Received: 30 March 2021 | Revised: 23 January 2022 | Accepted: 9 February 2022

DOI: 10.1002/tea.21765

RESEARCH ARTICLE

JRST | WILEY

Game-based learning has good chemistry with chemistry education: A three-level meta-analysis

Yuanyuan Hu[1] | **Timothy Gallagher**[1] | **Pieter Wouters**[1] | **Marieke van der Schaaf**[2] | **Liesbeth Kester**[1]

[1]Department of Education, Utrecht University, Utrecht, The Netherlands

[2]Center for Research and Development of Education, University Medical Center Utrecht, Utrecht, The Netherlands

Correspondence
Yuanyuan Hu, Department of Education, Utrecht University, Heidelberglaan 1, 3584 CS, Utrecht, The Netherlands.
Email: y.hu3@uu.nl

Funding information
This work was supported by the European Union's Framework Programme for Research and Innovation Horizon 2020 (Grant Agreement No 812716).

Abstract

Game-based learning (GBL) may address the unique characteristics of a single subject such as chemistry. Previous systematic reviews on the effects of GBL have yielded contradictory results concerning cognitive and motivational outcomes. This meta-analysis aims to: (a) estimate the overall effect size of GBL in chemistry education on cognitive, motivational, and emotional outcomes compared with non-GBL (i.e., media comparison); (b) examine possible moderators of the effects; and (c) identify the more effective game design and instructional design features (i.e., value-added comparison). We screened 842 articles and included 34 studies. This study is the first GBL meta-analysis that employed a three-level random-effects model for the overall effects. Moderator analysis used a mixed-effects meta-regression model. Results from the media comparison suggest chemistry GBL was more effective for cognition ($g = 0.70$, $k = 30$, $N = 4155$), retention ($g = 0.59$, $k = 20$, $N = 2860$), and motivation ($g = 0.35$, $k = 7$, $N = 974$) than non-GBL and the substantial

heterogeneity ($I^2 = 86\%$) for cognitive outcomes. No study reported emotional outcomes, and studies considering value-added comparisons of GBL with versus without specific design features ($k = 3$) were too few to perform a meta-analysis. Moderator analyses implied that except for publication source and sample size, no other moderator was related to effect sizes. There may be the small-study effects, particularly publication bias. Although we conclude that GBL enhances chemistry learning more than non-GBL, the results also make clear that additional high-quality value-added research is needed to identify design guidelines that may further improve chemistry GBL. More GBL meta-analyses on subjects other than chemistry are also needed. As the first GBL meta-analysis that emphasizes emotion, we call for more research on emotion and on relationships between cognition, motivation, and emotion in GBL.

KEYWORDS

chemistry, cognition, game-based learning, meta-analysis, motivation

1 | INTRODUCTION

Every time a new medium emerges, stakeholders expect new opportunities for education (Kirschner & Hendrick, 2020). This also applies to game-based learning (GBL)—a type of learning pedagogy with game play, accompanied by learning goals, learning outcomes, game goals, and game outcomes, in which a game is the medium for learning (Plass et al., 2015). Unlike simulations such as scientific computer models that represent real-world phenomena, games are defined by essential features such as play (Clark et al., 2009; Homer et al., 2020; Ke, 2016), goals (Malone, 1981), rules (Garris et al., 2002), interactivity (Salen & Zimmerman, 2004; Vogel et al., 2006), challenges (Shute & Ke, 2012), and feedback (Plass et al., 2015; Prensky, 2001), as displayed in Figure 1. The application of immersive learning technologies such as virtual reality (VR), augmented reality (AR), or mixed reality (MR) in GBL strengthens this expectation (Checa & Bustillo, 2019; Cummings & Bailenson, 2016; Di Natale et al., 2020; Garzón & Acevedo, 2019; Garzón et al., 2019; Laffey et al., 2019; Merchant et al., 2014; Moreno & Mayer, 2002, 2004; Parong & Mayer, 2018, 2021; Pellas et al., 2018). However, what supports learning, when, and for whom? Game comparison researchers seek to answer two questions: (1) do students learn better from GBL than non-GBL? (media comparison research; Mayer, 2020); and (2) which design features improve the effectiveness of GBL? (value-added research; Mayer, 2020).

The results from 19 systematic reviews (Supplementary material S1) and six meta-analyses across multiple subjects (Table 1) reveal mixed outcomes: some reviews conclude with caution

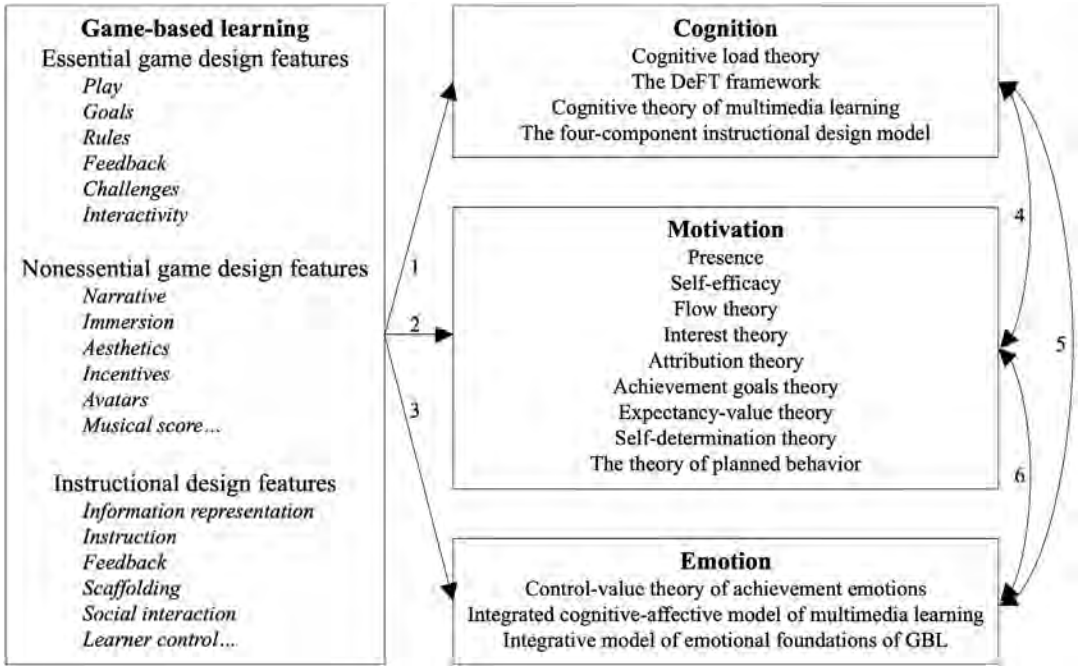

FIGURE 1 The theoretical framework of game-based learning. Single-headed arrows represent causal relations; double-headed arrows represent correlation; rectangles represent independent and dependent variables; all numbers are explained in the text

on the effectiveness of GBL and call for more empirical evidence (Boyle et al., 2016; Connolly et al., 2012; Girard et al., 2013; Martinez-Garza et al., 2013; Mayer, 2019, 2020; National Research Council, 2011a; Young et al., 2012), whereas previous meta-analyses support its cognitive benefits (Clark et al., 2016; Karakoç et al., 2020; Lamb et al., 2018; Sitzmann, 2011; Vogel et al., 2006 ; Wouters et al., 2013), particularly retention (Sitzmann, 2011; Wouters et al., 2013), but not its motivational benefits (Sitzmann, 2011; Vogel et al., 2006; Wouters et al., 2013). This inconsistency may be due to differences among the empirical studies (NRC, 2011a; Vogel et al., 2006; Wouters et al., 2013; Young et al., 2012). Most important, the effectiveness of GBL may depend on the subjects or nature of the content (Acquah & Katz, 2020; Hung et al., 2018; Rutten et al., 2012; Wouters et al., 2013; Young et al., 2012). That is, GBL should address the unique characteristics of each subject. In line with discipline-based educational research (NRC, 2012a; Rahman & Lewis, 2020), meta-analyses examining a particular subject (see Table 1), such as math (Byun & Joung, 2018; Tokac et al., 2019), second language (e.g., English; Chen, Tseng, & Hsiao, 2018; Thompson & von Gillern, 2020), and science (Riopel et al., 2020; Setiawan & Phillipson, 2019; Tsai & Tsai, 2020) will add value to our existing knowledge about GBL.

