Review of Educational Research
Month 202X, Vol. XX, No. X, pp. 1–47
DOI: 10.3102/00346543251389930
Article reuse guidelines: sagepub.com/journals-permissions
© 2025 AERA. https://journals.sagepub.com/home/rer

Virtual Characters Help K–12 Students Learn and Improve Motivation: A Meta-Analysis

Noah L. Schroeder
University of Florida

Shan Zhang
University of Florida

Chris Davis Jaldi
Wright State University

Jessica R. Gladstone
University of Illinois Urbana-Champaign

Alexis A. López
Educational Testing Service

Emmanuel Dorley
University of Florida

Pedagogical agents, conversational agents, motivational agents, and other virtual characters have long been used in educational technologies. We built and analyzed the most comprehensive dataset to date of studies examining how virtual characters influence K–12 students' learning and learning-related outcomes using three-level meta-analytic procedures. The results from five three-level meta-analyses indicate that virtual characters helped K–12 students learn ($g = 0.42$, $p < .001$, $k = 70$) and improved their motivation ($g = 0.48$, $p = .001$, $k = 47$) but did not have any significant effects on emotions ($g = 0.60$, $p = .20$, $k = 15$), perceptions ($g = 0.05$, $p = .88$, $k = 34$), or cognitive load ($g = -0.09$, $p = .84$, $k = 5$) compared to systems without a virtual character present. We conclude that virtual characters can provide a meaningful addition to learning environments for K–12 learners.

Keywords: instructional technologies, instructional design/development, computers and learning, pedagogical agent, K–12 education, human-computer interaction, virtual human, conversational agent, meta-analysis

Educational games, mobile phone applications, and web-based learning programs often use virtual characters (VCs) in their learning materials. In fact, there are services that will use generative artificial intelligence to not only create an educational video but also generate a human-like VC to present the material (Deepbrain AI, 2023). Although research around the use of VCs has proliferated over the last 25 years (Siegle et al., 2023), the literature presents conflicting views as to whether they are effective for supporting learning, and if so, for whom and under what conditions. Despite this, we continue to see VCs appear in more educational technologies as time goes on. As such, it is important to understand their effects on learners, particularly at the K–12 level, where students are developing a broad, foundational knowledge base that will support them for the rest of their lives.

Virtual characters can come in many different forms and can play many different roles in the learning environment (Clarebout et al., 2002; Schroeder & Gotch, 2015). Recent studies have examined various forms of VCs, such as pandas (Jing et al., 2022) and virtual humans (Bøg Petersen et al., 2022). The role of the VC can vary significantly, including lecturing or providing scaffolding via feedback. Siegle et al. (2023) discussed how the term pedagogical agent has typically referred to any type of VC within a learning environment that is designed to support learning. They explained how the term pedagogical agent is a term that encompasses other types of characters, such as conversational agents (that converse with the learner in natural language), motivational agents (that deliver motivational messages to the learner), or virtual humans (human-like characters within the learning environment). It should be of no surprise then that a seminal systematic review of pedagogical agents across all age groups found no significant effects on learning or motivation, concluding that the overall effects of agents were too broad a question, and instead we should focus on more detailed aspects of agent implementation (Heidig & Clarebout, 2011). In this study, we statistically examine the impacts of various aspects of agent design and implementation on learning outcomes following Heidig and Clarebout's (2011) frameworks.

Since the publication of Heidig and Clarebout's (2011) review, there has been a dramatic increase in the number of studies published around the use of VCs, particularly studies conducted with young learners in the K–12 grades. This is important because while Schroeder et al.'s (2013) meta-analysis found that pedagogical agents were particularly effective for K–12 learners, the number of studies in their analysis was quite limited. A more recent meta-analysis by Castro-Alonso et al. (2021) did not find the same advantage for K–12 students compared to other age groups; however, they only analyzed studies published since Schroeder et al.'s analysis, and their sample was also quite small for this age range. As a result, it is not known what the true, overall effects of VCs on

K–12 learners are, leaving critical questions unanswered: What are the effects of VCs on K–12 learners, and equally important, what factors may moderate these effects?

In addition to the conflicting results in existing meta-analyses, the research synthesis in the field of VCs has other notable limitations. First, none of the previously cited reviews focused specifically on K–12 learners. Rather, the inferences from Schroeder et al.'s (2013) and Castro-Alonso et al.'s (2021) reviews were the result of moderator analyses. Second, all of the previously cited reviews used relatively narrow search strings and focused on only one or two outcomes (learning and motivation). While we believe the cited reviews were comprehensive within their stated goals, we do not believe they were as broadly scoped as they could have been. In addition, they omitted studies of other learning-relevant processes and outcomes, such as learners' motivation, emotions, and perceptions, which we know can be critical to the learning process (Pintrich et al., 1993; Sinatra, 2005). Finally, all the existing meta-analyses of pedagogical agent research have used conventional, two-level meta-analytic methods. The inherent limitations of this approach necessarily exclude data that could otherwise be included in a more complex three-level model. As such, the purpose of this study is to examine, in the most comprehensive analysis to date, how VCs influence K–12 learners' learning and learning-relevant outcomes using three-level meta-analytic methods.

Literature Review

Existing meta-analyses and systematic reviews examining the influence of VCs have generally focused on learning outcomes and motivational outcomes (Castro-Alonso et al., 2021; Davis et al., 2023; Guo & Goh, 2015; Heidig & Clarebout, 2011; Peng & Wang, 2022; Schroeder & Adesope, 2014; Schroeder et al., 2013; Wang et al., 2023). Although these outcomes are, of course, important, these analyses omit other learning relevant processes, outcomes, and perceptions that we know can be important for learning. For example, models of conceptual change discuss how various aspects of the learner and their experience, other than only their prior knowledge, can influence learning (Pintrich et al., 1993; Sinatra, 2005). In other words, we believe it is important to consider the "whole learner" rather than a purely cognitive perspective on learning. In this section, we briefly review major theoretical frameworks that have contributed to the VC literature. We then outline what is known about the impact of VCs on a variety of theoretically driven variables that may influence learning.

Theoretical Approaches to Virtual Character Research in Education

Given that the vast majority of VC studies in the field of education are situated around learning outcomes, it is important to consider the theoretical frameworks that have driven this work. While we acknowledge that VC research has been approached from a wide variety of theoretical perspectives, there are two widely cited theories in this area of research we wish to draw attention to due to their popularity within the field: the cognitive theory of multimedia learning (CTML; Mayer, 2014a, 2024) and social agency theory (Atkinson et al., 2005; Mayer et al., 2003). We close the section by discussing a newer theoretical framework that addresses the limitations of CTML and social agency theory.

The CTML is a cognition-oriented theory that describes how we bring in information through both our eyes and ears and emphasizes that we have working memory processes for both verbal and pictorial representations that work together to facilitate deep learning (Mayer, 2014a, 2024). Critical to this theoretical perspective is the concept of a limited working memory, which has been supported by copious amounts of research (Cowan, 2001, 2010; Mayer et al., 2003). According to CTML, it is plausible that VCs may not aid learning because they might create processing that is not relevant to the learning task—a concept that aligns with the "image principle," which states that an image of the instructor does not improve learning (Mayer, 2014b).

One constraint of CTML is that it is specifically focused on cognitive processes and therefore does not really speak to the impacts of social, motivational, meta-cognitive, and other processes. An early attempt to bridge this gap was termed social agency theory (Mayer et al., 2003). According to social agency theory, a learning situation can enact the social conversation schema and thus increase learning (Atkinson et al., 2005; Mayer et al., 2003). This theoretical perspective opened the door to a number of studies around the use and design of VCs in learning environments. However, this theory has not been greatly elaborated on in the last 20 years, perhaps due to its lack of mechanistic specificity in contrast to a theory like CTML. In addition, this theoretical perspective is also limited as it does not consider processes such as motivation, emotions, etc. As a consequence, we do not have much specific instructional design guidance from social agency theory specifically, even if the driving idea behind it is well-supported.

A New Theoretical Perspective

Recently, Schneider et al. (2022) proposed the cognitive-affective-social theory of learning in digital environments (CASTLE). CASTLE provides a more mechanistically specific model of how social and other processes can influence the learning process and, consequently, learning outcomes. Under this conceptual framework, learning is more aligned with what researchers in the area of conceptual change have envisioned for as long as pedagogical agent research has existed—we learn based on our beliefs about ourselves and our environment, our emotions, our social interactions, etc. (Pintrich et al., 1993; Sinatra, 2005). More specifically, CASTLE describes how social cues, such as those from feedback, social comparison, and other social interactions, can influence learning and related processes, such as motivation and emotions (Schneider et al., 2022). Since VCs have long been posited to influence learning due to their socially relevant characteristics and interactions with learners, CASTLE represents a well-grounded theory from which to orient ourselves when exploring how social cues can influence the learning process. Accordingly, in this study, we examine the impact of VCs on all learning-relevant outcomes we identified in the literature base focused on educational or motivational contexts in order to build a more comprehensive understanding of how VCs can influence K–12 students' learning. These theoretical perspectives also support our examination of various moderator variables related to participant characteristics, VC design, and research design, as all of these factors can influence the social aspects of a learning experience. However, before discussing these issues, it is important to understand what is known about the impacts of VCs on K–12 students' learning processes.

The Impacts of Virtual Characters on K–12 Students' Learning

So, what is known about how VCs influence K–12 students' learning? In this section, we explore the variety of learning-relevant outcomes we identified in the literature and briefly review existing findings for each.

Learning

Reviews of VC research generally examine learning outcomes as their primary contribution to the literature. Overall, the literature to date has generally shown that VCs in the form of pedagogical agents can lead to small to moderate effects on learning (Castro-Alonso et al., 2021; Davis et al., 2023; Guo & Goh, 2015; Peng & Wang, 2022; Schroeder et al., 2013; Wang et al., 2023). However, meta-analyses have shown that these effects can be quite nuanced, meaning that the specific use case matters. In this study, we focus on learner age. Schroeder et al.'s (2013) meta-analysis indicated that K–12 learners find more benefit from pedagogical agents than other age groups. They hypothesized three plausible reasons why this may have occurred: 1) K–12 learners could receive more motivational benefits from VCs than older learners, 2) K–12 learners may be more influenced by the social aspects of the interaction than older learners, or 3) it could be a simple novelty effect. However, the nature of their analysis did not allow for concrete conclusions as to the causality of these findings, and the number of K–12 studies in the analysis was relatively small ($k = 6$). Castro-Alonso et al. (2021) analyzed a more recent, but not comprehensive (they reviewed studies published since Schroeder et al.'s work), set of studies and did not find similar results around learner age. It is therefore an open question as to whether VCs are more effective than non-VC conditions for K–12 learners, and what, if any, moderating factors may exist for this population.

Motivation

One rationale for including VCs within a learning environment may be that they may help motivate a learner (van der Meij, 2013). However, reviews of the literature around the extent to which VCs can motivate students have shown mixed results (Heidig & Clarebout, 2011), although two meta-analyses did find small positive effects (Guo & Goh, 2015; Wang et al., 2023). A notable challenge in relation to whether VCs provide a motivating effect is the control group: some advantages are seen compared to text-only control groups, but no notable advantages are seen in relation to voice-only conditions (Schroeder & Adesope, 2014). Furthermore, none of these cited reviews examined the effects of VCs on K–12 students' motivation specifically; rather, these were broad reviews of the literature around the use of VCs. It seems quite plausible that young students may engage differently with VC than older or adult learners due to their cognitive development and interests, and therefore, it remains an open question to what extent VCs may influence K–12 students' motivation.

Emotions

How do VCs influence learners' emotions during and after learning? Research to date is largely inconclusive. Wang et al.'s (2023) meta-analysis found a small, positive effect on learners' positive emotions. Primary studies, however, have found mixed results. For example, Beege and Schneider (2023) studied how VCs expressing enthusiasm influenced German secondary students' emotions and found no significant differences between those who learned with an expressive and non-expressive agent. Meanwhile, Wang et al. (2022) conducted three experiments with undergraduate students and found that happy VCs led to more positive emotions than neutral VCs. It is noteworthy that both of these sample studies were comparing different VC designs rather than the presence or absence of VCs. As such, the influence of a VC itself on young learners' emotions is still largely unknown.

Perceptions

Researchers have long posited that VCs can influence learners' perceptions of their learning experience (Lester et al., 1997), although these results have not always been consistent in the literature (Van Mulken et al., 1998). One challenge associated with the concept of learners' perceptions is the wide variety of perceptions one could be asked about. For example, Sinoo et al. (2018) examined learners' perceptions of friendship with a VC, while Plant et al. (2009) examined learners' perceptions of gender stereotypes. These perceptions stand in contrast to those examined in Daradoumis and Arguedas's (2020) study on students' perceptions of how well a VC enabled their personal growth. This being the case, it is not surprising that no synthesis exists examining how VCs influence learners' perceptions. In this study, we examine perceptions broadly when conducting an overall meta-analysis and intend to categorize learners' perceptions more specifically into coherent subgroupings should data allow.

Cognitive Load

Finally, the concept of cognitive load has been researched inside (Schroeder, 2017; Yung & Paas, 2015) and widely outside (Paas & Sweller, 2014; Sweller, 2010, 2020) of VC research. The simplified premise of cognitive load theory is that mental effort can be allocated to tasks good for learning or extraneous to learning; thus, an instructional designer should minimize extraneous elements in the learning environment (Paas & Sweller, 2014; Sweller, 2010, 2020). In the context of VC research, there has been concern that VCs present a source of extraneous cognitive processing (Clark & Choi, 2007). However, there does not seem to be a clear answer in the literature as to whether or not this is actually the case. Schroeder (2017) examined the effects of a VC in a learning environment compared to an environment without a VC and found that undergraduate students did not perceive a difference in measures of cognitive load. Yung and Paas (2015) conducted a similar study with seventh-grade students and also found no significant results. While these results are consistent, the extant studies in the area are limited, and the research in the area has never been synthesized.

Summary

Researchers have explored a variety of outcome variables as potentially being influenced by VCs, and modern learning theories demonstrate why each may influence the learning. Despite this, it is largely unknown how VCs influence young learners' learning and learning-relevant outcomes, as the work has never been comprehensively synthesized. Furthermore, as Heidig and Clarebout (2011) and Schroeder et al. (2013) demonstrated, the question of whether VCs are effective, while important, is not nuanced enough to really understand the effects of VCs on learning. Rather, we must consider aspects of VC design and implementation.

