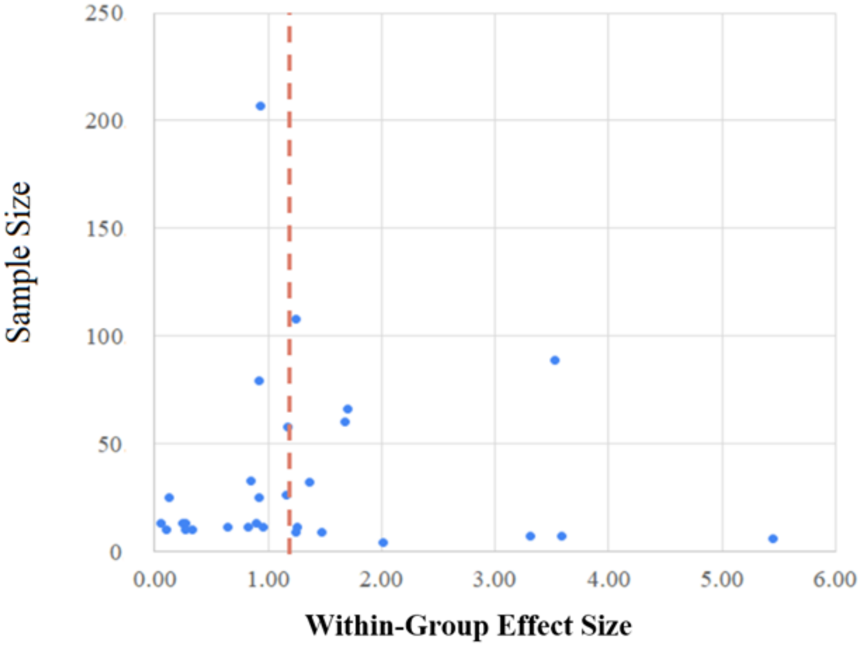


Figure C2

Within-Group Effects and Sample Sizes Scatterplot


Sample Size
Within-Group Effect Size