Perhaps due to the unique characteristics of physics, chemistry, and biology, this inconsistency is even more obvious in science GBL (Cheng et al., 2015; Klopfer & Thompson, 2020; Li & Tsai, 2013; Mayer, 2014b, 2020; NRC, 2011a; Riopel et al., 2020; Setiawan & Phillipson, 2019; Tsai & Tsai, 2020; Wouters et al., 2013; Young et al., 2012). Physics is mainly characterized by highly abstract and idealized mathematical expressions (Docktor & Mestre, 2014; Duit et al., 2007; Opfermann et al., 2017), and biology is mainly characterized by multiple and

TABLE 1 Overview of previous meta-analyses in game-based learning

Study	Year range	Target	Subject	Comparison	K1	Independent variable	Dependent variable	K2	ES
Byun and Joung (2018)	2000–2014	K-12	Math	Media	17	Digital game-based learning	Achievement	25	$d = 0.37^{na}$
Chen et al. (2020)	2008–2019	K-16	All	Value	25	Competition	Cognitive outcomes	25	$g = 0\ .37$*
							Non-cognitive outcomes	82	$g = 0.40$*
M.H. Chen et al. (2018)	2003–2014	All	English	Media	10	Digital game-based learning	Vocabulary acquisition	10	$d = 1.03$*
Clark et al. (2016)	2000–2012	K-16	All	Media value	69	Digital games	Cognitive learning outcomes	173	$g = 0.35$*
							Intrapersonal learning outcomes	35	$g = 0.35$*
						Enhanced design	Learning outcomes	40	$g = 0.34$*
Karakoç et al. (2020)	2000–2018	K-16	All	Media	35	Game-based learning	Achievement	38	$g = 1.70$*
Lamb et al. (2018)	2002–2015	K-14	All	Media	28	Serious educational games	Cognition	na	$d = 0.67^{na}$
						Serious games	Affect	na	$d = 0.51^{na}$
						Simulations	Behavior	na	$d = 0.04^{na}$
Riopel et al. (2020)	2020	All	Science	Media	79	Serious games incl. simulations	Declarative knowledge	65	$d = 0.34$*
							Retention	8	$d = 0.31$*
							Procedural knowledge	7	$d = 0.41$*
Setiawan and Phillipson (2019)	2010–2017	K-8	Science	Media	12	Digital games	Cognitive outcomes	12	$g = 0.67$*
Sitzmann (2011)	1976–2009	>18 years old	All	Media	65	Computer-based simulation games	Declarative knowledge	39	$d = 0.28^{na}$
							Retention	8	$d = 0.22^{na}$
							Procedural knowledge	22	$d = 0.37^{na}$
							Self-efficacy	8	$d = 0\ .52^{na}$
Thompson & von Gillern, 2020)	−2017	K-16	English	Media	19	Video game-based learning	Vocabulary acquisition	20	$d = 0.70^{na}$
Tokac et al. (2019)	2000–2012	K-12	Math	Media	24	Game-based learning	Achievement	39	$d = 0.13$*
Tsai and Tsai (2020)	2000–2018	K-16	Science	Media	26	Digital games incl. simulations	Knowledge acquisition	14	$g = 0.65$*
				Value				13	$g = 0.41$*

TABLE 1 (Continued)

Study	Year range	Target	Subject	Comparison	K1	Independent variable	Dependent variable	K2	ES
Vogel et al. (2006)	1986–2003	All	All	Media	32	Games	Cognitive gains	na	$z = 6.05$*
						Interactive simulations	Attitude	na	$z = 13.74$*
Wouter & Oostendorp (2013)	1990–2012	All	All	Media	39	Serious games	Knowledge	25	$d = 0.27$*
							Skills	52	$d = 0.29$*
							Retention	16	$d = 0.36$*
							Motivation	31	$d = 0.26$
Wouters et al. (2013)	1990–2012	All	All	Value	29	Instructional support	Knowledge	36	$d = 0.33$*
							Skill	32	$d = 0.62$*
							In-game performance	38	$d = 0.19$*

Notes: ES, effect size; na, not available; K1, Number of primary studies included in the meta-analysis; K2, Number of pairwise comparisons in each category.
*$p < .05$.

hierarchical levels of organization in living organisms (NRC, 2009; Tsui & Treagust, 2013; Wandersee et al., 2000). Although multilevel thinking plays a role in many STEM subjects such as physics and biology, chemistry is mainly characterized by multilevel thinking (American Chemical Society, 2018; de Jong & Taber, 2007; Gilbert & Treagust, 2009; NRC, 2009; NRC, 2012a). The triple nature of chemistry is difficult to learn mostly because students struggle to coordinate thinking within three unique levels of chemical knowledge: (1) *macro*—tangible and visible phenomena, such as chemical reactions; (2) *submicro*—invisible atoms, ions, molecules, or structures; and (3) *symbolic*—representational symbols, formulas, or equations (called Johnstone's triangle or chemistry triplet; de Jong & Taber, 2007; Gilbert & Treagust, 2009; Johnstone, 1991, 2000; Sirhan, 2007; Taber, 2009, 2013; Talanquer, 2011; Towns & Kraft, 2011).

To date, the biggest challenge in education such as chemistry is how to create effective, efficient, motivating, and enjoyable learning experiences (Neelen & Kirschner, 2020), particularly how to increase chemistry literacy, motivate learners to learn chemistry, and/or pursue chemistry-related advanced degrees and careers (European Commission, 2015; NRC, 2011a, 2011b, 2012b, 2012c, 2014). Theoretically, this challenge could be addressed by instructional methods such as GBL (Cooper & Stowe, 2018; Klopfer & Thompson, 2020). First, interactivity in combination with multiple representations in GBL requires learners to connect all three levels of chemistry knowledge and switch from one level to another, which may help overcome learning difficulties in Johnstone's triangle (de Jong & Taber, 2007). For example, to figure out how suspects make fake coins, players must watch an animation of zinc, water, and chloride (submicro), write its chemical equation (symbolic), and conduct a gold rush experiment in virtual labs (macro; Hodges et al., 2018). In this process, GBL can also demonstrate the chemical phenomena, visualize the underlying submicroscopic processes, and show symbolic representations. Second, real-time feedback in GBL enables learners to identify chemistry content that they may be struggling with. Third, as the essential activity of GBL (Sicart, 2014), play is critical for cognitive and emotional development (Homer et al., 2020). For example, play allow learners to retain multiple representations of the same subject (Plass et al., 2015), which may also help multilevel thinking. Fourth, challenges in GBL that are neither too easy nor too difficult ask players to master certain chemistry content before moving to next level and playing games (e.g., *Sokobond*) is usually fun, which may help support a zone of proximal development (Vygotsky, 1978), motivate, and enjoy learning chemistry (Homer et al., 2020; Malone, 1981). Fifth, through GBL, learners can enjoy experiences free of real-life constraints and practice repeatedly. For example, although lab works play a central role in secondary (ACS, 2018) and higher education (ACS, 2015a, 2015b), sometimes they are rarely implemented due to limited curriculum time or costly infrastructures such as nuclear magnetic resonance (NRC, 2011a, 2014). GBL can create virtual laboratories or scenes to conduct scientific inquiry (e.g., *HoloLAB Champions*), particularly in dangerous experiments and environments that are physically inaccessible (Parker et al., 2008), which may help develop chemical practices (NRC, 2011a, 2012b). Thus, GBL has great potential to boost chemistry education.