The Design of Virtual Characters

Researchers have explored a variety of different ways to design VCs, such as manipulating their gender (Plant et al., 2009; Schroeder & Adesope, 2015), voice (Chiou et al., 2020; Craig & Schroeder, 2017), contextual relevance (Veletsianos, 2010), role (Baylor & Kim, 2005), or their human likeness (Jing et al., 2022; Wu et al., 2023). This variety can make it difficult to parse what features are effective for supporting learning and which are not. Fortunately, Heidig and Clarebout (2011) created a model that delineates different decisions that must be made when designing VCs. This framework provides an appropriate starting point for examining what variables may moderate the effects of VCs in learning situations.

Global Level

The global level of design refers to whether the VC is humanlike or not (Heidig & Clarebout, 2011). Social agency theory and CASTLE would posit that we may have different social conversation schemas activated depending on the VC's appearance. For example, a young student may socially engage differently with a virtual human as opposed to an animal-like VC. Accordingly, in our analyses, we consider the high-level design of the VC to see if using a humanlike or non-humanlike appearance influences K–12 learners.

Medium Level

At the medium level of design, Heidig and Clarebout (2011) encourage researchers to consider technical decisions, such as the VC's animation level and voice type, as well as the choice of the character, such as its role in the learning environment. All of these factors can influence how one perceives the VC and may influence the extent to which social conversation schemas are activated. Both social agency theory and CASTLE posit that this can influence learning (Mayer et al., 2003; Schneider et al., 2022). Accordingly, in our analyses, we consider whether the VC used gestures, what type of voice it used, and what type of VC it was characterized as by the authors in order to determine if any of these features moderate the effects of VCs on young learners.

Detail Level

The final level of design is the detail level, which examines appearance-related aspects of the VC, such as age, gender, or clothing (Heidig & Clarebout, 2011).

Similar to the medium level of design, these are critical factors that may influence how the VC is perceived and engaged with socially. For example, research has shown that VCs are stereotyped by learners, and these detail-level factors can influence stereotypes (Veletsianos, 2010). Given the importance of social processes and their influence on learning to social agency theory and CASTLE, in our analyses we examine the age and gender of the VCs used in primary studies to see if these design aspects moderate the effects of VCs on K–12 learners.

The Influence of Study Design on Learning with Virtual Characters

While learners' interactions with VCs can certainly moderate the effects of VCs, as can the design of the VCs, it would be remiss to ignore critical components of study design that could influence the results. For example, a biased participant assignment procedure could influence the findings of a study, or a pretest could sensitize participants to the important content in the learning materials, thereby influencing the results (Cohen et al., 2007).

This being the case, it is in our view critical that meta-analyses consider aspects of study design in their analyses. In this study, we included a variety of variables that have been explored in prior reviews of the literature, as well as a few additional ones that could plausibly moderate the effects of VCs. Specifically, we examine if the learning experience is a collaborative or individual activity, how participants were assigned to conditions, what type of control condition was used, if the control group experienced the same base learning materials as the experimental group, if a pretest was used, and the type of test used and the domain of the test, as well as the domain of the learning materials. We also considered the learners' age and grade level, and if studies were conducted with specific populations (e.g., high or low prior knowledge, learners with autism, or children with specific learning disabilities). Finally, we considered whether studies were between- or within-subject designs, and what type of pacing was used in the learning system. We also created a metric for study quality to see if lower-quality studies were associated with different effects than higher-quality studies (explained in detail in the methods section).

The Present Study

Existing meta-analyses around the impact of VCs on learning are conventional, two-level meta-analyses (Castro-Alonso et al., 2021; Davis et al., 2023; Schroeder et al., 2013; Wang et al., 2023). While these methods are appropriate and useful for summarizing the state of the literature, conventional meta-analytic methods are inherently limited by the fact that the comparisons must be statistically independent from one another. In short, this means that it is possible, and in educational technology fields quite common, that not all data present in the primary study are actually analyzed.

Consider the following example: A study contains two independent groups, a VC group and a non-VC control group, and there are three learning tests, an immediate learning test, a one-week delayed test, and a two-month delayed test. A conventional meta-analysis can only analyze one outcome from this study (or else participants would be counted twice, creating dependencies in the data that the analysis cannot account for). As a result, the meta-analyst is left with a choice: Do they pick one of the tests to use in their analysis (hopefully following a set criterion they set up in advance for such selection), or do they compute a weighted mean and pooled standard deviation of two or more measures? Neither of these approaches is optimal. The former leaves data on the table, while the latter conflates potentially moderating variables between the conditions that were combined.

A solution to this problem is three-level meta-analysis (3LMA). A 3LMA accounts for the dependencies in the data (Assink & Wibbelink, 2016), allowing for multiple comparisons (commonly thought of as multiple outcome measures) from the same participants. Using 3LMA can provide deeper insights into a phenomenon than conventional meta-analysis by not only including all outcome variables of interest due to accounting for the dependencies in the data, but also allowing for less conflated moderator analyses due to not having to weight means and pool standard deviations from multiple comparisons. While conventional meta-analyses are quite common in the field of education, perhaps due to the presence of numerous graphic user interface-based programs to help with their analysis, 3LMA is growing in popularity as more researchers learn about the methodology and become proficient in statistical programming.

In this study, we conduct a series of five 3LMAs to explore the following research questions:

RQ1: What is the impact of VCs on K–12 students' learning, and what moderates this effect?
RQ2: What is the impact of VCs on K–12 students' motivation, and what moderates this effect?
RQ3: What is the impact of VCs on K–12 students' emotions, and what moderates this effect?
RQ4: What is the impact of VCs on K–12 students' perceptions, and what moderates this effect?
RQ5: What is the impact of VCs on K–12 students' cognitive load, and what moderates this effect?

Methods

Transparency and Openness

We followed the PRISMA 2020 guidelines (Page et al., 2021) for systematic reviews and meta-analyses. All of the data and R code used for the analyses are available at https://osf.io/4duxj/?view_only=fe0e91572b354c8691f7f5a3c365 abf8. Specific information about the R packages and analytical approaches used is provided in the data analysis section of the methods. This review was not preregistered.

Literature Search

This meta-analysis is part of a large-scale, multifaceted review of how VCs influence K–12 students' learning. Accordingly, the studies identified for this analysis were based on the literature search from Zhang, Jaldi, Schroeder, and Gladstone's (2024) scoping review.

To briefly summarize the literature search, in September 2023, we searched nine major databases spanning education, psychology, social sciences, computing, and medicine and health profession fields using the search string (“virtual human”* OR ”embodied agent”* OR “virtual character”* OR “pedagogical agent”* OR “conversational agent”* OR “motivational agent”*) AND (K–12 OR elementary OR primary OR secondary OR middle OR high) AND (learn* OR motivat* OR self-efficacy OR self-confidence OR ability belief* OR self-concept OR interest* OR engag* OR value* OR util* OR “sense of belonging” OR belong* OR achiev* OR develop*). The search string was developed based on relevant literature in the area and prominent theoretical perspectives in the field. In addition, we added the relevant studies included in three existing meta-analyses of pedagogical agent research (Castro-Alonso et al., 2021; Schroeder et al., 2013; Wang et al., 2023). After removing duplicates, this resulted in 1374 abstracts for consideration. We used ASReview (https://asreview.nl/) and its natural language processing capabilities to help narrow down our sample of relevant studies during abstract screening (fully described in Zhang, Jaldi, Schroeder, & Gladstone, 2024). Of these 1374 abstracts, 166 full texts were examined, which resulted in 112 studies that met the inclusion criteria for their scoping review.

To locate studies for the meta-analysis, we examined the full text of the 112 studies that were included in Zhang, Jaldi, Schroeder, and Gladstone’s (2024) scoping review and then applied the inclusion and exclusion criteria as follows.

Inclusion and Exclusion Criteria

In order to be included in these meta-analyses, studies had to include a group learning with or from a VC compared to a no-character condition. Importantly, images of actual humans were not considered VCs, meaning that only computer-generated characters are included in this analysis. In addition, studies must report enough data to calculate an effect size, such as means, standard deviations, and sample sizes, or t or F statistics. Studies must also be publicly available through library resources or inter-library loan.

Studies of log file data of student interactions were excluded from the analyses as we were interested in outcome measures rather than process measures (including, but not limited to, eye-tracking data, unless the eye-tracking was measuring a learning outcome as in Grynszpan et al., 2022). We also excluded studies in which the no-character condition contained any part of the instructional sequence that contained a VC (e.g., Johnson et al., 2013).

Study Screening

We began by examining the full text of the 112 studies included in Zhang, Jaldi, Schroeder, and Gladstone’s (2024) scoping review. Of these 112 studies, 30 met our inclusion criteria and were included in the analyses (Figure 1).

Data Extraction

Selecting Relevant Comparisons

While 3LMA accounts for dependencies within the data, we prioritized reducing the number of potentially confounding variables in the dataset. As a result, if more than two groups were present, we only coded those with the least

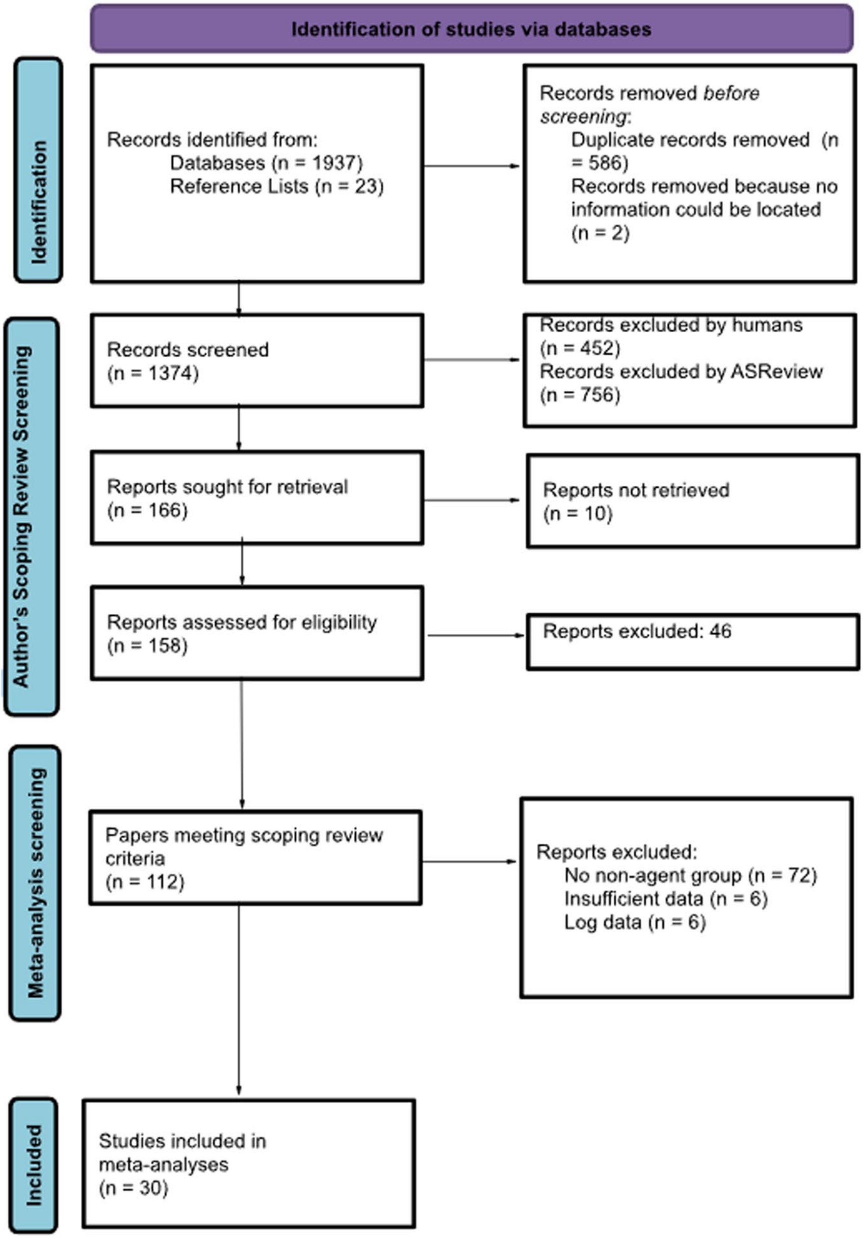
Identification of studies via databases
Identification
Author's Scoping Review Screening
Meta-analysis screening
Included
Records identified from:
Databases (n = 1937)
Reference Lists (n = 23)
Records removed *before screening*:
Duplicate records removed (n = 586)
Records removed because no information could be located (n = 2)
Records screened (n = 1374)
Records excluded by humans (n = 452)
Records excluded by ASReview (n = 756)
Reports sought for retrieval (n = 166)
Reports not retrieved (n = 10)
Reports assessed for eligibility (n = 158)
Reports excluded: 46
Papers meeting scoping review criteria (n = 112)
Reports excluded:
No non-agent group (n = 72)
Insufficient data (n = 6)
Log data (n = 6)
Studies included in meta-analyses (n = 30)

FIGURE 1. *PRISMA diagram, adapted from Page et al. (2021) and Zhang, Jaldi, Schroeder, and Gladstone (2024).*

confounding variables. For example, in Chen and Chen's (2014) study, there were three groups: a control group, a VC group, and a VC + competition group. For this study, we ignored the VC + competition group due to competition representing a confounding variable compared to the control condition. Similarly, Lee and Mustapha (2017) also had three groups in their experiment: a control group, a peer character group, and a teacher character group. For this study, we ignored the peer character group because the teacher character group was more closely aligned with the control group that learned from their teacher. While this approach does decrease some of the data that could have been included in the analyses, it makes for more interpretable analyses by removing potentially confounding variables.

We also encountered a few studies with two experimental groups and one control group, but neither experimental group introduced an additional confounding variable. For example, Plant et al. (2009) included male and female characters in contrast to a control group. In cases like this, we included all groups in our analyses.

Variables Extracted

We extracted many different variables from each study in order to create a comprehensive picture of how VCs may influence K–12 students' learning. We separate these variables into the following categories: study descriptives, study design, system design, character design, assessment design, outcome measure, and effect size data. Each of these categories is described in the following sections. Note that any variable, except effect size data, had an option for "not reported" that was used when the authors did not explicitly provide the information, and it could not be inferred from figures in the study.