Empirically, little attention has been paid to chemistry (Cheng et al., 2015)—the central science that connects physics and biology (Brown et al., 2018). In previous systematic reviews, most primary research in science GBL was conducted in physics or biology (Cheng et al., 2015; Li & Tsai, 2013), with only five studies in chemistry. This may be due to limitations such as examining media comparison research without value-added research (e.g., Riopel et al., 2020; Setiawan & Phillipson, 2019), a single learning outcome (e.g., achievement; Riopel et al., 2020; Setiawan & Phillipson, 2019; Tsai & Tsai, 2020; Young et al., 2012), a narrow publication source

(e.g., peer-reviewed articles; Setiawan & Phillipson, 2019), and/or narrow range of grades (e.g., K-8; Setiawan & Phillipson, 2019; K-12, Young et al., 2012).

Although GBL has been emerging in chemistry education over the past 20 years, little is known about its effectiveness. Therefore, systematic knowledge is needed about whether GBL makes a difference in chemistry education and how to support GBL chemistry (Bellou et al., 2018). Hence, this meta-analysis investigates the effects of GBL (media comparison) and game and instructional design features (value-added comparison) on chemistry learning; that is, to estimate the effect size, indicate whether the effect size is consistent across empirical studies, and/or to identify more sources of diversity (Borenstein et al., 2009). To include exhaustive studies, we broadened the learning outcomes, age groups, and publication sources.

1.1 | Learning outcomes in chemistry game-based learning

Chemistry learning involves not only scientific practices (e.g., ask questions; develop and use models; plan and carry out investigations), crosscutting concepts that bridge across other disciplines (e.g., patterns; cause and effect; scale, proportion, and quantity), and chemistry core ideas (e.g., matter and its interactions; energy) but also motivation (e.g., attitude; interest) and feelings toward chemistry (e.g., emotions; ACS, 2018; European Commission, 2015; Forsthuber et al., 2011; NRC, 2012a, 2012b, 2014, 2016; Schola Europaea, 2019). Theoretically, GBL can impact chemistry learning by affecting cognitive processes, motivation to learn, and/or emotion (NRC, 2011a; Plass et al., 2015; Plass et al., 2020). Design features of GBL match multiple theories of learning (cognitive, motivational, and emotional) to a larger extent than other media or methods (e.g., studying a web lecture). We developed the general theoretical framework of GBL (see Figure 1). Features of GBL include instructional design features (e.g., teacher support, task design, and peer interactions) and game design features (e.g., the design of the interface, number of sessions, and challenges) which are further divided into essential game design features (e.g., rules) and nonessential game design features (e.g., narratives). To solve the aforementioned challenges regarding low motivation to learn chemistry, we focus on motivation to learn a subject (e.g., chemistry) instead of motivation to play the game.

1.1.1 | Cognition in chemistry game-based learning

From the cognitive perspective, the goal of chemistry education involves scientific practices related to chemistry core ideas and crosscutting concepts (ACS, 2018; NRC, 2012a, 2012b, 2014; Schola Europaea, 2019). Generally, GBL may affect cognitive processes underlying learning chemistry such as schema construction and schema automation, which is grounded in learning theories such as cognitive load theory (CTL; Sweller et al., 1998, 2019), cognitive theory of multimedia learning (CTML; Mayer, 2014a, 2014b, 2020), and the four-component instructional design model (4C/ID model; van Merriënboer & Kirschner, 2018), as displayed in Figure 1 line 1. According to CTML, meaningful GBL happens when players learn by active processing, namely selecting relevant information in the game, mentally organizing it as visual and verbal representations into a coherent structure, and integrating these representations with prior knowledge (schema construction; Moreno & Mayer, 2007). Take the previous gold rush game for example. GBL can provide multiple representations: Learners learn different representations, mentally relate representations to one another, and integrate them into coherent mental

models (multimedia principle; Mayer, 2014a), which may foster the aforementioned multilevel thinking (Chiu & Wu, 2009; Wu & Shah, 2004). According to the DeFT framework (Ainsworth, 2006), multilevel thinking requires not only multiple representations but also dynamic linking between these representations (multiple representation principle; Ainsworth, 2014), which can be facilitated by interactivity in GBL. Furthermore, learning chemistry such as multilevel thinking may pose considerable cognitive demands on learners and GBL may affect three demands: essential processing aiming at mentally representing the essential material, generative processing aiming at making sense of materials, and extraneous processing that does not contribute to learning Mayer, 2014b, 2020). For example, in HoloLAB Champions, narrative by the virtual host provides a relevant and meaningful context for scientific practices (situational learning; Plass et al., 2015; Prensky, 2001), which may facilitate essential processing; Game interactivity allows learners to learn chemistry lab skills by doing, which may facilitate generative processing (e.g., Moreno & Mayer, 2005); and Ongoing feedback assesses learners' performance and directs their attention to relevant information, which may reduce extraneous processing (Johnson et al., 2017). Given that players' cognitive capacity is limited, complex GBL, particularly when multiple representations are involved, is demanding. Thus, game design and instructional design aim to optimize cognitive processes and outcomes via managing essential processing (e.g., reduce game complexity by pretraining), minimizing extraneous processing (e.g., remove seductive details), and fostering generative processing (e.g., scaffolding; Mayer, 2014b, 2020), which is the focus of value-added research.

1.1.2 | Motivation in chemistry game-based learning

From the motivational perspective, the goal of chemistry education is to increase motivation to learn, complete degrees, or pursue careers in chemistry (ACS, 2018; European Commission, 2015; NRC, 2012a, 2012b, 2014; Schola Europaea, 2019). Generally, GBL may affect players' values, needs, beliefs, attributions, and goals of learning chemistry (for an overview and comparison, see Cook & Artino, 2016; de Brabander & Martens, 2014; Mayer, 2014b, 2020; Plass et al., 2015), which is grounded in motivation theories such as self-determination theory (Deci & Ryan, 2000), expectancy-value theory (Wigfield & Eccles, 2000), achievement goal theory (Elliot et al., 2011), self-efficacy (Bandura, 1986), attribution theory (Weiner, 1985), the theory of planned behavior (Ajzen, 1991), flow theory (Csikszentmihalyi, 1975, 1990), presence (Cummings & Bailenson, 2016; Lee, 2004), and interest theory (Hidi & Renninger, 2006; Schiefele, 2009), as displayed in Figure 1 line 2. For example, according to the player experience of the need satisfaction model, features of GBL environments such as HoloLAB Champions can support basic psychological needs for autonomy (e.g., choices regarding the level of challenge, strategies, or tools), competence (e.g., experience growth and leveling up by optimal challenge), and relatedness (e.g., opportunities to contribute, communicate, and cooperate with the virtual host), resulting in intrinsic motivation and cognition (Ryan & Rigby, 2020). Thus, game design and instructional design also aim to increase motivation to learn (Plass et al., 2020), such as using an incentive system and motivating music in HoloLAB Champions.

1.1.3 | Emotion in chemistry game-based learning

Emotions are also involved in chemistry learning (Jaber & Hammer, 2016; King et al., 2017; Maria et al., 2003; NRC, 2012a; Raker et al., 2019; Sinatra et al., 2014). Generally, GBL can

induce different types of emotions in chemistry learning (e.g., achievement emotions, epistemic emotions, and topic emotions) by shaping their antecedents (e.g., perceived control and perceived value of the learning tasks, cognitive incongruity; for details, see Loderer et al., 2020; Plass et al., 2019). These assumptions are grounded in emotion theories such as the control-value theory of achievement emotions (CVT; Pekrun & Perry, 2014), integrated cognitive-affective model of media (ICALM; Plass & Kaplan, 2015), and the integrative model of emotions in game-based learning (EoGBL; Loderer et al., 2020), as displayed in Figure 1 line 3. For example, according to CVT, the optimal level of challenges and scaffolding in HoloLAB Champions may promote a higher perceived control and value of learning chemistry, and, consequently, induce more positive achievement emotions (e.g., enjoyment) and less negative achievement emotions (e.g., boredom). Thus, game design and instructional design, particularly emotional design, aim to trigger more positive emotions and less negative emotions (Loderer et al., 2020; Plass et al., 2015, 2019), such as using happy expression, warm color, and round shape rather than sad and neutral expression, cold color, and square shape (Park et al., 2015; Plass et al., 2014, 2015, 2019; Plass & Kaplan, 2015; Um et al., 2012).