Study descriptives. We coded the year the study was published, the publication type (journal article, conference proceeding, dissertation, or other), and the publication name.

Study design. We coded the control condition (software program, traditional teaching, video, robot interaction, or no intervention control group), how participants were randomized to condition (individual, group, not reported), and whether it was a between- or within-subjects research design. We also coded the duration of the intervention (1 hour or less, >1 hour to 2 hours, >2 hours to 5 hours, >5 hours to 10 hours, >10 hours to 20 hours, >20 hours), the duration of the study (<1 day, 1 day to 1 week, >1 week to 4 weeks, >4 weeks), and the domain of the learning materials (mathematics, computer skills, language, science, interview skills, social skills, health). Finally, we also coded the setting (lab or classroom), grade level of the participants aligned with the International Standard Classification of Education United States (US) equivalents (National Center for Education Statistics, n.d.; primary [US grades 1–6], lower secondary US grades 7–9, upper secondary (US grades 10-12)), and the location where the data were collected.

System design. We coded the task type (individual or collaborative), the pacing of the system (learner-paced or system-paced), and the media type (computer, VR, AR, phone, tablet, other). We also coded the number of agents with which the student interacted.

Character design. We coded the type of voice the character had (text-to-speech, human voice, voice not specified, text communication, voice and text communication, student choice between narration or text communication) and if the agent used gestures (yes, no). We also coded the type of agent the authors claimed their VC was (pedagogical, conversational, motivational, multiple agent types), its form (human-inspired, non-human, mixed—multiple agents), gender[1] (male, female, gender neutral, multiple agents, nonhuman), age (child, adult, irrelevant nonhuman, multiple agents—multiple ages, unknown) and the agent's role (demonstrating/modeling, coaching/scaffolding, information source, testing, multiple roles).

Assessment design. We coded the item type (free response, performance, Likert scale, multiple choice, true/false, multiple item types), test type (transfer, retention, self-report, or general learning test), and whether a pretest was present or not (present or absent).

Outcome measure. We coded the type of outcome measure(s) reported in the study (learning, motivation, perceptions, emotions, attitudes, or cognitive load). Importantly, we classified measures as the authors did in primary studies. For example, self-efficacy was considered a measure of motivation, as were constructs such as interest and confidence.

Study quality. When we examined the literature, we were unable to locate a largely agreed-upon measure of study quality. As such, within our team, we came up with a list of qualities we expect to see in studies to meet a "minimum" criterion of causal intervention study quality in the types of studies qualifying for these meta-analyses. We coded whether the control condition had the same learning materials as the experimental group other than the experimental manipulation, if there was an abnormally high attrition rate (we defined this as more than 20% of the sample), if participants were randomized or stratified to condition or not to create equivalent groups, if the outcome measure was clearly understood by the reader,[2] and if there was some form of reliability reported for the outcome measure. Each of these was coded initially as "yes" or "no," which was then transcribed to either a "1" or "0," where 1 indicated the more desired outcome. For example, if both conditions had the same materials except for the experimental manipulation, it was coded as "1." We then calculated an overall quality metric with a maximum score of 5 based on these variables, with 5 being the highest number of quality indicators and 0 being the lowest.

Effect size data. We coded the means, standard deviations, and sample sizes for each comparison included in the analysis. Importantly, we excluded total score variables when the components of those scores were available, as we opted to include the more fine-grained data to create a more insightful analysis. When assessments were negatively valanced, we reversed the effect size to be consistent with other measures included in the analyses.

Inter-rater agreement

Some data used in the analyses reported here were coded by Zhang, Jaldi, Schroeder, and Gladstone (2024), where inter-rater agreement was determined to be 84%. The authors then refined the coding scheme and coded an additional five studies in which their inter-rater agreement was 94% (Zhang, Jaldi, Schroeder, & Gladstone, 2024). Newly coded data for this study, including all effect size information, was coded by one researcher who extracted data from all of the studies. A second author coded 20% of the sample in order to calculate inter-rater agreement. Inter-rater agreement was determined to be 90%. Any discrepancies were reconciled through discussion and by reexamining the primary study.

Data Analysis

We conducted a series of 3LMAs using restricted maximum likelihood estimation and used *t*- and *F*- distributions for making inferences (Viechtbauer, 2022). The analyses were conducted using the metafor package (Viechtbauer, 2010) for R (R Core Team, 2022), and the analysis code was adapted from the code provided by Schroeder (2024). We interpreted effect sizes using Hattie's (2015) criteria for educational studies, with $g = .20$ being a small effect, $g = .40$ being a moderate effect, and $g = .60$ being a large effect. In cases where moderator analyses had only one comparison for a specific level of the potentially moderating variable, we dropped the comparison from the analysis due to concerns around statistical power.

Outliers and influential cases

We checked for outliers and influential cases for each 3LMA. We checked for outliers using van Lissa's (n.d.) method, which looks for comparisons that have confidence intervals that do not overlap with the overall meta-analytic effect size.

We then looked for influential cases by examining the Cook's distance, DFBETAS, and hat values (Viechtbauer, n.d.). A Cook's distance over .50 was more closely examined, as was a DFBETAS value over 1 (Viechtbauer & Cheung, 2010). Finally, hat values over $3 \times \left(number\ of\ model\ coefficients \div number\ of\ studies \right)$ were also closely examined (Viechtbauer, n.d.).

Data Visualization

We used Fernández-Castilla et al.'s (2020) approach to visualizing the dependent data within the 3LMAs. Specifically, the forest plots show the study precision (black lines), median precision of one effect size (gray lines), study weight (size of black box for each study), and the number of effect sizes derived from each study (J) (Fernández-Castilla et al., 2020).

Publication Bias

While 3LMA has existed for quite a while, there are relatively few approaches to computing publication bias in 3LMA, and existing methods used to assess

TABLE 1
Results of the five 3LMA

Outcome	$k_{comparisons}$	$k_{studies}$	g	p	Q	$Q\,p$	τ^2_{within}	$\tau^2_{between}$
Learning	70	25	.42	< .001	229.60	< .001	.08	.14
Motivation	47	17	.48	.001	396.16	< .001	.61	.01
Emotion	15	5	.60	.20	113.28	< .001	.56	.48
Perception	34	7	.05	.88	210.95	< .001	.28	.40
Cognitive load	5	3	−.09	.84	25.53	< .001	.43	.02

publication bias in conventional meta-analysis tend to lead to inflated type 1 error when used in 3LMA models (Rodgers & Pustejovsky, 2021). Consequently, it can be challenging to know to what extent specific types of publication bias may exist or to what extent they may influence the results. We used Fernández-Castilla et al.'s (2020) approach to creating funnel plots for dependent data meta-analyses and examined the plots for asymmetry. Given the context of this study, analyzing data from 30 primary studies with numerous outcomes and effect sizes, we feel the tradeoffs of including a more complete dataset through 3LMA outweigh the limitations of not being able to make many claims about specific types of publication bias.

Results

The overall results of each 3LMA are provided in Table 1. The results of all outlier and influence analyses, as well as funnel plots to assess publication bias, are available on OSF (https://osf.io/4duxj/?view_only=fe0e91572b354c8691f7f5a3c365abf8). In total, 30 studies were included in the analyses. Depending on the outcomes measured in each study, individual studies could be included in the analyses to answer more than one research question.

RQ1: What Is the Impact of VCs on K–12 Students' Learning, and What Moderates This Effect?

Our random-effects 3LMA of 70 comparisons from 25 studies produced a moderate, statistically significant effect indicating that VCs improved K–12 students' learning ($g = 0.42$, $p < .001$). The heterogeneity statistics indicated that there was significant heterogeneity, $Q(69) = 229.60$, $p < .001$, $\tau^2_{within =}.08$, $\tau^2_{between =}.14$. Overall, the model accounts for 74.65% of the variance (I^2), with 27.82% within studies and 46.83% between studies. A 3LMA forest plot is presented in Figure 2.

We next checked the data for outliers or studies with significant influence on the result. Using van Lissa's (n.d.) method, eight potential outliers were identified. However, examining the Cook's distance, DFBETAS, and hat values indicated that none were influential. As such, all eight potential outliers were retained in the analyses.

We also examined our data to see if publication bias was a significant concern with our sample. Examination of the funnel plots shows that they are reasonably symmetrical, indicating that publication bias is unlikely to be a significant concern.

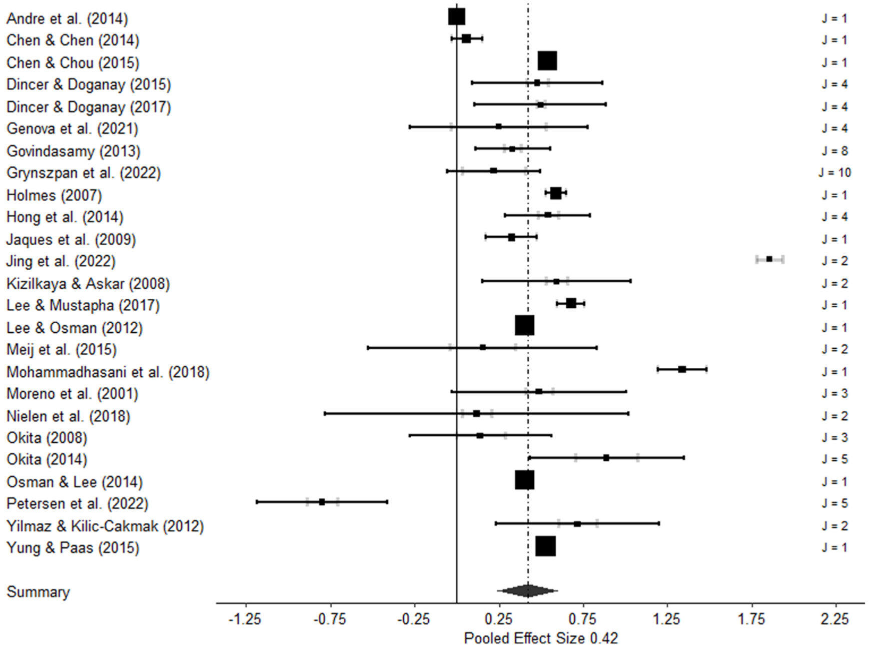

FIGURE 2. *A 3LMA forest plot of learning outcomes. The forest plot shows the study precision (black lines), median precision of one effect size (gray lines), study weight (size of black box for each study), and the number of effect sizes derived from each study (J) (Fernández-Castilla et al., 2020).*

Finally, we conducted a series of moderator analyses, as described below. All the results for the moderation analyses in regards to learning are available in Appendix A.

Publication Type

The type of publication did not significantly moderate the effects of VCs on K–12 students' learning $Q_b(2, 22) = .21$, $p = .812$.

Study Design

The control condition used in the study significantly moderated the effects of VCs on K–12 students' learning, $Q_b(2, 22) = 4.91$, $p = .017$. Specifically, VCs were significantly more effective when compared to video instruction ($g = 0.93$, $p < .001$) and software programs ($g = .47$, $p < .001$), and they were not significantly different from traditional teaching ($g = .09$, $p = .51$).

We also found that the location where the study was conducted significantly moderated the effects of VCs on K–12 students' learning, $Q_b(7, 15) = 9.51$, $p < .001$. Specifically, two comparisons from China were associated with the strongest effects ($g = 1.65$, $p < .001$). The weakest effect was found from five comparisons in Denmark ($g = -.80$, $p < .001$). However, 33 comparisons did not report where the study took place ($g = .37$, $p < .001$). The most comparisons that did state where the data were collected were from the United States, which reported 11 comparisons ($g = 0.54$, $p < .01$).

Meanwhile, the randomization strategy, the duration of the intervention, the duration of the study, the learning domain, the study setting, and the learners' grade level were not statistically significant moderators in regards to learning outcomes. We did not analyze the potential moderation by study design as all but one comparison used a between-subjects design.

System Design

The media in which the learners' interacted with the VCs significantly moderated the VCs' effects on learning outcomes, $Q_b(1, 21) = 11.75$, $p = .003$. Specifically, most studies took place on a computer (59 comparisons), and they were associated with a moderate effect size ($g = .51$, $p < .001$). VCs were not significantly more effective than the control conditions when they took place in virtual reality scenarios (nine comparisons).

The system's pacing and the number of characters in the system did not significantly moderate the effects of a VC on learning outcomes. We did not analyze the potential moderation by task type, as all but one comparison used an individual task.

Character Design

The role the character played within the learning environment significantly moderated its effects on learning, $Q_b(4, 19) = 5.48$, $p = .004$. Specifically, 33 comparisons used the character as an information source, which was associated with a moderate to large effect size ($g = .57$, $p < .001$). VCs acting as coaches or scaffolds (11 comparisons) were also effective for supporting learning ($g = .46$, $p = .01$). However, five comparisons used the VCs in demonstrating or modeling roles, and these were not effective for learning compared to noncharacter conditions ($g = -.80$, $p = .01$).

We did not find that the VCs' voice, whether it used gestures or not, what type it was considered (e.g., pedagogical, conversational, etc.), its form (i.e., human-inspired vs nonhuman), its gender, or its age significantly moderated the effects of VCs on learning for K–12 students.

Assessment Design

We found that the item type on the learning test significantly moderated the effects of VCs on K–12 students' learning, $Q_b(4, 64) = 3.21$, $p = .018$. Specifically, six comparisons used multiple types of items, and these were associated with large effects from the VC ($g = .90$, $p < .001$). Meanwhile, free-response tests were used in 10 comparisons and were associated with a moderate effect ($g = .49$, $p = .02$), and multiple-choice questions were used in 16 comparisons and were associated with small to moderate effects ($g = .32$, $p = .02$).

The presence of a pretest also significantly moderated the effects of VCs on learning, $Q_b(1, 68) = 5.70$, $p = .02$. Most comparisons ($k = 52$) included a pretest, and they were associated with moderate to large effects ($g = .52$, $p < .001$), while

18 comparisons did not include a pretest and no statistically significant effect was found ($g = .11$, $p = .46$).

The type of test used—whether it was a general learning test, retention or recall test, or transfer test—did not significantly moderate the effects of VCs on learning.