Furthermore, chemistry game designers and researchers must consider which design features facilitate cognitive process, motivation, and emotions in players (Figure 1 lines 4, 5, and 6) because they may influence each other (e.g., Pekrun & Linnenbrink-Garcia, 2014; Robbins et al., 2004; Talsma et al., 2018; Valentine et al., 2004). Such influence may apply to chemistry GBL contexts. First, past performance in GBL (e.g., success or failure) may be the sources of motivation (e.g., self-efficacy; Bandura, 1986) and (achievement) emotions (e.g., enjoyment; Pekrun & Perry, 2014). For example, a successful performance on one chemistry game level is likely to promote higher motivation and more positive emotions on the following level. Second, GBL motivates players to invest sustained effort and time to engage in selecting, organizing, and integrating information, improving learning and emotion (Mayer, 2014b, 2019). For example, a higher motivation on a chemistry game is likely to promote higher performance and more positive emotions. Third, positive emotions in GBL induce intrinsic motivation to invest effort (Loderer et al., 2018), reduce cognitive load (Plass & Kaplan, 2015), sustain attention on relevant information (Park et al., 2015), lead to flexible and creative learning strategies (Fiedler & Beier, 2014), facilitate self-regulated learning (Artino & Jones, 2012; Pekrun & Perry, 2014), and, consequently, increase performance (Loderer et al., 2020; Sabourin & Lester, 2014). For example, more positive emotions on a chemistry game are likely to promote higher motivation and performance. Unfortunately, data from included studies in this meta-analysis do not allow us to formulate a research question on relations between cognition, motivation, and emotion in chemistry GBL. Further research on chemistry GBL is needed to confirm this assumption.

1.2 | The present study

Although GBL and chemistry education align, overview research regarding the learning effects and the determining factors is limited. Despite focusing on chemistry education, this meta-analysis builds on and considers some limitations from previous meta-analyses. First, we focus on unresolved issues: whether GBL is more motivating (Sitzmann, 2011; Vogel et al., 2006; Wouters et al., 2013) than *non-GBL* (i.e., any type of learning activities without using games, such as learning with lectures) and which design features enhance GBL (Chen et al., 2020; Clark et al., 2016; Tsai & Tsai, 2020; Wouters & van Oostendorp, 2013). Second, to avoid deviating definitions of key concepts including cognitive gains (Vogel et al., 2006), motivation (Clark

et al., 2016; Lamb et al., 2018), or simulation games (Riopel et al., 2020; Setiawan & Phillipson, 2019; Sitzmann, 2011), we define GBL narrowly by the aforementioned essential game design features such as play (i.e., exclude pure simulations) and classify learning outcomes into cognitive, motivational, and emotional outcomes. Third, as learning content varies, game genres vary and, consequently, game effectiveness may also vary (Wouters et al., 2013). Given that game genre is critical in game design and depends on to-be-learned knowledge, we examine it and follow Chen et al. (2020) and Ke's (2016) classification of game genre such as role-playing (see Table 2). Fourth, some studies only included attitude (Vogel et al., 2006), self-efficacy (Sitzmann, 2011), interest, and engagement (Wouters et al., 2013) as motivational outcomes, whereas we include other motivation theories, such as expectancy-value theory and achievement goal theory. Fifth, the standard (two-level) meta-analytic model used by previous meta-analyses did not consider dependency of effect sizes within studies (e.g., one study reported multiple comparisons, or multiple measurements for the same outcome; Cheung, 2014, 2019; López-López et al., 2018). Instead, we use a three-level meta-analytic model to estimate between- and within-study variance. Sixth, although publication bias (i.e., studies with statistically significant or positive results tend to be published more often than those with not statistically significant or negative results; Rosenthal, 1979) is one of the biggest issues in meta-analyses (Fernández-Castilla et al., 2021), some studies missed many commonly used methods when detecting and correcting publication bias (e.g., Riopel et al., 2020; Sitzmann, 2011), such as the Egger test which has been shown a better correction for publication bias than other methods (Stanley & Doucouliagos, 2014; for detailed comparison of methods, see Fernández-Castilla et al., 2021; Kromrey & Rendina-Gobioff, 2006). Progress in statistical analysis techniques enables us to use more recent and reliable meta-analysis methods, such as meta-regression to detect publication bias and investigate continuous moderators such as sample size.

This meta-analysis systematically synthesizes all experimental studies that applied GBL in K-16 chemistry education by addressing the following questions:

> **RQ1.** Is the effect of chemistry GBL on cognitive (including retention), motivational, and emotional outcomes larger than for non-GBL (media comparison)?

For cognitive outcomes, GBL changes academic knowledge, which is often measured by immediate and/or delayed tests (Mayer, 2020). As learning and instruction seek to promote the storage of learned knowledge in long-term memory that can be retrieved when needed (Bennett & Rebello, 2012; Paas & Sweller, 2014), delayed tests are advocated to determine the long-lasting impact of GBL (i.e., long-term retention) instead of fleeting knowledge improvement due to arousal (Mayer, 2014b). Furthermore, retention is a key learning outcome in chemistry education (NRC, 2012a, 2014, 2015). For motivational outcomes, that GBL are motivating is the most frequent appeal of GBL (Malone, 1981; Plass et al., 2015; Wouters et al., 2013). For emotional outcomes, decreasing boredom and increasing enjoyment is another appeal of GBL (Loderer et al., 2020). Previous meta-analyses generally found small to large effect sizes for the different cognitive outcomes (Karakoç et al., 2020; Lamb et al., 2018; Riopel et al., 2020; Setiawan & Phillipson, 2019; Sitzmann, 2011; Tsai & Tsai, 2020; Wouters et al., 2013) and retention (Riopel et al., 2020; Sitzmann, 2011; Wouters et al., 2013) in favor of GBL relative to non-GBL but disagree about the effects on motivational outcomes (Sitzmann, 2011; Vogel et al., 2006; Wouters et al., 2013), and little is known about emotional outcomes (see Table 1). Fortunately, a meta-analysis on emotions in technology-based learning concludes enjoyment and curiosity are positively related to achievement in GBL (Loderer et al., 2018). Based on this evidence, we hypothesize the following:

TABLE 2 Coding for basic information, learning outcomes, and moderator variables

Variables	Categories
Basic information	Author, publication year, grade, country, comparison, game name, chemistry topic, and assessment method
Game genre	1. Puzzle game with logic thinking, pattern recognition, objects matching, or questions answering (e.g., trivia) 2. Action game with combat and physical challenges such as shooting 3. Adventure game with exploring, gathering items, and solving puzzles driven by story 4. Strategy game with system thinking and via a decision tree 5. Role-playing game when players assume the roles of characters 6. Simulation game which simulates reality
Learning outcomes	1. Cognition incl. factual, conceptual, procedural, and/or strategic knowledge 2. Retention when cognition was measured in a delayed test 3. Motivation in chemistry incl. learning attitude, interest, intrinsic motivation, self-determination, achievement goal, task value, flow, presence, and/or self-efficacy 4. Emotion incl. enjoyment, pride, hope, anxiety, anger, shame, boredom, or hopeles
level of control group	1. Active incl. doing experiments, computer-based tutorials, assignments, or exercises 2. Passive incl. reading textbooks, listening to lectures, or watching videos
Additional instruction	1. The game with additional instructions * (e.g., pretraining before GBL, debriefing after GBL) 2. The game without additional instructions (i.e., GBL is standalone)
User grouping	1. Single: play the game individually 2. Multiple: play the game in groups
No. of game sessions	1. Single: play the game once 2. Multiple: play the game repeatedly
Sample size	The actual number of participants
Publication source	1. Gray literature incl. theses or conference proceedings 2. Published incl. books or peer-reviewed journals
Randomization	1. Random controlled trial: randomly assign the participants (not the class) to groups 2. Quasi-experiment design: no random assignment
Assessment type	1. Closed incl. only multiple-choice questions 2. Non-closed incl. short-answer questions or open-ended questions with or without multiple-choice questions 3. Mix incl. closed questions together with non-closed questions

Notes: For example, if a lecture is given before GBL, it counts as additional instruction (e.g., pretraining); if it is given without GBL, it counts as non-GBL. na, unknown was coded when relevant information is missing.

Hypothesis 1. Chemistry GBL yields higher cognitive outcomes than non-GBL.

Hypothesis 2. Chemistry GBL yields higher retention than non-GBL.

Hypothesis 3. Chemistry GBL yields higher motivational outcomes than non-GBL.

Hypothesis 4. Chemistry GBL induces more positive emotions and less negative emotions than non-GBL.

RQ2. Do instruction characteristics (activity level of control group, additional instruction, user grouping, and number of game sessions) and methodology characteristics (randomization, sample size, publication source, and assessment type) moderate the effect?

We identified the following instruction characteristics as moderators based on the aforementioned cognitive foundations of GBL and inconclusive results from previous meta-analyses in GBL (see Table 2). First, *activity level of control group*. According to CTML (Mayer, 2014a, 2014b, 2020) and CTL (Sweller et al., 2019), GBL fosters generative processing in which learners actively engage in selecting, organizing, and integrating new information, but this is also true for non-GBL. Active processing is key to learning; the deeper the processing of information, the more that will be retained and encoded into memory (Craik & Lockhart, 1972); thus, the difference between GBL and non-GBL may decline when non-GBL uses active instead of passive instruction. Second, *additional instructions* (i.e., instructions that are used together with GBL rather than non-GBL, such as pretraining before GBL). Organizing and integrating information is critical for learning, but these do not occur automatically (Mayer, 2014a, 2014b). Integrating GBL with non-game instructions (e.g., pretraining;Clark et al., 2016 ; Wouters et al., 2013) may facilitate articulating and integrating new knowledge with prior knowledge, leading to higher recall, transfer, and retention than standalone GBL (Merrill, 2012; Wouters et al., 2008; Wouters et al., 2013; Young et al., 2012). Third, *user grouping*. Playing games in groups and explaining things to each other may also facilitate knowledge articulation and, thus, the organization and integration of new information. This point is supported by the collaboration principle in multimedia learning, also known as the collective working memory effect (Kirschner et al., 2009, 2011), which states it is better to assign complex learning tasks in groups (van Merriënboer & Kester, 2014). Fourth*, number of game sessions*. GBL can be complex for novices. When they start playing a game, they must learn technological knowledge—game information (extraneous processing because it does not contribute to learning) and content knowledge (essential and generative processing); thus, they may easily become overwhelmed. Multiple game sessions allow the players to get familiar with the game.

Theoretically, the efficacy of GBL relative to non-GBL may improve in specific learning arrangements—such as additional instruction, group gameplay, multiple sessions, or passive instruction in non-GBL—that facilitate information processing. Empirically, previous meta-analyses only agree on the role of number of game sessions; that is, compared with non-GBL, learners benefited more when GBL involved multiple sessions, and no difference was found between single sessions and non-GBL (Clark et al., 2016; Wouters et al., 2013). However, the meta-studies are unsure regarding additional instruction (Clark et al., 2016; Sitzmann, 2011; Wouters et al., 2013), user grouping (Tsai & Tsai, 2020; Vogel et al., 2006; Wouters et al., 2013), and activity level of control group (Riopel et al., 2020; Sitzmann, 2011; Wouters et al., 2013). Therefore, we only formulated a hypothesis regarding the number of game sessions. For other variables, we investigated whether and to what extent they moderate the overall effect.

Hypothesis 5. Relative to non-GBL, chemistry GBL with multiple game sessions yield higher learning outcomes than those with single sessions.

To check study quality, we included the following methodology characteristics as moderators: randomization, sample size, publication source, and assessment type (see Table 2). Ideally, large sample sizes and randomized controlled trials (RCTs) are recommended by review organizations, such as What Works Clearinghouse (WWC, 2019). Practically, effect sizes were found to be systematically higher in quasi-experiments designs (QEDs) than in RCTs, in smaller than larger studies, and in published studies than gray literature due to methodological weaknesses and small-study effects (Cheung & Slavin, 2016; Slavin, 2008; Slavin & Smith, 2009). Although closed assessment (e.g., multiple-choice questions) is easier to implement than non-closed assessment (e.g., open-ended questions), effect sizes seem larger when using non-closed assessment (Tsai & Tsai, 2018). Again, previous GBL meta-analyses provide inconclusive results on the moderating effects of these methodology characteristics (Karakoç et al., 2020; Riopel et al., 2020; Sitzmann, 2011; Tsai & Tsai, 2018; Wouters et al., 2013). Therefore, it is valuable to check publication bias. We further evaluate the extent of bias, estimate unbiased effects, and suggest improvements for future research; simply excluding unpublished studies would ignore publication bias and overestimate the overall effects.

RQ3. Which game design or instructional design features improve chemistry GBL (value-added comparison)?

As displayed in Table 3, value-added research pinpoints design features that promote GBL by reducing extraneous processing (e.g., redundancy), managing essential processing (e.g., modality), and/or fostering generative processing (e.g., personalization; Mayer, 2020). Previous meta-analyses have suggested that some instructional design features such as modality, personalization, feedback (Tsai & Tsai, 2020; Wouters & van Oostendorp, 2013), competition (Chen et al., 2020), and/or enhanced scaffolding (e.g., personalized scaffolding based on individual learner needs; Clark et al., 2016) enhance GBL. However, other features remain unsettled, including game design features such as narrative (integrate a storyline; Wouters & van Oostendorp, 2017) or immersion (use VR), and instructional design features such as collaboration (play in groups; Clark et al., 2016), learner control (allow learners to choose game levels), or segmenting (break the materials into parts; Mayer, 2020). Given the limited research evidence, we do not formulate a hypothesis but explore the efficacy of these features in chemistry GBL.

TABLE 3 Features that promote GBL

Features	Descriptions	Cognitive processing
Pretraining	Provide trainings on key concepts and characteristics before gameplay	Essential processing
Modality	Present words in spoken rather than written forms	
Personalization	Use conversational rather than formal styles	Generative processing
Feedback	Add explanations and advice to corrective feedback	
Self-explanation	Provide prompts to self-explain the performance during gameplay	
Competition	Play the game against virtual components, time, or other players	
Redundancy	Eliminate redundant information, such as written text from spoken texts or pictures	Extraneous processing

2 | METHOD

2.1 | Literature search

The PRISMA flow diagram (Moher et al., 2009) in Figure 2 summarizes the process of the literature search and selection. The search comprised three parts. First, searching databases: Web of Science, Scopus, Eric, and PsycINFO. The search terms were combined in English: "(chemistry or chemical) AND (game or immersive learning or virtual reality or virtual environment or augmented reality or augmenting reality or mixed reality) AND (learning or cognit* or achievement or interest or attitude or engage* or motivation* or involvement or enjoy* or emotion* or affective)", which should be in the title, abstract, or the list of keywords. In case the term "game" was not in the title, abstract, or keywords, we included immersive learning technologies commonly used in games as the search terms, such as VR. The period searched was from 2000, when GBL studies changed dramatically after that due to technological development (Parker et al., 2008) and GBL started becoming popular in chemistry, to January 2020. Second, a Google Scholar search for gray literature (Haddaway et al., 2015). Finally, a snowball search in the reference lists and citations of the aforementioned meta-analyses and systematic reviews. Overall, 1156 records came out, and 842 remained after removing duplicates.