Study Quality

Nineteen comparisons included all five of our quality indicators and were associated with the strongest effect sizes ($g = .74$, $p < .001$). Interestingly, 27 comparisons contained three quality indicators, and they were associated with moderate effects ($g = .43$, $p = .01$). Meanwhile, 24 comparisons contained four quality indicators, and they were not associated with statistically significant effects ($g = .16$, $p = .17$).

Whether or not the studies used the same base learning materials in the control condition as the experimental condition, whether they used random or stratified assignment to conditions, and whether they reported reliability metrics for the learning measures did not significantly moderate the effects of VCs on learning. Notably, no comparison reported more than 20% attrition in the sample. We did not analyze the potential moderation by whether the outcome was understandable, as all but one comparison had understandable measures.

RQ2: What Is the Impact of VCs on K–12 Students' Motivation, and What Moderates This Effect?

The random-effects 3LMA of 47 comparisons from 17 studies produced a moderate, statistically significant effect size showing that VC aids K–12 students' motivation ($g = .48$, $p = .001$). We found that there was significant heterogeneity within the sample, $Q(46) = 396.16$, $p < .001$, $\tau^2_{\text{within}} = .61$, $\tau^2_{\text{between} =} .01$. The model accounts for 92.21% of the variance (I^2), with 91.42% of the variance explained by within-study heterogeneity and .79% between studies. A 3LMA forest plot is presented in Figure 3.

Next, we checked for outliers and studies that had a significant influence. Using van Lissa's (n.d.) method, we found nine potential outliers. However, examining the Cook's distance, DFBETAS, and hat values indicated that none were influential. As such, they were retained in the analyses.

To check for publication bias, we again examined funnel plots. Examination of the funnel plots shows that they are reasonably symmetrical, although there was some asymmetry with regard to comparisons with small standard errors. There seems to be a slight skew toward larger positive effect sizes; however, there is also a concentration of studies smaller than the meta-analytical effect size. Accordingly, we acknowledge that while publication bias is a possible concern in this analysis, it is likely not extreme.

Finally, we considered whether moderator analyses were warranted. Although the Q statistic indicates there was significant heterogeneity, the I^2 statistic indicates that this heterogeneity is within studies rather than between studies. As such, we did not conduct moderator analyses.

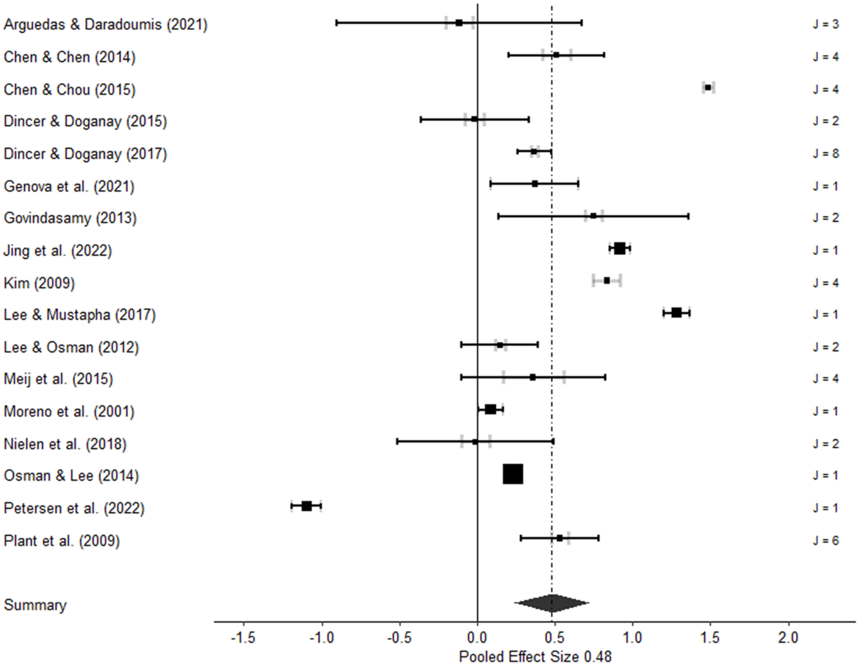

FIGURE 3. *A 3LMA forest plot of motivation outcomes. The forest plot shows the study precision (black lines), median precision of one effect size (gray lines), study weight (size of black box for each study), and the number of effect sizes derived from each study (J) (Fernández-Castilla et al., 2020).*

RQ3: What Is the Impact of VCs on K–12 Students' Emotions, and What Moderates This Effect?

The random effects 3LMA of emotional outcomes produced a large, nonsignificant effect ($g = .60$, $p = .20$). Regretfully, there was not much data for this analysis, with only 15 comparisons extracted from five studies. The heterogeneity statistics indicated there was significant heterogeneity within the sample, $Q(14) = 113.28$, $p < .001$, $\tau^2_{\text{within}=}.56$, $\tau^2_{\text{between}=}.48$. The model accounts for 91.97% of the variance (I^2), with 49.50% of the variance due to within-study heterogeneity, and 42.47% between studies. A 3LMA forest plot is presented in Figure 4.

When we searched for outliers, we located only one potential outlier. However, examining the Cook's distance, DFBETAS, and hat value indicated that it was not influential. It was therefore retained in the analysis.

Examination of the funnel plots showed that they were reasonably symmetrical, although it is difficult to address this with much confidence due to the small number of comparisons in the sample. Due to this, we suggest that publication bias may not be a significant concern but caution that this conclusion

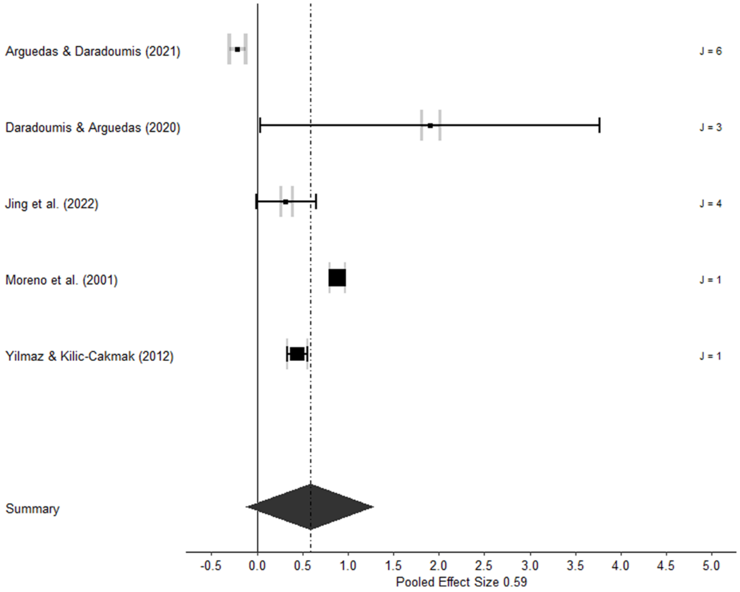

FIGURE 4. *A 3LMA forest plot of emotion outcomes. The forest plot shows the study precision (black lines), median precision of one effect size (gray lines), study weight (size of black box for each study), and the number of effect sizes derived from each study (J) (Fernández-Castilla et al., 2020).*

is hard to support with high confidence due to there only being five studies in the sample.

We next considered whether moderator analyses were warranted. Although the heterogeneity statistics rationalize the use of moderator analyses, due to the small number of comparisons and nonsignificant overall effect size, we did not run moderation analyses.

RQ4: What Is the Impact of VCs on K–12 Students' Perceptions, and What Moderates This Effect?

The random effects 3LMA of the impact of VCs on K–12 students' perceptions found a negligible, nonsignificant effect ($g = .05$, $p = .88$). Notably, this analysis contained 34 comparisons extracted from seven studies. Heterogeneity statistics indicated significant heterogeneity within the sample, $Q(33) = 210.95$, $p < .001$, $\tau^2_{\text{within}=}.28$, $\tau^2_{\text{between}=}.40$. The model accounts for 88.91% of the variance (I^2), with 37.16% of the heterogeneity within studies and 51.75% between studies. A 3LMA forest plot is presented in Figure 5.

Our check for outliers indicated three potential outliers. When examining the Cook's distance and DFBETAS, none were considered influential. Hat values indicated that two of the outliers were influential. However, since the other two tests indicated they were not, these comparisons were retained in the analyses.

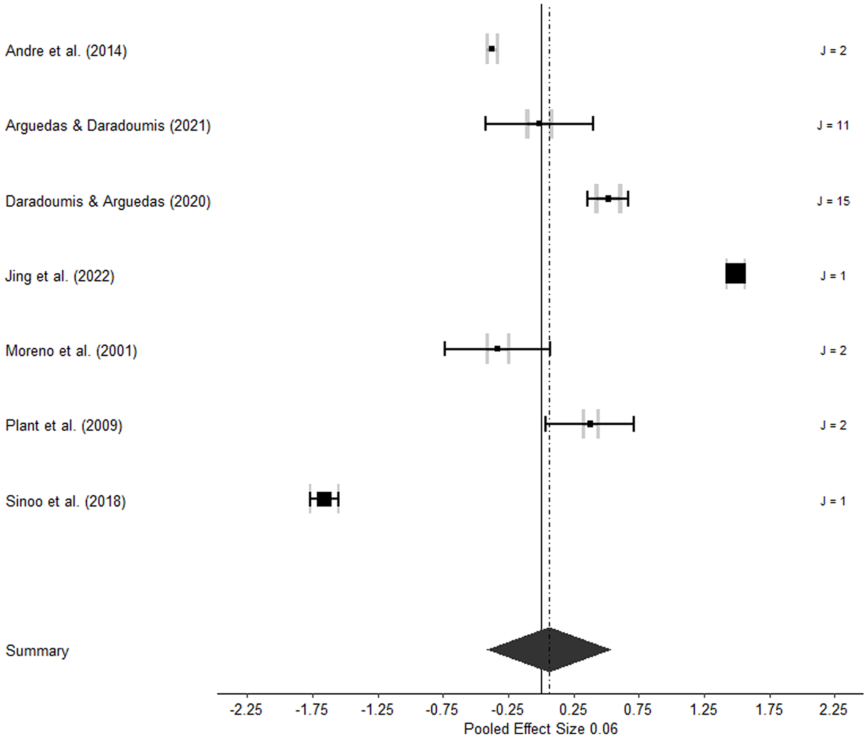

FIGURE 5. *A 3LMA forest plot of perception outcomes. The forest plot shows the study precision (black lines), median precision of one effect size (gray lines), study weight (size of black box for each study), and the number of effect sizes derived from each study (J) (Fernández-Castilla et al., 2020).*

The funnel plots show that the data are quite symmetrical. Accordingly, publication bias was not viewed as a significant concern with this analysis.

We next considered whether moderator analyses were warranted. The heterogeneity statistics indicated that they were, and the fact that there were 34 comparisons suggests there may be potential to find interesting moderators. However, these 34 comparisons were extracted from only seven studies. Furthermore, the vast majority (25 of 34) of comparisons came from only two studies, both of which had difficult-to-understand outcome measures. Based on this, we did not conduct moderator analyses because we felt they may lead to conclusions with substantial limitations.

RQ5: What Is the Impact of VCs on K–12 Students' Cognitive Load, and What Moderates This Effect?

The random effects 3LMA of the impact of VCs on cognitive load outcomes was quite small, with only five comparisons extracted from three studies. The

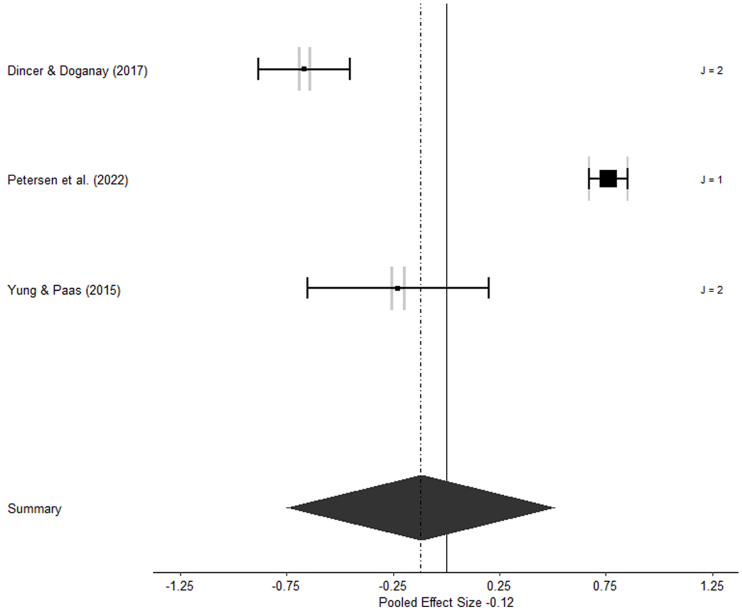

FIGURE 6. *A 3LMA forest plot of cognitive load outcomes. The forest plot shows the study precision (black lines), median precision of one effect size (gray lines), study weight (size of black box for each study), and the number of effect sizes derived from each study (J) (Fernández-Castilla et al., 2020).*

overall effect size was not significant, but recall that negative effect sizes for this analysis indicates a reduction of cognitive load ($g = -.09$, $p = .84$). Heterogeneity statistics indicated there was significant heterogeneity within the sample, $Q(4) = 25.53$, $p < .001$, $\tau^2_{\text{between}=}.43$, $\tau^2_{\text{within}=}.02$. The model accounts for 93.30% of the variance (I^2), with 4.21% within studies and 89.00% between studies. A 3LMA forest plot is presented in Figure 6.

Due to the very small number of comparisons and studies included in this analysis, we did not search for outliers, examine data for indicators of publication bias, or conduct moderation analyses.

Discussion

RQ1: What Is the Impact of VCs on K–12 Students' Learning, and What Moderates This Effect?

Previous conventional meta-analyses of the impacts of VCs on learning outcomes have produced small to moderate overall effect sizes ranging from $g = .19$ (Schroeder et al., 2013) to $g = .42$ (Peng & Wang, 2022), although effect sizes generally range from $g = .20$ to $g = .30$ (Castro-Alonso et al., 2021; Guo & Goh, 2015; Wang et al., 2023). When we examine existing meta-analytic results in more depth, we can see that these meta-analyses only examined a small number of studies of K–12 participants, and the effect sizes varied (Castro-Alonso et al., 2021; Schroeder et al., 2013; Wang et al., 2023). In our 3LMA specifically examining the impact of VC on K–12 students' learning, we found a statistically significant, moderate effect size ($g = .42$, $p < .001$). Importantly, this was derived from 70 comparisons extracted from 25 studies, well more than previous analyses had located and analyzed. This result shows that VCs can, generally speaking, improve K–12 student learning compared to non-VCs learning conditions.