2.2 | Inclusion and exclusion criteria

Studies were evaluated based on (a) language, (b) subject, (c) participants, (d) accessibility, (e) comparison, (f) independent variable, (g) dependent variable, (h) data, and (i) others

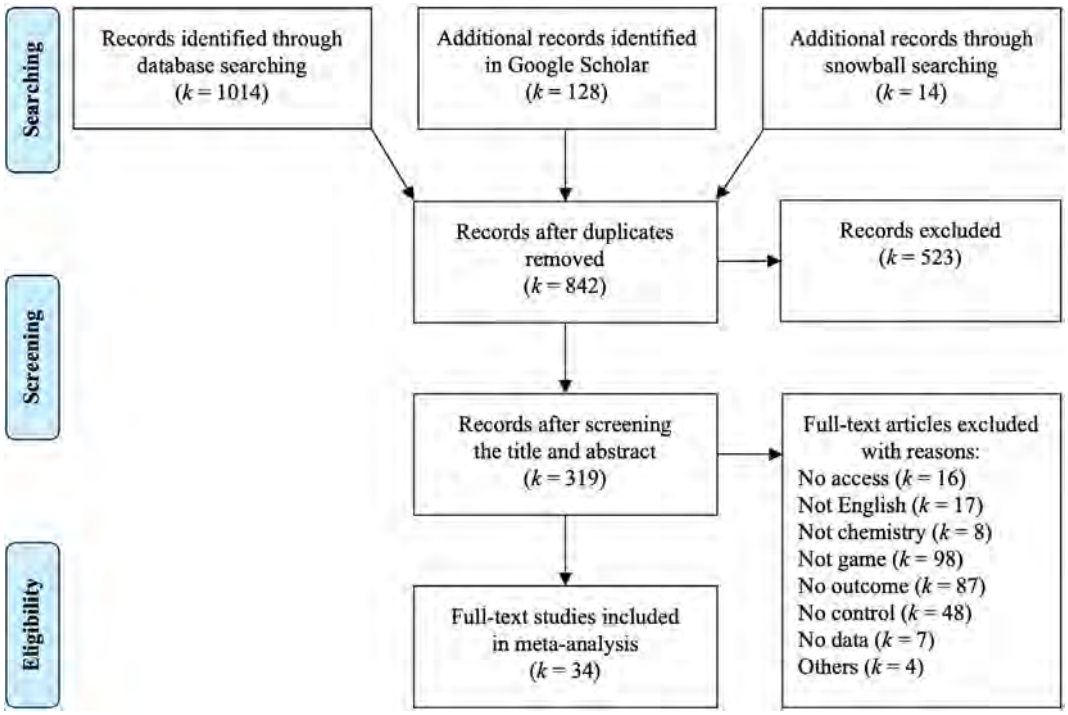

FIGURE 2 PRISMA flow diagram of literature search and records selection process

(Table 4). Study inclusion followed two stages (Figure 2). First, in the screening stage, all the titles and abstracts were screened based on a, b, f, and h, which led to 319 records being included and 523 excluded. If the study did not state whether it fulfilled these four criteria, the researchers included it and made a further decision based on its full text in the next stage. To assess the inter-rater reliability (IRR), 91 records (>10% of 842) were randomly chosen and screened independently by the first two authors, with Cohen's $k = 0.83$. Any disagreements were resolved by discussing and consulting with the third author. Second, the full texts of 319 records were retrieved and evaluated using all nine criteria. The first two authors randomly selected 34 full texts (>10% of 319) and assessed them independently. Despite two disagreements regarding the reason for exclusion, a perfect IRR was reached (Cohen's $k = 1$). In total, we included 34 articles.

2.3 | Coding

To conduct a quantitative analysis and provide a qualitative description of all the included studies, we collected the following data: basic information, game genre, learning outcomes, and moderator variables (Table 2). The coding process was performed in a standardized way. First, a trial coding with five studies was run to evaluate whether all possible situations for a moderator variable were covered by a category. Then, a random sample of four studies (>10% of 34) was coded independently by the first two authors. A satisfactory IRR with Cohen's $k = 1$ was reached for all variables, except for number of game sessions (Cohen's $k = 0.5$). Any disagreements were discussed until agreement was reached. For this variable, another four articles were

TABLE 4 Inclusion and exclusion criteria

Criteria	Inclusion	Exclusion
(a). Language	English	Language other than English
(b). Subject	Chemistry	Integrated science
(c). Participants	Nondisabled K-16 students	Preservice teachers, student teachers, or employees
(d). Accessibility	The full text of the study is accessible	No access via on Internet or contacting authors
(e). Comparison	Game vs. nongame or game with vs. without specific features	No control group, control group without learning the same subject matter
(f). Independent variable	Game-based learning or serious games with the term "game" appearing in the article	Games for entertainment, or educational applications, technologies, or tools (e.g., VR, AR, MR) without using the term "game"
(g). Dependent variable	Cognitive, motivational, or emotional outcome	Introduction, assessment, motivation, or perception of the game
(h). Data	Sufficient data to calculate effect size	Case study, no empirical data, or enough data
(i). Others		The study published in conferences or theses was updated and later published in journals

randomly chosen and coded by the first two authors with IRR of Cohen's $k = 1$. Finally, the first author coded the remaining studies. A detailed coding for moderators is available in Table S1.

2.4 | Calculating effect sizes

Standardized mean difference (mean difference between experiment groups and control groups divided by the pooled standard deviation) was adopted as effect size (Borenstein et al., 2009), and *g* (Hedges, 1981) was calculated using Comprehensive Meta-Analysis (CMA v3; Borenstein et al., 2013). The effect size for each study was computed in a hierarchical order: the first is raw data (mean and standard deviation) and the second is data from inferential statistics (e.g., *t* or *F* value; Wouters et al., 2013).

The following complex cases were evaluated cautiously. First, when studies used a pretest–posttest control group design ($k = 23$), the preexisting difference between the experimental and control groups should be considered, and thus, the effect size was estimated on both pretest and posttest data (Morris, 2008). In practice, effect sizes were computed by post-pre mean difference of experimental group minus post-pre mean difference of control group divided by the pooled standard deviation of post-scores. In addition, when different sample sizes in the pretest and posttest were reported, the smaller sample size was adopted (da Silva Júnior et al., 2018).

Second, when studies used multiple experimental or control groups ($k = 5$), the recommended solution was to form multiple pairwise comparisons and calculate multiple effect sizes (e.g., GBL vs. concept mapping and GBL vs. conventional lecture; Okonkwo, 2012).

Third, when studies recorded multiple measurements of the same outcome ($k = 8$), multiple effect sizes were calculated, as suggested by Cheung (2014, 2019) and López-López et al. (2018)). For instance, knowledge comprehension and knowledge application that were assessed separately were calculated separately as the indicator for cognitive outcomes (e.g., Chen et al., 2014; Chen & Liao, 2015). Intrinsic motivation, self-determination, self-efficacy, grade motivation, and career motivation were calculated separately as the indicator for motivation in chemistry (e.g., Meesuk & Srisawasdi, 2014; Srisawasdi & Panjaburee, 2019).

Another special case was Johnson-Glenberg et al. (2014) using the AB-BA design: Two groups were given GBL intervention (A) and regular instruction (B) in different sequences. Group 1 received pretest, SMALLab GBL intervention, mid-test, regular instruction, and posttest, while group 2 received regular instruction first after pretest and SMALLab GBL intervention after mid-test. When the mid-test was conducted, groups 1 and 2 had only received GBL intervention (SMALLab) and regular instruction, respectively. Therefore, the treatment before the mid-test was considered as a single-pair comparison (group 1 as GBL group and group 2 as control group) and the mid-test instead of the "posttest" at the end of the study was taken as the posttest.

2.5 | Data analysis

All statistical analyses were run via the "metafor" package (version 3.1.8; Viechtbauer, 2010) in R (version 4.1.0), except the distribution of variance across levels that was run via the "dmetar" package (Harrer et al., 2019). Because some studies reported multiple measures of the same construct (e.g., cognition was measured by knowledge comprehension and knowledge

application) or multiple comparisons (e.g., two GBL groups vs. one non-GBL group), multiple effect sizes could arise per study and these effect sizes are dependent within studies. Separate meta-analyses were performed for cognition, retention, and motivation using the random-effects three-level meta-analytic model (Cheung, 2014, 2019; López-López et al., 2018). The three-level meta-analytic model includes sampling variance (level 1), within-study variance (level 2), and between-study variance (level 3). Heterogeneity was assessed using Cochran's Q test, and I^2 and τ^2 statistics. We used the Knapp and Hartung adjustment (Knapp & Hartung, 2003) to control the Type I error rate (Viechtbauer et al., 2015) and the restricted maximum likelihood method (López-López et al., 2014).