One should keep in mind that while we ran a considerable number of moderator analyses, running many moderator analyses can increase the Type I error rate (false-positive rate). As a consequence, it is possible that the significantly moderating variables we found are, in fact, Type I errors and thus should be interpreted with caution. However, we feel one should interpret them with caution anyway, given that they were derived from 25 studies. With those caveats in mind, we consider pressing questions in the field.

One question that comes to mind is, are VCs more effective at promoting learning than traditional classroom teaching? Our results indicate that they are not statistically different from one another in the studies within our sample. However, under no circumstance should this be inferred to mean that VCs can or should replace human teachers. We unequivocally reject that notion, and we advise that VCs are simply one of many tools a teacher can use to help their students learn. This being the case, we turn our attention to online learning or computer-based learning, which has greatly expanded in the last 10 years, particularly during and after the COVID-19 pandemic. Two common ways of learning with computers are either video lectures or interactive software programs. We found that learning with a VC was significantly better than learning with videos that did not contain a VC ($g = .93$, $p < .001$), and learning with a VC was significantly better than learning with interactive software that did not contain a VC ($g = .47$, $p < .001$). We found no control condition that performed significantly better than a VC condition. In short, VCs were effective across a range of computer-based learning environments, regardless of what they were compared against.

While we found that the study's location significantly moderated the effects of a VC, this result is difficult to explain. One possible explanation for this finding is that cultural norms and expectations influence how students perceive and interact with a VC. A few studies have indicated that cultural references and social cues, such as the visual appearance and characteristics of the VC (Veletsianos, 2010); the facial expressions, gestures, and postures of the VC (Rehm et al., 2012); the conversational styles used by the VCs (Lin et al., 2020); and the language used by the VCs (López et al., 2021) can impact the quality of the interactions between students and VC. Thus, designing VCs that acknowledge and value the students' cultural backgrounds is critical to avoid miscommunication or misunderstandings when interacting with students. The development of culturally responsive VCs that can adapt to students' cultural norms is crucial given the diverse backgrounds of students (e.g., race, ethnicity, language background, country of origin) in today's K–12 classrooms. The use of culturally relevant VCs could create learning environments that are more inclusive and engaging for all students. Additional work would need to be done to investigate this specifically.

Another significantly moderating variable was in relation to the media in which the VC appeared. While VCs were significantly more effective than non-VC conditions when shown on a computer screen ($g = .51, p < .001$), these effects were not generalizable to virtual reality scenarios, which showed no significant difference. While there were only nine comparisons that used virtual reality, the effect size was negative. We question whether there would have been more comparisons in this category if the result had been statistically significant. If nothing else, this trend in the data shows the need for more studies using VCs in virtual reality environments. It is quite possible that the same results found on computer interfaces may not be found in virtual reality environments due to the affordances of each. For example, the computer-based environment may benefit from the social cues fostered by a VC (Atkinson et al., 2005; Mayer et al., 2003; Schneider et al., 2022). Is it possible that the increased level of immersion offered by virtual reality negates the social cues offered by the integration of a VC? An alternative explanation is that both studies that used virtual reality (Bøg Petersen et al., 2022; Genova et al., 2021) used traditional teaching as the control condition. As discussed, traditional teaching was not significantly different from learning with a VC. Regardless, the nature of the trend in relation to virtual reality, as well as the small number of studies, indicates that it may be beneficial to examine the use of VCs in virtual reality environments compared to virtual reality environments that don't contain VCs to examine the impacts on learning.

Another issue highlighted in previous reviews was that VCs do not typically play many different roles in the learning environment. A previous review found that nearly all VCs served as information sources, with fewer VCs providing coaching or scaffolding (Schroeder & Gotch, 2015). They also found that no VCs played the roles of testing, demonstrating, or modeling a task. This pattern was relatively consistent in our sample. We found most comparisons used the VC as an information source, and they were quite effective in this role ($g = .57, p < .001$). Fewer comparisons examined the VC in a coaching or scaffolding role ($n = 11$), but they were quite effective at this role as well ($g = .46, p = .01$). Perhaps the biggest surprise to our team was that five comparisons used VCs for demonstrating or modeling tasks, and these were associated with negative effects ($g = -.80$, $p = .01$). In other words, the control condition (in this case, traditional teaching) was significantly more effective, but all five comparisons came from the same study that used VCs in a virtual reality environment (Bøg Petersen et al., 2022). This raises the question of whether VCs that demonstrate or model tasks are generally less effective than other conditions, or whether this study is unique in its findings and context. More research is needed to investigate this phenomenon, particularly since demonstrating or modeling was one of the first uses for VCs in learning environments (Johnson & Rickel, 1997).

One finding in our results may cause one to question whether the results reflect a true overall effect or are a result of methodological choices in primary studies. We found that VCs were more effective than control conditions in comparisons that included a pretest ($g = .52, p < .001$), while they were not significantly better in comparisons that did not include pretests. These results make one question if there could be a pretest sensitization effect occurring (Cohen et al., 2007). However, this seems unlikely given that both the VC and control conditions included pretests. Moreover, our study quality metrics indicated that comparisons with all five quality indicators—that is, those that had low attrition, similar materials between conditions, randomized or stratified sampling strategies, used measures that were easily understandable to a reader, and reported reliability metrics—were associated with large effects compared to control conditions ($g = .74$, $p < .001$), while those with less quality indicators were associated with either no statistically significant advantage for the use of VCs compared to control conditions or a smaller significant effect. However, the rationale for why studies with three quality indicators resulted in a statistically significant effect, while those with four did not have a statistically significant effect, is difficult to determine. Examining the data, we can see that the majority of comparisons that found an advantage for the non-VC condition occurred in studies with four quality indicators, and five of them came from the same study, which also happens to be the study that compared a VC in virtual reality environments to a real human teacher (Bøg Petersen et al., 2022). Taken together, we can draw a few specific conclusions. First, we do not believe our result indicates a pretest sensitization effect in our studies. However, if one wanted to investigate it further, experiments with Solomon four-group designs could be warranted. Second, it is apparent that the study by Bøg Petersen et al. (2022) could contribute to why we found that studies with four quality indicators did not produce a statistically significant result compared to non-VC control conditions. We can safely conclude that studies with all five of our quality indicators produced the largest effect sizes; however, we are hesitant to make generalizable claims about why studies with fewer quality indicators may have shown smaller or nonsignificant effects, particularly given that our list of quality indicators could be viewed as incomplete depending on one's personal opinions about research quality.

Finally, it is important to acknowledge the number of moderating analyses that were not statistically significant. One could interpret this finding as VCs broadly supported K–12 students' learning, with few factors having any notable influence on this. However, this must be seen through the context of statistical power—while we had 70 comparisons, they came from 25 studies that measured learning. We may have lacked the statistical power necessary to find significant differences in moderators. When we parse our findings further, we can see that two of the moderating variables (media and agent role) were notably impacted by one study (Bøg Petersen et al., 2022). We do not highlight this as a flaw in the study or to criticize it. On the contrary, this study was rather unique in our sample, and it shows that more studies like it are necessary to build a deeper understanding of whether the key variables in the study (virtual reality and an agent that demonstrates or models tasks) are associated with effects consistent with those found by the authors.

To summarize, VCs broadly supported K–12 students' learning. There were almost no exceptions to this, as there were few situations in which statistically significant negative effects were found in moderator analyses. We also suggest that a key takeaway point made in earlier work, that the question of whether VCs are effective is not nuanced enough (Heidig & Clarebout, 2011), is now clearly supported. Our work shows that VCs support K–12 learning. Consistent with Heidig and Clarebout's (2011) conclusion, our findings show that we now need research synthesis to explore how different design features of VCs influence learning. Specifically, the meta-analysis reported here examined the impact of VC compared to no-VC conditions. We see the need for a research synthesis that examines the design features of VCs—that is, studies that compare a VC of one design to a VC of another design. This work will lead to insights into *how* VCs should be designed rather than *if* VCs should be integrated into learning environments. A systematic review of this work has recently been published (Zhang, Jaldi, Schroeder, López, et al., 2024); however, a meta-analysis may be appropriate in the future.

RQ2: What Is the Impact of VCs on K–12 Students' Motivation, and What Moderates This Effect?

Existing syntheses of the VC literature have shown mixed results on whether they can influence one's motivation to learn. Heidig and Clarebout's (2011) systematic review found only four studies that examined the impact of VC on learners' motivation, and three of them found no significant difference. Meanwhile, a more recent meta-analysis of 26 comparisons found that VCs can improve learners' intrinsic motivation (Wang et al., 2023). While these results were promising, they were not restricted to K–12 learners, so it was unclear what motivational effects may occur with this population specifically.

In our 3LMA of 47 comparisons, we found a statistically significant, moderate effect of VCs on learners' motivation ($g = .48$, $p = .001$). These results imply that K–12 learners find more motivational benefits from VCs than previous research has found with broader participant groups (e.g., Wang et al., 2023).

One consequential methodological decision we made was to combine all measures of motivation located in the primary studies. For example, our overall meta-analytical effect size is the amalgamation of measures of, for example, self-efficacy (Bøg Petersen et al., 2022; Genova et al., 2021; van der Meij et al., 2015), interest (Jing et al., 2022), utility (Plant et al., 2009), and motivation to continue using the software (Moreno et al., 2001). We suspect this may be why we found quite high within-study heterogeneity and very low between-study heterogeneity. For example, if a study were to measure intrinsic motivation and extrinsic motivation, we would not theoretically expect these to be consistent with one another, thus leading to within-study heterogeneity. However, it is important to note that the decision to combine all measures of motivation was a practical consideration. Although we had 47 comparisons from 17 studies when examining learners' motivation, the ability to compare across specific motivation constructs significantly drops the number of comparisons and studies we can draw from. A recently published meta-analysis (Gladstone et al., 2025) examined studies from a broader sample (i.e., not restricted to K–12 students) and was able to parse apart how specific motivation constructs (e.g., self-efficacy, interest, value, utility, etc.) are impacted by characteristics of the VC and of the students themselves (e.g., demographic characteristics).

RQ3: What Is the Impact of VCs on K–12 Students' Emotions, and What Moderates This Effect?

A recent meta-analysis found that VCs can have statistically significant, positive effects on learners' positive emotions (Wang et al., 2023). As such, we also examined the impact of VCs on K–12 students' emotions. However, we found no statistically significant effects ($g = .60$, $p = .20$). We acknowledge that, like motivation, this analysis contains many different types of variables, such as insecurity (Arguedas & Daradoumis, 2021), arousal (Daradoumis & Arguedas, 2020), and attitude (Yilmaz et al., 2018). Future research should examine the impact of VCs on specific positive and negative emotions (e.g., enjoyment and frustration).

It is also important to note, similar to what we note for perception measures below, that a substantial portion of the measures in this analysis (nine out of 15) came from two studies with difficult-to-understand outcome measures. This presents an issue in regards to interpretation. For example, if we had found a statistically significant effect, what would the result mean with 60% of the primary measures contributing to the effect size being difficult for readers to understand?

While we would not argue that this particular analysis is unimportant, we also do not believe the results are particularly meaningful due to the specific measures contributing data to the overall effect size. Not only were many difficult for our team to actually understand what was measured, but there was a wide variety of outcomes measured, with little consistency between studies. We encourage researchers to report their measurement techniques clearly and to ensure their measures are grounded in the literature and existing theory. This process can help not only develop more generalizable knowledge but also help create a consistent body of literature to summarize. At the present time, we do not feel there are enough consistent operationalizations of emotions within the VC field to make any particularly broad conclusions about the impacts of VCs on K–12 students' emotions, beyond finding no statistically significant effect. However, this lack of operationalization clarity in the VC literature highlights the importance and need for more interdisciplinary work, as there are widely accepted frameworks of achievement emotions in the educational psychology literature. One prominent framework in the educational psychology literature is Pekrun's control-value theory of achievement emotions (Pekrun, 2006; Pekrun et al., 2007). This framework provides a comprehensive approach for understanding and analyzing various emotions students may experience in learning contexts. This framework has also been used to develop the widely used Achievement Emotions Questionnaire (Pekrun et al., 2011). Thus, researchers interested in examining how student emotions may moderate the effectiveness of VCs on students' learning would benefit from using this prominent framework and well-validated measure of students' emotions.

RQ4: What Is the Impact of VCs on K–12 Students' Perceptions, and What Moderates This Effect?

Our analysis of the impact of VCs on the K–12 students' perceptions found no statistically significant overall effect. We think that there may be two reasons this is the case: the types of perceptions measured and the primary studies themselves.

First, we found many types of perceptions measured in the literature, ranging from the number of positive comments students provided (André et al., 2014) through gender stereotypes (Plant et al., 2009) and perceived experience (Jing et al., 2022). While it is beneficial that researchers are measuring so many different perceptions, it makes research synthesis like meta-analysis challenging. It is possible that the effect size found was not significant because of the inconsistent measures used in primary studies. Similarly, qualitative synthesis of these results is also likely challenging due to the disparate measures used between studies.

Second, it is important to acknowledge that 25 of the 34 comparisons of learners' perceptions came from only two studies (Arguedas & Daradoumis, 2021; Daradoumis & Arguedas, 2020), both of which had difficult-to-understand outcome measures. This presents two limitations. First, the vast majority of the data in this particular meta-analysis came from two studies, thereby limiting the generalizability of the analysis. In addition, since the measures were difficult to understand, they would be difficult to generalize even if a statistically significant overall effect were found.

We were surprised not to see any measures using the Agent Persona Instrument or Agent Persona Instrument – Revised (Ryu & Baylor, 2005; Schroeder et al., 2017, 2018). These instruments have been widely used with post-secondary populations to examine how learners perceive VCs (Davis et al., 2021). However, this may be due to the study's inclusion criteria. We searched for studies specifically examining the use of a VC in a non-VC learning experience. As such, the aforementioned perception instruments may not be as relevant as a study that examines the effects of different VC designs. In a systematic review on role models, which serve a similar role as VC, Gladstone and Cimpian (2021) found that the characteristics of the role model impact student motivation and learning. Therefore, given that VCs can aid K–12 students' learning, it seems pertinent to systematically review the design of VCs to examine the impacts on learning, perceptions, etc.

RQ5: What Is the Impact of VCs on K–12 Students' Cognitive Load, and What Moderates This Effect?

One critique of the use of VCs in educational technologies is the concern that they can be a source of extraneous cognitive processing or extraneous cognitive load (Clark & Choi, 2007). Interestingly, we located only three studies that measured cognitive load across five comparisons. Overall, our 3LMA showed that VCs had no significant impact on cognitive load outcomes for K–12 learners.

It is difficult to draw generalizable conclusions from the few studies included in our analysis. This is particularly the case given cognitive load measures in general, which have been criticized in the literature (De Jong, 2010; Schroeder & Cenkci, 2020). However, regardless of the strengths or weaknesses of these measures, at the present time, there is no discernible evidence for the claim that the mere presence of a VC causes extraneous cognitive load or increases cognitive load generally. With regards to future research, more rigorous methodological work needs to be done to develop a more reliable and valid measurement of cognitive load in this context. However, the results of this meta-analysis suggest that research efforts should instead be focused on other areas of the VC–child interaction, such as the impact of the VC on specific motivation constructs. Overall, unless there are substantial moves forward in the measurement of cognitive load, there may be other areas of VC–child interaction worth investigating that have stronger measures.

Implications for Theory

First, it is necessary to make it clear that this study was not designed to test whether certain theoretical perspectives better explain the effects of VCs compared to others. In other words, we did not set out to see if social agency theory (Mayer et al., 2003) or CASTLE (Schneider et al., 2022) better explains the effects of VCs on learning and learning related processes. However, our findings do have specific implications for both. Specifically, we found that VCs improve K–12 students' learning and motivation compared to conditions that did not contain a VC. This suggests, although we note it cannot be empirically shown because it was not directly measured, that the social cues provided by the VC influenced learning. This perspective is consistent with social agency theory (Mayer et al., 2003), although the theory is mechanistically vague. For example, critically examining the social agency theory perspective causes one to question *why* social cues influence learning. CASTLE suggests that social cues can influence the learners' social processes, which subsequently can influence learning (Schneider et al., 2022). This perspective seems more complete, yet we note that there are still notable empirical gaps in this theoretical explanation. For example, which social processes are VCs influencing? To what extent, and at which point in the cognitive process, are these social processes influencing learning? There is ample room for experimentation in answering these types of questions. In addition, it is necessary to explore the intersection of social processes in relation to agent design and implementation features. For example, based on extant research, we largely assume that it is the social processes initiated by VCs that improve learning outcomes if the effect is not attributed to specific, evidence-based pedagogical approaches, such as signaling the learners' attention to relevant parts of the screen. Yet to our knowledge, the field to date has not consistently measured how specific social processes may or may not be influenced by specific design or implementation features and the subsequent impacts on learning. In short, we have preliminary evidence that learners are invoking social processes that are influencing their learning based on these findings and extant theory in the area, but we lack concrete evidence that social processes are the reason why VCs improve learning, or that any specific design or implementation aspect of VCs influences this more or less than others.

Examining broader theoretical perspectives that highlight the importance of learners' emotions and perceptions in learning (Pintrich et al., 1993; Sinatra, 2005), our meta-analyses do not add specific extensions to these theories. Interestingly, we found that learners' emotions and perceptions were not significantly influenced by VCs. We suggest that there may be intersections between theoretical perspectives, namely CASTLE and conceptual change, that could open interesting lines of questioning. For example, if VCs were designed to specifically influence learners' emotions, would that subsequently be more effective for facilitating learning? Again, there is ample room for experimentation.

Finally, we sought to examine whether VCs significantly influenced learners' cognitive load. Although there were few comparisons, we found no significant effects from including VCs within the learning environment. This runs counter to some proposed negative impacts of VC (Clark & Choi, 2007) and is consistent with an experimental study with older learners (Schroeder, 2017). While researchers could continue the line of questioning of whether VCs influence learners' cognitive load, we do not believe this is the most fruitful path forward for the field for reasons explained previously, where we discussed the results of this meta-analysis.

Implications for Practice

The results of our meta-analyses tell a consistent story: VCs can improve K–12 students' learning and motivation in many situations, but they may not be more effective than traditional teaching approaches. Meanwhile, they did not significantly impact learners' emotions, perceptions, or cognitive load. Consequently, we suggest that adding VCs to computer-based trainings, videos, or other types of software may be effective, but one should not expect them to fully replace traditional teaching approaches, and we would not support this as an implication of this study.

When it comes to the VC design and behavior, our study has limited implications, but we can refer to the findings of a complementary systematic review that recently investigated this question. The authors reported that there were not consistent findings in relation to the VCs appearance influencing learning outcomes; however, learners did express preferences for certain VCs in studies (Zhang, Jaldi, Schroeder, López, et al., 2024). They concluded that there may be affective benefits to letting learners choose the VC they learn with, and based on the evidence available, it is unlikely to have any negative detriment on learning outcomes. Regarding the VCs' behavior or role in the environment, Zhang, Jaldi, Schroeder, López, et al. (2024) found that VCs were most effective when they embodied the effective pedagogies that teachers use in the classroom.

To summarize, VCs are best viewed as an effective teaching tool and should be used appropriately; they are not a panacea to all educational challenges, and we do not suggest that they should replace classroom teachers. Rather, teachers can use VCs to help improve their out-of-classroom teaching, such as computer-based homework, lecture videos viewed at home, intelligent tutoring systems, etc. These use cases for VCs, videos, and computer-based environments showed significant improvements in learning when VCs were included as opposed to when they were excluded. Evidence to date suggests that teachers should focus their design efforts on the pedagogy the VC uses during instruction, ensuring it aligns with best practices (Zhang, Jaldi, Schroeder, López, et al., 2024).

Limitations

Like all meta-analyses and systematic reviews, our work reported here is limited by the reporting in primary studies. For example, we had wanted to examine if VCs' gaze, gestures, and facial expressions influenced the variables examined. We coded this data, and it is reported in our coding forms (available in the OSF repository: https://osf.io/4duxj/?view_only=fe0e91572b354c8691f7f5a3c365abf8); however, few studies explicitly addressed these key design features, so we did not analyze the results. We strongly encourage researchers to report their VC's design in alignment with Heidig and Clarebout's (2011) framework so readers have a clear understanding of how the VC was designed.

Second, as noted in the respective analysis sections, we combined numerous measures of motivation, emotions, and perceptions that may not have been theoretically consistent with one another in order to calculate an overall effect size. This was done due to the limited number of consistent measures between studies, but it certainly could influence our results, particularly in relation to the emotion and perception analyses. For example, in the emotion analysis, measures of optimism, dominance, and calmness were statistically combined. While we positioned the positive effect size as a generally positive emotion, these emotions are certainly not the same thing, hence making the interpretation of the overall effect size complex. To provide more clarity in future synthesis efforts, more studies around each specific emotion would be needed.

We found analyzing study quality to be a particular challenge. We found no examples in the literature for what can or should be considered a good quality metric (as opposed to, say, a risk of bias assessment) for our use case. Accordingly, we considered the fundamental question: What are the bare minimum requirements we want to see in an intervention study when we consider whether we might be interested in using the intervention ourselves? Our quality indicators reflect these discussions. It is very important to note that we do not consider studies with all five quality indicators to, by default, be what one may term a "high-quality study," yet we also do not by default assume it is not. Rather, we specify it has quality *indicators*, not that it is or isn't a "quality" study. This process made clear to our team that the field would benefit from a theoretically guided and validated measure for evaluating primary study quality in educational intervention studies.

Finally, we acknowledge that the search terms used in Zhang, Jaldi, Schroeder, and Gladstone's (2024) scoping review were focused primarily on learning and motivational outcomes. This focus reflected our research aim: to synthesize studies focused on educational and motivation-related contexts among K–12 learners. While we also examined other variables (e.g., emotions and cognitive load), this was only done when they were measured in educational or motivational contexts, as these were considered secondary outcomes rather than central to the search strategy. As a result, we did not include these constructs as standalone search terms.

Implications for Future Research

We have outlined implications for future research in the previous sections. However, we have a few additional points we wish to raise as a result of this large-scale research synthesis.

First, we encourage researchers to ensure that they are fully reporting the results of their studies. In particular, the full reporting of descriptive statistics (at a minimum, mean, standard deviation, and sample size) makes it much easier for research synthesists to quantitatively summarize the literature. In addition, providing the data itself through repositories is also quite helpful for research synthesists.

Second, as noted previously, we agree with Heidig and Clarebout (2011) that we must examine the nuance involved with implementing VCs rather than simply whether they are effective or not. Consequently, we encourage researchers to provide more details about the VCs they are employing in their studies. While reviewing the literature, we found several studies (within and outside of our sample) that do not provide detailed descriptions of key aspects of VC design. We encourage researchers employing VCs to consider using Heidig and Clarebout's (2011) framework as a common grounding for describing their VC design. While this may add length to publications, we kindly remind researchers that they could include this type of information as supplementary materials or even provide it through a repository like the Open Science Framework (https://osf.io/). Doing so will allow for key insights into what aspects of VCs have the most influence on various outcomes. This becomes increasingly important as researchers begin to explore social cues and the intersection of social cues and cultural norms and expectations. Without detailed descriptions of the VC design, readers may struggle to build on findings.

Given the current challenges in creating VCs, particularly the time-consuming process and the technical skills required, it is imperative to develop new toolkits that simplify the creation and testing of VCs. While some toolkits exist, those we are aware of still require technical skills and considerable time investment (e.g., Hartholt et al., 2013). There is a need for toolkits that make the process more accessible for non-programmers, such as educators and researchers with limited programming expertise. One potential direction for future research is the creation of a user-friendly VC toolkit that aligns with Heidig and Clarebout's (2011) framework for designing VC. Such a toolkit could provide templates and customizable options based on established best practices in VC design. This approach would facilitate the consistent reporting and implementation of key design features, helping to enhance the comparability and generalizability of research findings. By lowering the barrier to entry for creating and deploying VCs, we can encourage more widespread experimentation and innovation in this field. This, in turn, could lead to more robust and nuanced understandings of the impacts of VCs on learners, particularly in the K–12 context.

Conclusion

As the use of VCs continues to increase (Zhang, Jaldi, Schroeder, & Gladstone, 2024), the results of this study provide an important overview and launching point for those looking to incorporate VCs in their teaching to help promote learning. We found that VCs can be a meaningful part of the learning environment, supporting K–12 students' learning and motivation, while not significantly impacting their emotions, perceptions, or cognitive load. While more research is needed in some areas, at this point it is quite clear that VCs can aid learning. We suggest that promising next steps include more experimentation to explore specifically what social processes VCs influence, as well as research synthesis around the impacts of VC design.

Appendix A

TABLE A1

Moderator analyses for learning outcomes

Publication Type								
Publication Type	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Conference proceeding	38	38	2	.52	.23	.42	[−.35, 1.39]	
Dissertation	36	66	3	.14	.78	.47	[−.85, 1.12]	
Journal article	1922	1886	65	.43*	<.001	.1	[.22, .64]	
								$Q_b(2, 22)=.21, p=.812$
Study Design								
Control condition	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Software program	1171	1174	47	.47*	<.001	.10	[.26, .67]	
Traditional teaching	635	616	19	.09	.51	.14	[−.19, .38]	
Video	190	200	4	.93*	<.001	.26	[.39, 1.47]	
								$Q_b(2, 22)=4.91, p=.017$
Randomization	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Group	723	703	13	.48*	.01	.18	[.11, .85]	
Individual	818	823	44	.41*	.01	.13	[.13, .68]	
Not reported	455	464	13	.35	.2	.27	[−.20, .91]	
								$Q_b(2, 22)=.10, p=.907$
Intervention duration	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
1 hour or less	371	369	16	.57*	<.001	.17	[.24, .91]	
>1 hour to 2 hours	198	214	9	−.06	.82	.24	[−.54, .43]	
>2 hours to 5 hours	455	470	11	.35	.14	.22	[−.12, .81]	
>5 hours to 10 hours	699	682	26	.37*	.04	.17	[.02, .72]	
>20 hours	124	128	4	.54	.17	.38	[−.26, 1.35]	
Not reported	149	127	4	.70*	.02	.26	[.15, 1.25]	
								$Q_b(5, 19)=1.28, p=.312$

(continued)

TABLE A1 (continued)

Study Design								
Duration of study	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
<1 day	361	376	15	.48*	.01	.17	[.13, .83]	
1 day to 1 week	580	586	19	.25	.21	.2	[−.14, .64]	
>1 week to 4 weeks	573	556	10	.53*	.03	.23	[.05, 1.01]	
>4 weeks	358	370	23	.32	.1	.19	[−.07, .72]	
Not reported	124	102	3	.71*	.04	.32	[.05, 1.37]	
								$Q_b(4, 20) = .57, p = .685$
Domain	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Computer skills	488	460	8	.49	.17	.34	[−.23, 1.20]	
Job skills	28	28	4	.25	.65	.54	[−.89, 1.38]	
Language	147	148	5	.36	.36	.38	[−.44, 1.16]	
Mathematics	166	196	10	.54	.07	.28	[−.04, 1.12]	
Reading	49	48	2	.08	.89	.53	[−1.04, 1.20]	
Science	1018	1000	31	.45*	.01	.15	[.14, .76]	
Social skills	100	110	10	.22	.65	.47	[−.78, 1.22]	
								$Q_b(6, 18) = .17, p = .983$
Setting	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Classroom	1600	1561	47	.34*	.01	.11	[.11, .58]	
Lab	217	241	11	.69*	.01	.23	[.21, 1.18]	
Not reported	179	188	12	.51	.09	.28	[−.08, 1.09]	
								$Q_b(2, 22) = .96, p = .398$
Grade level	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Primary	1162	1142	30	.50*	<.001	.13	[.24, .76]	
Primary/lower secondary	44	44	4	1.14*	<.001	.38	[.37, 1.91]	
Lower secondary	396	403	14	.48*	.01	.18	[.11, .84]	
Mixed lower-upper secondary	100	110	10	.22	.57	.38	[−.57, 1.01]	
Upper secondary	294	291	12	.08	.72	.21	[−.35, .51]	
								$Q_b(4, 20) = 1.69, p = .192$