Following previous meta-analyses, the interpretation of the magnitude of the overall effect size was based on the benchmark identified by Cohen (1988): 0.2 = small, 0.5 = medium, and 0.8 = large, although his standard has some limitations such as not considering methodological features (Cheung & Slavin, 2016; Lipsey et al., 2012). Considering these limitations of Cohen's (1988) criteria, the magnitude of effect size of individual studies was also evaluated based on the benchmarks identified by Cheung and Slavin (2016): 0.30 = average for studies with small sample size (<250) and 0.16 = average for studies with large sample size (≥250).

Sensitivity analysis was conducted by checking whether the study's confidence interval overlaps with that of the pooled effect size and calculating standard deviations ($z \leq -3$ or $z \geq 3.0$ are outliers). Publication bias was visualized by plotting the observed standardized mean differences against their standard errors and tested by funnel plot test (using sample size as a predictor of effect sizes; Macaskill et al., 2001), Begg's rank correlation test (using variance and sample size as a predictor of effect sizes; Begg & Mazumdar, 1994), trim-and-fill method (using L_0^+ as the number of unavailable effect sizes due to publication bias; Duval & Tweedie, 2000a, 2000b), and an adapted version of Egger's regression test (using sampling variance as a predictor of effect sizes; de Jong et al., 2021; Egger et al., 1997; Fernández-Castilla et al., 2021; Knapp et al., 2017; Sterne et al., 2011; Viechtbauer, 2017). The existence of publication bias is indicated by a statistically significant test and $L_0^+ > 3$ (see Fernández-Castilla et al., 2021). The adapted version of Egger's regression test models a quadratic relationship between the standard errors and the standardized mean differences and the intercept of this model is the overall effect size free of publication bias (see Stanley & Doucouliagos, 2014).

Following de Jong et al. (2021) and Knapp et al. (2017), a three-level mixed-effects model was run for moderator analysis. Two groups of moderators were analyzed to decrease the risk of making a Type I error or Type II error caused by testing moderators individually or simultaneously (Jansen et al., 2019; van Alten et al., 2019). Due to missing values, the activity level of control group, additional instruction, and assessment type were omitted from group analysis and analyzed individually.

3 | RESULTS

3.1 | Descriptive findings

This meta-analysis included 34 studies published from 2006 to 2020. Their characteristics are in Table 5. The sample sizes ranged from 40 to 470. Thirty studies used media comparisons, while only three made value-added comparisons. For learning outcomes, no study reported emotions, nine reported motivation, 33 reported cognition, and six reported both cognition and motivation, but only three reported both an immediate test and a delayed test (retention). For

educational level, all studies were implemented in secondary schools ($k = 21$) and universities ($k = 13$). Regarding country, one third was conducted in the United States ($k = 11$) and one in eight in China ($k = 4$). For chemistry content, the most common topics were nomenclature ($k = 8$), periodic table ($k = 4$), and organic chemistry ($k = 4$). Regarding assessment methods, tests and questionnaires were the most frequent measures for cognition and motivation; only five studies adopted mixed-method research (e.g., tests combined with interviews), among which one retrieved log data. Regarding game genre, the most used genres were puzzle ($k = 12$), simulation ($k = 7$), and role-playing games ($k = 6$). A detailed example of GBL activities for each game genre is available in Table S2.

3.2 | Research question 1: Media comparison

3.2.1 | Sensitivity analysis

One comparison by Okonkwo (2012)—GBL vs. conventional lecture—is an outlier based on its extremely large effect size ($g = 5.34$ and $g = 3.13$ for cognition and motivation) and 3 standard deviations larger than the mean ($z = 5.5$ for cognition and $z = 3.9$ for motivation). Furthermore, its 95% CI does not overlap with that of the summary effect. Thus, the study was excepted for further analysis.

3.2.2 | Distribution of effect sizes

Effect sizes of cognition and motivation for the individual studies and their distribution are presented in forest plots (Figures 3 and 4). No results were found for emotion. As displayed in Table 6, regarding cognitive outcomes, the effect sizes ($k = 30$, $\#ES = 57$) vary substantially across the studies, from −0.62 to 1.84. Among the 53 positive outcomes, 40 are statistically significant. Among the five large-scale studies (sample size ≥250), three reported an equal or above average effect size (≥0.16) and among the 52 small-scale studies (sample size <250), 46 reported above average effect size (≥0.3), according to Cheung and Slavin (2016). The mean effect size is statistically significant and medium ($g = 0.70$; 95% CI [0.51; 0.89]), according to Cohen (1988). Regarding retention, a similar effect ($g = 0.59$; 95% CI [0.35; 0.83]; $k = 20$, $\#ES = 31$) is found. Regarding motivation, the effect sizes ($k = 7$, $\#ES = 21$) vary from −0.09 to 1.18. Among the 19 positive outcomes, eight are statistically significant. Only two reported a negative but not statistically significant effect; the mean effect is statistically significant but small ($g = 0.35$; 95% CI [0.19; 0.50]). The model fit of this three-level model was statistically significantly better than the two-level model than the two-level model that does not consider within-study variance (cognition: $\chi^2 = 17.20$, $p < .001$; retention: $\chi^2 = 5.88$, $p = .01$; and motivation: $\chi^2 = 6.88$, $p = .009$).

3.2.3 | Heterogeneity

As displayed in Table 6, the statistically significant Q values reveal that effect sizes vary. For cognition, The I^2 reflects that sampling variance, within-study variance, and between-study variance can explain 14%, 33%, and 53% of the observed variance, respectively (Borenstein, 2019). The similar results are found for retention. For motivation, I^2 indicates that sampling variance,

TABLE 5 Characteristics of included studies in the meta-analysis ($n = 34$)

Study	*N*	Outcome	Comparison	Grade	Country	Game name	Game genre	Topic	Assessment method
Akkuzu and Uyulgan (2016)	62	Cognition	Media	Higher	Turkey	OrCheTaboo	Puzzle	Functional group	Test interview
Cahyana et al. (2017)	40	Cognition	Media	Secondary	Indonesia	na	na	Reaction rate	Test
Cha et al. (2017)	198	Cognition motivation	Media	Higher	Malaysia	Brainteaser	Puzzle	Organic chemistry	Test questionnaire
Chee and Tan (2012)	77	Cognition	Media	Secondary	Singapore	Legends of Alkhimia	Role-playing	Properties of substances	Test
Chen and Liao (2015)	76	Cognition	Dynamic-AR strategy	Secondary	China	Manufacturing iron man	Simulation	Chemical cell	Test
Chen et al. (2014)	105	Cognition motivation	Worked example	Secondary	China	The Alchemist's Fort	Role-playing adventure	Chemical reactions	Test questionnaire
Chimeno et al. (2006)	40	Cognition	Media	Higher	US	The rainbow wheel The rainbow matrix	Puzzle	Nomenclature	Test
da Silva Júnior et al. (2018)	246	Cognition	Media	Secondary	Brazil	Say my name	Puzzle	Organic nomenclature	Test
Daubenfeld and Zenker (2015)	46	Cognition	Media	Higher	Germany	na	Adventure	Equilibria	Test
Fatokun et al. (2016)	96	Cognition retention	Media	Secondary	Nigeria	Element Card I Atomic radius card Ionization card Group fixing SPD – game Sorting – out Transition element card Throw and answer	Puzzle	Periodicity (Periodic table)	Test
Gupta (2019)	67	Cognition Retention	Media	Higher	US	Molebots	Action	Nomenclature	Test questionnaire

(Continues)

TABLE 5 (Continued)