(continued)

TABLE A1 (continued)

Study Design								
Location	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
China	66	60	2	1.65*	<.001	.29	[1.04, 2.26]	
Denmark	113	112	5	−.80*	<.001	.18	[−1.19, −.41]	
Malaysia	89	88	2	.51	.06	.25	[−.02, 1.04]	
Netherlands	70	68	4	.10	.66	.22	[−.37, .58]	
Taiwan	271	288	7	.49*	.01	.13	[.20, .77]	
Turkey	103	92	4	.65*	.01	.20	[.22, 1.07]	
United States	172	202	11	.54*	<.01	.14	[.25, .83]	
Not reported	1046	1014	33	.37*	<.001	.07	[.23, .52]	
								$Q_b(7, 15)=9.51, p<.001$
Levels with only one comparison removed from the analysis: France, Iran								
System Design								
Pacing	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Learner-paced	1155	1167	45	.32*	.01	.12	[.08, .56]	
System-paced	166	170	12	.83*	.01	.31	[.19, 1.47]	
Unclear	675	653	13	.54*	.01	.19	[.14, .93]	
								$Q_b(2, 22)=1.45, p=.256$
Media	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Computer	1740	1736	59	.51*	<.001	.08	[.34, .69]	
Virtual reality	141	140	9	−.39	.13	.25	[−.91, .13]	
								$Q_b(1, 21)=11.75, p=.003$
Number of agents	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
One	1727	1728	55	.42*	<.001	.11	[.20, .64]	
Two	100	110	10	.22	.61	.43	[−.67, 1.11]	
Not reported	169	152	5	.56	.12	.34	[−.15, 1.27]	
								$Q_b(2, 22)=.19, p=.825$

(continued)

TABLE A1 (continued)

Character Design								
Voice type	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Human voice	421	428	13	−.20	.47	.27	[−.76, .37]	
TTS	21	20	2	.15	.76	.51	[−.91, 1.22]	
Voice (not specified)	557	560	21	.46*	.01	.17	[.12, .79]	
Text communication	287	294	13	.53*	.01	.19	[.13, .93]	
Student choice between narration and text	246	238	4	.91*	<.001	.24	[.44, 1.38]	
Voice and text communication	198	204	6	.33	.18	.23	[−.16, .82]	
Not reported	266	246	11	.57*	<.001	.19	[.19, .95]	
								$Q_b(6, 18)=1.77, p=.162$
Agent gestures	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Yes	806	807	28	.36*	.03	.15	[.04, .68]	
No	81	91	3	.36	.35	.37	[−.42, 1.13]	
Not reported	1109	1092	39	.47*	<.001	.13	[.2, .75]	
								$Q_b(2, 22)=.16, p=.849$
Agent type	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Pedagogical	1695	1728	56	.45*	<.001	.11	[.21, .69]	
Motivational	139	124	6	.29	.33	.29	[−.31, .89]	
Multiple roles	88	67	6	.64	.06	.32	[−.02, 1.30]	
Not reported	74	71	2	.03	.95	.39	[−−.78, .83]	
								$Q_b(3, 21)=.59, p=.625$
Agent form	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Human-inspired	1143	1142	40	.28*	.02	.12	[.04, .51]	
Mixed—multiple agents	440	426	10	.57*	.01	.21	[.15, .99]	
Nonhuman	413	422	20	.60*	<.001	.16	[.26, .93]	
								$Q_b(2, 22)=1.81, p=.188$

(continued)

TABLE A1 (continued)

Character Design

Agent gender	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Female	506	512	17	−.05	.79	.19	[−.45, .35]	
Gender neutral	28	28	4	.25	.58	.44	[−.68, 1.17]	
Male	164	180	4	.71*	.01	.25	[.19, 1.24]	
Multiple agents	713	686	23	.47*	<.001	.15	[.16, .77]	
Nonhuman	413	422	20	.60*	<.001	.15	[.28, .92]	
Not reported	172	162	2	.18	.57	.31	[−.45, .8]	
								$Q_b(5, 19) = 1.96$, $p = .131$
Agent age	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Adult	755	755	31	.24	.13	.15	[−.08, .55]	
Irrelevant (nonhuman agent)	413	422	20	.60*	<.01	.16	[.26, .93]	
Multiple agents and/or multiple ages	568	552	12	.53*	.01	.19	[.16, .91]	
Unknown	260	261	7	.29	.22	.24	[−.18, .77]	
								$Q_b(3, 21) = 1.19$, $p = .338$
Agent role	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Coaching/scaffolding	211	219	11	.46*	.01	.17	[.11, .81]	
Demonstrating/ modeling	113	112	5	−.80*	.01	.28	[−1.38, −.21]	
Information source	1216	1200	33	.57*	<.001	.1	[.37, .77]	
Multiple roles	273	286	16	.30	.09	.17	[−.05, .66]	
Not reported	132	122	4	.48	.11	.28	[−.11, 1.07]	
								$Q_b(4, 19) = 5.48$, $p = .004$

(continued)

TABLE A1 (continued)

Assessment Design								
Item type	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Free response	155	186	10	.49*	.02	.2	[.09, .89]	
Multiple choice	457	465	16	.32*	.02	.13	[.06, .59]	
Multiple item types	223	195	6	.90*	<.001	.21	[.48, 1.31]	
Performance	430	436	17	.03	.83	.15	[−.27, .33]	
Not reported	709	685	20	.51*	<.001	.14	[.24, .79]	
								$Q_b(4, 64) = 3.21, p = .018$
Test type	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
General learning test	514	490	14	.45*	.01	.15	[.13, .77]	
Retention/recall	787	784	23	.33*	.03	.15	[.04, .63]	
Transfer	695	716	33	.45*	<.001	.14	[.17, .73]	
								$Q_b(2, 67) = .38, p = .687$
Pretest present	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
No	600	597	18	.11	.46	.15	[−.19, .42]	
Yes	1396	1393	52	.52*	<.001	.1	[.32, .72]	
								$Q_b(1, 68) = 5.70, p = .02$
Study Quality								
Control same materials	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
No	619	606	25	.21	.17	.15	[−.09, .5]	
Yes	1377	1384	45	.51*	<.001	.11	[.30, .73]	
								$Q_b(1, 68) = 3.58, p = .063$
Randomized or stratified	n_{agent}	$n_{control}$	k_{comp}	*g*	*p*	SE	95% CI	
Not reported	455	464	13	.35	.19	.26	[−.19, .89]	
Yes	1541	1526	57	.43*	<.001	.1	[.22, .65]	
								$Q_b(1, 23) = .08, p = .779$

(continued)

TABLE A1 (continued)

Study Quality

Reliability reported	n_{agent}	$n_{control}$	k_{comp}	g	p	SE	95% CI	
No	960	998	39	.25	.06	.13	[−.01, .52]	
Yes	1036	992	31	.57*	<.001	.12	[.31, .83]	
								$Q_b(1, 23) = 3.13, p = .09$
Study quality	n_{agent}	$n_{control}$	k_{comp}	g	p	SE	95% CI	
3	642	661	27	.43*	.01	.15	[.11, .74]	
4	765	761	24	.16	.17	.11	[−.07, .39]	
5	589	568	19	.74*	<.001	.13	[.48, 1.00]	
								$Q_b(2, 67) = 7.09, p = .002$

*Indicates $p < .05$.

The n_{agent} indicates the number of participants in the virtual character group. Note that individual participants could be counted more than once, depending on the number of comparisons and outcome variables from within the same study. For example, a study with 30 participants in the experimental group that contains two relevant outcome variables would be counted as two comparisons, and thus as 60 participants rather than 30 participants.

The $n_{control}$ indicates the number of participants in the control group. Note that individual participants could be counted more than once, depending on the number of comparisons and outcome variables from within the same study. For example, a study with 30 participants in the experimental group that contains two relevant outcome variables would be counted as two comparisons, and thus as 60 participants rather than 30.

The k_{comp} indicates the number of comparisons.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant DRL-2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

ORCID iDs

Noah L. Schroeder https://orcid.org/0000-0002-3281-2594
Jessica R. Gladstone https://orcid.org/0000-0001-6030-6288

Notes

[1] In this study, we use the term "gender" to refer to the perceived social characteristics of pedagogical agents. However, prior research and coding schemes often categorize agents as "male" or "female," which aligns more with sex-based labels. We acknowledge this discrepancy and encourage future work to explore broader, more inclusive gender representations in virtual characters.

[2] We defined studies that were not easily understood as those that had two or more people from our research team not clearly understand what was actually being measured in the study. There were few instances of this throughout the coding process. Examples are mentioned in the discussion section, where they become more contextually relevant to the results found.

References

*Indicates study included in the meta-analyses

*André, V., Jost, C., Hausberger, M., Le Pévédic, B., Jubin, R., Duhaut, D., & Lemasson, A. (2014). Ethorobotics applied to human behaviour: Can animated objects influence children's behaviour in cognitive tasks? *Animal Behaviour*, *96*, 69–77. https://doi.org/10.1016/j.anbehav.2014.07.020

*Arguedas, M., & Daradoumis, T. (2021). Analysing the role of a pedagogical agent in psychological and cognitive preparatory activities. *Journal of Computer Assisted Learning*, *37*(4), 1167–1180. https://doi.org/10.1111/jcal.12556

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, *30*(1), 117–139. https://doi.org/10.1016/j.cedpsych.2004.07.001

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, *15*(2), 95–115. https://doi.org/10.3233/irg-2005-15(2)02

Beege, M., & Schneider, S. (2023). Emotional design of pedagogical agents: The influence of enthusiasm and model-observer similarity. *Educational Technology Research and Development*, *71*(3), 859–880. https://doi.org/10.1007/s11423-023-10213-4

*Bøg Petersen, G., Klingenberg, S., & Makransky, G. (2022). Pipetting in virtual reality can predict real-life pipetting performance. *Technology, Mind, and Behavior*, *3*(3). https://doi.org/10.1037/tmb0000076

Castro-Alonso, J. C., Wong, R. M., Adesope, O. O., & Paas, F. (2021). Effectiveness of multimedia pedagogical agents predicted by diverse theories: A meta-analysis. *Educational Psychology Review*, *33*(3), 989–1015. https://doi.org/10.1007/s10648-020-09587-1

*Chen, C., & Chou, M. (2015). Enhancing middle school students' scientific learning and motivation through agent-based learning. *Journal of Computer Assisted Learning*, *31*(5), 481–492. https://doi.org/10.1111/jcal.12094

*Chen, Z.-H., & Chen, S. Y. (2014). When educational agents meet surrogate competition: Impacts of competitive educational agents on students' motivation and performance. *Computers & Education*, *75*, 274–281. https://doi.org/10.1016/j.compedu.2014.02.014

Chiou, E. K., Schroeder, N. L., & Craig, S. D. (2020). How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education*, *146*, 103756. https://doi.org/10.1016/j.compedu.2019.103756

Clarebout, G., Elen, J., Johnson, W. L., & Shaw, E. (2002). Animated pedagogical agents: An opportunity to be grasped? *Journal of Educational Multimedia and Hypermedia*, *11*(3), 267–286.

Clark, R. E., & Choi, S. (2007). The questionable benefits of pedagogical agents: Response to veletsianos. *Journal of Educational Computing Research*, *36*(4), 379–381. https://doi.org/10.2190/2781-3471-67MG-5033

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/s0140525x01003922

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, *114*, 193–205. https://doi.org/10.1016/j.compedu.2017.07.003

*Daradoumis, T., & Arguedas, M. (2020). Cultivating students' reflective learning in metacognitive activities through an affective pedagogical agent. *Journal of Educational Technology & Society*, *23*(2), 19–31.

Davis, R. O., Park, T., & Vincent, J. (2021). A systematic narrative review of agent persona on learning outcomes and design variables to enhance personification. *Journal of Research on Technology in Education*, *53*(1), 89–106. https://doi.org/10.1080/15391523.2020.1830894

Davis, R. O., Park, T., & Vincent, J. (2023). A meta-analytic review on embodied pedagogical agent design and testing formats. *Journal of Educational Computing Research*, *61*(1), 30–67. https://doi.org/10.1177/07356331221100556

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, *38*(2), 105–134. https://doi.org/10.1007/s11251-009-9110-0

Deepbrain AI. (2023). *Best ai video generator | deepbrain ai*. Best Ai Video Generator | Deepbrain Ai. https://www.deepbrain.io/

*Dincer, S., & Doganay, A. (2015). The impact of pedagogical agent on learners' motivation and academic success. *Practice and Theory in Systems of Education*, *10*(4), 329–348. https://doi.org/10.1515/ptse-2015-0032

*Dinçer, S., & Doğanay, A. (2017). The effects of multiple-pedagogical agents on learners' academic success, motivation, and cognitive load. *Computers & Education*, *111*, 74–100. https://doi.org/10.1016/j.compedu.2017.04.005

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). Visual representations of meta-analyses of multiple outcomes: Extensions to forest plots, funnel plots, and caterpillar plots. *Methodology*, *16*(4), 299–315. https://doi.org/10.5964/meth.4013

*Genova, H. M., Lancaster, K., Morecraft, J., Haas, M., Edwards, A., DiBenedetto, M., Krch, D., DeLuca, J., & Smith, M. J. (2021). A pilot RCT of virtual reality job interview training in transition-age youth on the autism spectrum. *Research In Autism Spectrum Disorders*, *89*, 101878. https://doi.org/10.1016/j.rasd.2021.101878

Gladstone, J. R., & Cimpian, A. (2021). Which role models are effective for which students? A systematic review and four recommendations for maximizing the effectiveness of role models in STEM. *International Journal of STEM Education*, *8*, 1–20. https://doi.org/10.1186/s40594-021-00315-x

Gladstone, J. R., Schroeder, N. L., Heidig, S., Zhang, S., Palaguachi, C., & Pitera, M. (2025). Do pedagogical agents enhance student motivation? Unraveling the evidence through meta-analysis. *Educational Psychology Review*, *37*(3), 72. https://doi.org/10.1007/s10648-025-10050-2

*Govindasamy, M. K. (2013). Embodied agent in tutor role: Effects on field dependent and independent low achiever's retention and perceived science self-efficacy beliefs. *Journal of Educational Multimedia and Hypermedia*, *22*(3), 273–297.

*Grynszpan, O., Bouteiller, J., Grynszpan, S., Martin, J.-C., & Nadel, J. (2022). Social gaze training for Autism Spectrum Disorder using eye-tracking and virtual humans. *Interaction Studies*, *23*(1), 89–115. https://doi.org/10.1075/is.21022.gry

Guo, Y. R., & Goh, D. H.-L. (2015). Affect in embodied pedagogical agents: Meta-analytic review. *Journal of Educational Computing Research*, *53*(1), 124–149. https://doi.org/10.1177/0735633115588774

Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., & Gratch, J. (2013). All together now. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *Intelligent virtual agents. IVA 2013. Lecture notes in computer science* (*Vol. 8108*, pp. 368–381). Springer.

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 79–91. https://doi.org/10.1037/stl0000021

Heidig, S., & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review*, *6*(1), 27–54. https://doi.org/10.1016/j.edurev.2010.07.004

*Holmes, J. (2007). Designing agents to support learning by explaining. *Computers & Education*, *48*(4), 523–547. https://doi.org/10.1016/j.compedu.2005.02.007

*Hong, Z.-W., Chen, Y.-L., & Lan, C.-H. (2014). A courseware to script animated pedagogical agents in instructional material for elementary students in English education. *Computer Assisted Language Learning*, *27*(5), 379–394. https://doi.org/10.1080/09588221.2012.733712

*Jaques, P. A., Lehmann, M., & Pesty, S. (2009). *Evaluating the affective tactics of an emotional pedagogical agent* [Conference session]. Proceedings of the 2009 ACM Symposium on Applied Computing, 104–109. Association for Computing Machinery. https://doi.org/10.1145/1529282.1529304

*Jing, B., Liu, J., Gong, X., Zhang, Y., Wang, H., & Wu, C. (2022). Pedagogical agents in learning videos: Which one is best for children? *Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2022.2141787

Johnson, A. M., Ozogul, G., Moreno, R., & Reisslein, M. (2013). Pedagogical agent signaling of multiple visual engineering representations: The case of the young female agent. *Journal of Engineering Education*, *102*(2), 319–337. https://doi.org/10.1002/jee.20009

Johnson, W. L., & Rickel, J. (1997). Steve: An animated pedagogical agent for procedural training in virtual environments. *ACM SIGART Bulletin*, *8*(1–4), 16–21. https://doi.org/10.1145/272874.272877

*Kim, Y. (2009). The role of learner attributes and affect determining the impact of agent presence. *International Journal of Learning Technology*, *4*(3/4), 2–2. https://doi.org/10.1504/IJLT.2009.028808

*Kizilkaya, G., & Askar, P. (2008). The effect of an embedded pedagogical agent on the students' science achievement. *Interactive Technology and Smart Education*, *5*(4), 208–216. https://doi.org/10.1108/17415650810930893

*Lee, T. T., & Mustapha, N. H. (2017). Who is more efficient: Teacher or pedagogical agents? In M. Puteh, N. AbdHamid, & N. Adenan (Eds.), *Proceedings of the International Conference on Education, Mathematics, and Science 2016* (*Vol. 1847*). AIP Publishing. https://doi.org/10.1063/1.4983903

*Lee, T. T., & Osman, K. (2012). Interactive multimedia module in the learning of electrochemistry: Effects on students' understanding and motivation. In G. Baskan, F. Ozdamli, S. Kanbul, & D. Ozcan (Eds.), *4th World Conference on Educational Sciences* (Vol. 46, pp. 1323–1327). Elsevier Ltd. https://doi.org/10.1016/j.sbspro.2012.05.295

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). *The persona effect: Affective impact of animated pedagogical agents* [Conference session]. Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, 359–366, Springer-Verlag, Berlin, Heidelberg.

Lin, L., Ginns, P., Wang, T., & Zhang, P. (2020). Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education*, *143*, 103658. https://doi.org/10.1016/j.compedu.2019.103658

López, A. A., Guzman-Orth, D., Zapata-Rivera, D., Forsyth, C. M., & Luce, C. (2021). Examining the accuracy of a conversation-based assessment in interpreting English learners' written responses. *ETS Research Report Series*, *2021*(1), 1–15. https://doi.org/10.1002/ets2.12315

Mayer, R. E. (2014a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press.

Mayer, R. E. (2014b). Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 345–368). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.017

Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, *36*(1), 8. https://doi.org/10.1007/s10648-023-09842-1

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, *95*(2), 419–425. https://doi.org/10.1037/0022-0663.95.2.419

*Mohammadhasani, N., Fardanesh, H., Hatami, J., Mozayani, N., & Fabio, R. A. (2018). The pedagogical agent enhances mathematics learning in ADHD students. *Education & Information Technologies*, *23*(6), 2299–2308. https://doi.org/10.1007/s10639-018-9710-x

*Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177–213. https://doi.org/10.1207/S1532690XCI1902_02

National Center for Education Statistics. (n.d.). *Education indicators: An international perspective / indicator 1 side bar*. ISCED Levels of Education. https://nces.ed.gov/pubs/eiip/eiip1s01.asp

*Nielen, T. M. J., Smith, G. G., Sikkema-de Jong, M. T., Drobisz, J., van Horne, B., & Bus, A. G. (2018). Digital guidance for susceptible readers: Effects on fifth graders' reading motivation and incidental vocabulary learning. *Journal of Educational Computing Research*, *56*(1), 48–73. https://doi.org/10.1177/0735633117708283

*Okita, S. Y. (2008). *Learn wisdom by the folly of others: Children learning to self correct by monitoring the reasoning of projective pedagogical agents* (PQDT:64921281) [Doctoral dissertation]. Stanford University.

*Okita, S. Y. (2014). Learning from the folly of others: Learning to self-correct by monitoring the reasoning of virtual characters in a computer-supported mathematics learning environment. *Computers & Education*, *71*, 257–278. https://doi.org/10.1016/j.compedu.2013.09.018

*Osman, K., & Lee, T. (2014). Impact of interactive multimedia module with pedagogical agents on students' understanding and motivation in the learning of electrochemistry. *International Journal of Science & Mathematics Education*, *12*(2), 395–421. https://doi.org/10.1007/s10763-013-9407-y

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In *The Cambridge handbook of multimedia learning* (2nd ed., pp. 27–42). Cambridge University Press.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–36). Elsevier. https://doi.org/10.1016/B978-012372545-5/50003-4

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions

Questionnaire (AEQ). *Contemporary Educational Psychology*, *36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002

Peng, T.-H., & Wang, T.-H. (2022). Developing an analysis framework for studies on pedagogical agent in an e-learning environment. *Journal of Educational Computing Research*, *60*(3), 547–578. https://doi.org/10.1177/07356331211041701

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, *63*(2), 167–199. https://doi.org/10.2307/1170472

*Plant, E. A., Baylor, A. L., Doerr, C. E., & Rosenberg-Kima, R. B. (2009). Changing middle-school students' attitudes and performance regarding engineering with computer-based social models. *Computers & Education*, *53*(2), 209–215. https://doi.org/10.1016/j.compedu.2009.01.013

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Rehm, M., Nakano, Y., Koda, T., & Winschiers-Theophilus, H. (2012). Culturally aware agent communication. In M. Zacarias & J. V. de Oliveira (Eds.), *Human-computer interaction: The agency perspective* (pp. 411–436). Springer.

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141–160. https://doi.org/10.1037/met0000300

Ryu, J., & Baylor, A. L. (2005). The psychometric structure of pedagogical agent persona. *Technology, Instruction, Cognition and Learning*, *2*(4), 291–315.

Schneider, S., Beege, M., Nebel, S., Schnaubert, L., & Rey, G. D. (2022). The cognitive-affective-social theory of learning in digital environments (CASTLE). *Educational Psychology Review*, *34*(1), 1–38. https://doi.org/10.1007/s10648-021-09626-5

Schroeder, N. L. (2017). The influence of a pedagogical agent on learners' cognitive load. *Journal of Educational Technology & Society*, *20*(4), 138–147.

Schroeder, N. L. (2024). *A beginner's guide to systematic review and meta-analysis*. https://noah-schroeder.github.io/reviewbook/

Schroeder, N. L., & Adesope, O. O. (2014). A systematic review of pedagogical agents' persona, motivation, and cognitive load implications for learners. *Journal of Research on Technology in Education*, *46*(3), 229–251. https://doi.org/10.1080/15391523.2014.888265

Schroeder, N. L., & Adesope, O. O. (2015). Impacts of pedagogical agent gender in an accessible learning environment. *Educational Technology & Society*, *18*(4), 401–411.

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, *49*(1), 1–39. https://doi.org/10.2190/EC.49.1.a

Schroeder, N. L., & Cenkci, A. T. (2020). Do measures of cognitive load explain the spatial split-attention principle in multimedia learning environments? A systematic review. *Journal of Educational Psychology*, *112*(2), 254–270. https://doi.org/10.1037/edu0000372.

Schroeder, N. L., & Gotch, C. M. (2015). Persisting issues in pedagogical agent research. *Journal of Educational Computing Research*, *53*(2), 183–204. https://doi.org/10.1177/0735633115597625

Schroeder, N. L., Romine, W. L., & Craig, S. D. (2017). Measuring pedagogical agent persona and the influence of agent persona on learning. *Computers & Education*, *109*, 176–186. https://doi.org/10.1016/j.compedu.2017.02.015

Schroeder, N. L., Yang, F., Banerjee, T., Romine, W. L., & Craig, S. D. (2018). The influence of learners' perceptions of virtual humans on learning transfer. *Computers & Education*, *126*, 170–182. https://doi.org/10.1016/j.compedu.2018.07.005

Siegle, R. F., Schroeder, N. L., Lane, H. C., & Craig, S. D. (2023). Twenty-five years of learning with pedagogical agents: History, barriers, and opportunities. *TechTrends*, *67*(5), 851–864. https://doi.org/10.1007/s11528-023-00869-3

Sinatra, G. M. (2005). The "warming trend" in conceptual change research: The legacy of Paul R. Pintrich. *Educational Psychologist*, *40*(2), 107–115. https://doi.org/10.1207/s15326985ep4002_5

*Sinoo, C., van der Pal, S., Henkemans, O. A. B., Keizer, A., Bierman, B. P. B., Looije, R., & Neerincx, M. A. (2018). Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management. *Patient Education and Counseling*, *101*(7), 1248–1255. https://doi.org/10.1016/j.pec.2018.02.008

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

van der Meij, H. (2013). Motivating agents in software tutorials. *Computers in Human Behavior*, *29*(3), 845–857. https://doi.org/10.1016/j.chb.2012.10.018

*van der Meij, H., van der Meij, J., & Harmsen, R. (2015). Animated pedagogical agents effects on enhancing student motivation and learning in a science inquiry learning environment. *Educational Technology Research & Development*, *63*(3), 381–403. https://doi.org/10.1007/s11423-015-9378-5

van Lissa, C. (n.d.). *Doing Meta-Analysis in R and exploring heterogeneity using metaforest*. https://cjvanlissa.github.io/Doing-Meta-Analysis-in-R/index.html

Van Mulken, S., Andre, E., & Müller, J. (1998). The persona effect: How substantial is it? In H. Johnson, L. Nigay, & C. Roast (Eds.), *People and computers XIII* (pp. 53–66). Springer.

Veletsianos, G. (2010). Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education*, *55*(2), 576–585. https://doi.org/10.1016/j.compedu.2010.02.019

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2022). *Package "metafor."* https://cran.r-project.org/web/packages/metafor/metafor.pdf

Viechtbauer, W. (n.d.). *Model diagnostics for "rma.mv" objects—Influence.rma.mv*. https://wviechtb.github.io/metafor/reference/influence.rma.mv.html

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Wang, Y., Feng, X., Guo, J., Gong, S., Wu, Y., & Wang, J. (2022). Benefits of affective pedagogical agents in multimedia instruction. *Frontiers in Psychology*, *12*, Article 797236. https://doi.org/10.3389/fpsyg.2021.797236

Wang, Y., Gong, S., Cao, Y., Lang, Y., & Xu, X. (2023). The effects of affective pedagogical agent in multimedia learning environments: A meta-analysis. *Educational Research Review*, *38*, N.PAG-N.PAG. https://doi.org/10.1016/j.edurev.2022.100506

Wu, C., Jing, B., Gong, X., & Ma, X. (2023). The zoomorphic effect: A contribution to the study of images of pedagogical agents for children's learning in instructional videos. *Journal of Computer Assisted Learning*, *39*(5), 1620–1635. https://doi.org/10.1111/jcal.12822

Yilmaz, F. G. K., Olpak, Y. Z., & Yilmaz, R. (2018). The effect of the metacognitive support via pedagogical agent on self-regulation skills. *Journal of Educational Computing Research*, *56*(2), 159–180. https://doi.org/10.1177/0735633117707696

*Yılmaz, R., & Kılıç-Çakmak, E. (2012). Educational interface agents as social models to influence learner achievement, attitude and retention of learning. *Computers & Education*, *59*(2), 828–838. https://doi.org/10.1016/j.compedu.2012.03.020

*Yung, H. I., & Paas, F. (2015). Effects of cueing by a pedagogical agent in an instructional animation: A cognitive load approach. *Educational Technology & Society*, *18*(3), 153–160.

Zhang, S., Jaldi, C. D., Schroeder, N. L., & Gladstone, J. R. (2024). Pedagogical agents in K–12 education: A scoping review. *Journal of Research on Technology in Education*, 1–28. https://doi.org/10.1080/15391523.2024.2381229

Zhang, S., Jaldi, C. D., Schroeder, N. L., López, A. A., Gladstone, J. R., & Heidig, S. (2024). Pedagogical agent design for K–12 education: A systematic review. *Computers & Education*, *223*, 1–13. https://doi.org/10.1016/j.compedu.2024.105165

Authors

NOAH L. SCHROEDER, PhD, is a research scientist at the University of Florida.

SHAN ZHANG is a doctoral student in curriculum and instruction, specializing in educational technology at the University of Florida.

CHRIS DAVIS JALDI is a master's student in computer science at Wright State University.

JESSICA R. GLADSTONE, PhD, is an assistant professor of educational psychology at The University of Illinois Urbana-Champaign.

ALEXIS A. LÓPEZ, PhD, is at Educational Testing Service.

EMMANUEL DORLEY, PhD, is an assistant professor in the Computer & Information Science & Engineering Department at The University of Florida.