Study	N	Outcome	Comparison	Grade	Country	Game name	Game genre	Topic	Assessment method
Halpern et al. (2012)	136	Cognition	Media	Higher	US	Operation ARA	Role-playing	na	Test
Hodges et al. (2018)	351	Cognition	Media	Secondary	US	Blended reality environment	Simulation Role-playing	Redox reaction	Test Interview Data logs
Jagodziński and Wolski (2015)	200	Cognition Retention	Media	Secondary	Poland	Virtual chemical laboratory	Simulation	Inorganic acids	Test
Joag (2014)	104	Cognition	Media	Secondary	India	na	Puzzle	Periodic table	Test
Johnson-Glenberg et al. (2014)	51	Cognition	Media	Secondary	US	SMALLab	Simulation	Titration	Test
Kavak (2012)	49	Cognition	Media	Secondary	Turkey	ChemOkey	Puzzle	Nomenclature	Test
Lay and Osman (2018)	138	Cognition Motivation	Media	Secondary	Malaysia	MyKimDG	na	Precipitation reaction	Test Questionnaire
le Maire et al. (2018)	210	Cognition	Media	Higher	Belgium	Clash of Chemists	na	Stoichiometry	Test
Low (2010)	75	Cognition	Media	Secondary	Singapore	SynTactic©	Strategy	Organic synthesis	Test questionnaire Interview observation
Martin and Shen (2014)	70 61 68 69	Cognition	Media aesthetic Choice Competition	Higher	US	Element Solitaire	Puzzle	Periodic table	Test
Martinez-Hernandez (2010)	40	Cognition	Media	Higher	US	Electrolysis room Ammonia synthesis The hidden key Marble blocks Light sensor challenge Processing room	Adventure	State of matter Stoichiometry Chemical equilibrium Neutralization Redox reaction	Test questionnaire Interview
Meesuk and Srisawasdi (2014)	87	Motivation	Media	Secondary	Thailand	SAGOI The IE war	Action	Ionization	Questionnaire

TABLE 5 (Continued)

Study	*N*	Outcome	Comparison	Grade	Country	Game name	Game genre	Topic	Assessment method
Merchant et al. (2013)	382	Cognition	Media	Higher	US	Second Life® The molecule game Chemist as an artist The tower of VSEPR theory	Simulation	VSEPR theory	Test
Okonkwo (2012) 1 2	234 233	Cognition Motivation	Media	Secondary	Nigeria	Simulation-game	Role-playing	Pollution and waste management	Test Questionnaire
Rastegarpour and Marashi (2012)	105	Cognition	Media	Secondary	Iran	na	na	Nomenclature	Test
Renner (2014)	78	Cognition	Media	Secondary	US	na	Simulation	Alpha, beta, and gamma radiation	Test
Sousa Lima et al. (2019)	144	Cognition	Media	Secondary	Brazil	Chemical Nomenclature	Puzzle	Organic nomenclature	Test
Srisawasdi and Panjaburee (2019)	62	Cognition Motivation	Media	Secondary	Thailand	Factory Game	Role-playing	Properties of substances	Test Questionnaire
Stringfield and Kramer (2014)	120	Cognition Motivation	Media	Higher	US	Who wants an A in general chem	Puzzle	General organic biochemistry	Test Questionnaire
Su and Cheng (2019)	72	Cognition	Media	Secondary	China	Virtual chemical laboratory	Simulation	CO_2 gas collection	Test
Sugiyarto et al. (2018)	64	Cognition	Media	Secondary	Indonesia	Chemondro	Puzzle	Nomenclature	Test
Weng et al. (2015)	135	Cognition Motivation	Media	Secondary	China	na	Action	Periodic table	Test Questionnaire
Wood and Donnelly-Hermosillo (2019)	470	Cognition	Media	Higher	US	Topinomica	Puzzle	Nomenclature	Test Questionnaire Observation

Notes: *g*, effect size; *N*, total sample size; na, not available; 1, game versus concept mapping; 2, game versus conventional lecture.

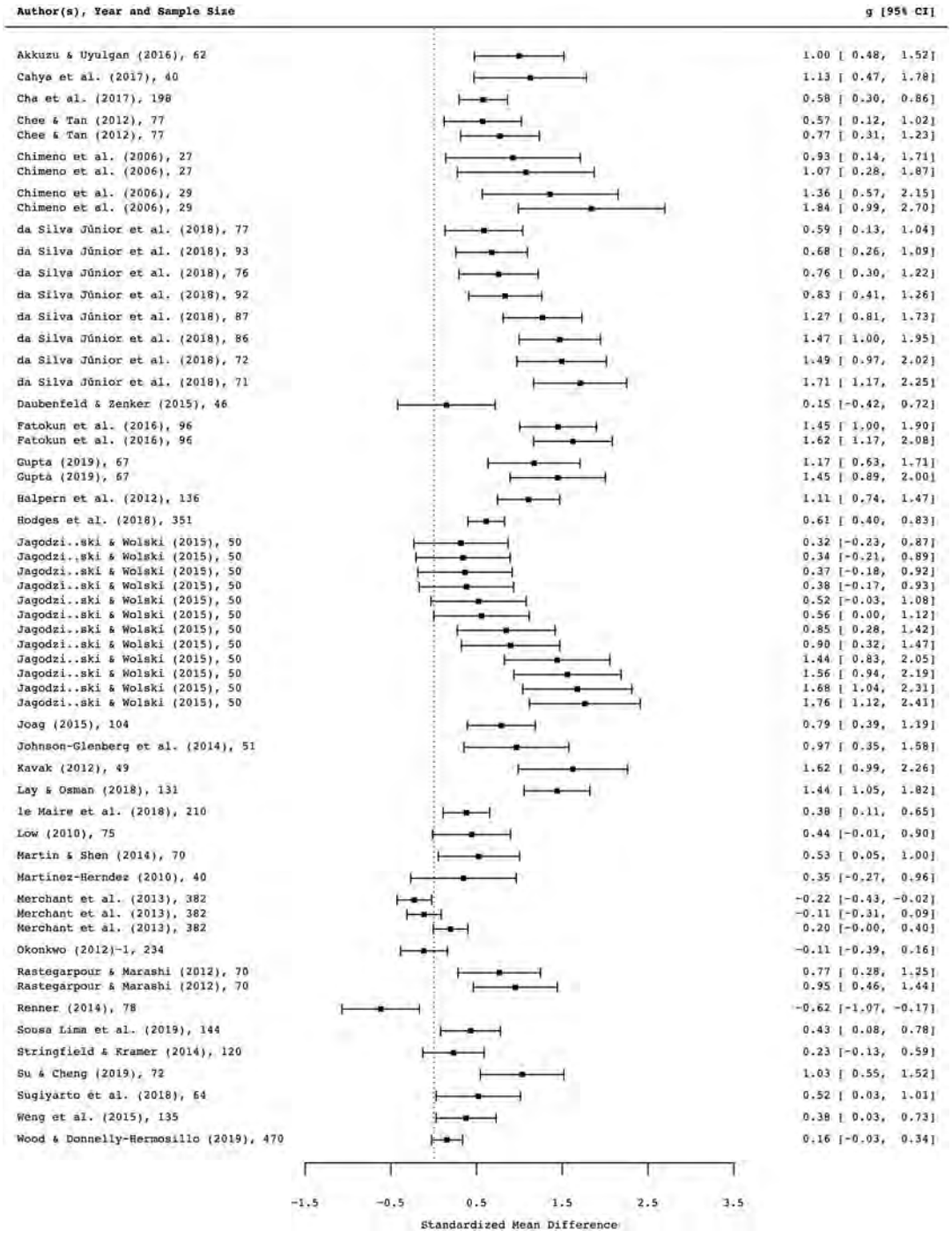

FIGURE 3 Forest plot for cognitive outcomes

within-study variance, and between-study variance could explain 39%, 49%, and 12% of the observed variance, respectively. Thus, moderator analysis is required to inspect sources for heterogeneity for cognition but not for motivation since it is not informative to analyze only seven studies.

3.2.4 | Publication bias

Similarly, the risk of publication bias was assessed for cognition (Banks et al., 2012; Sterne et al., 2011). In the funnel plot of Figure S1, an obvious asymmetry was found: the outlying

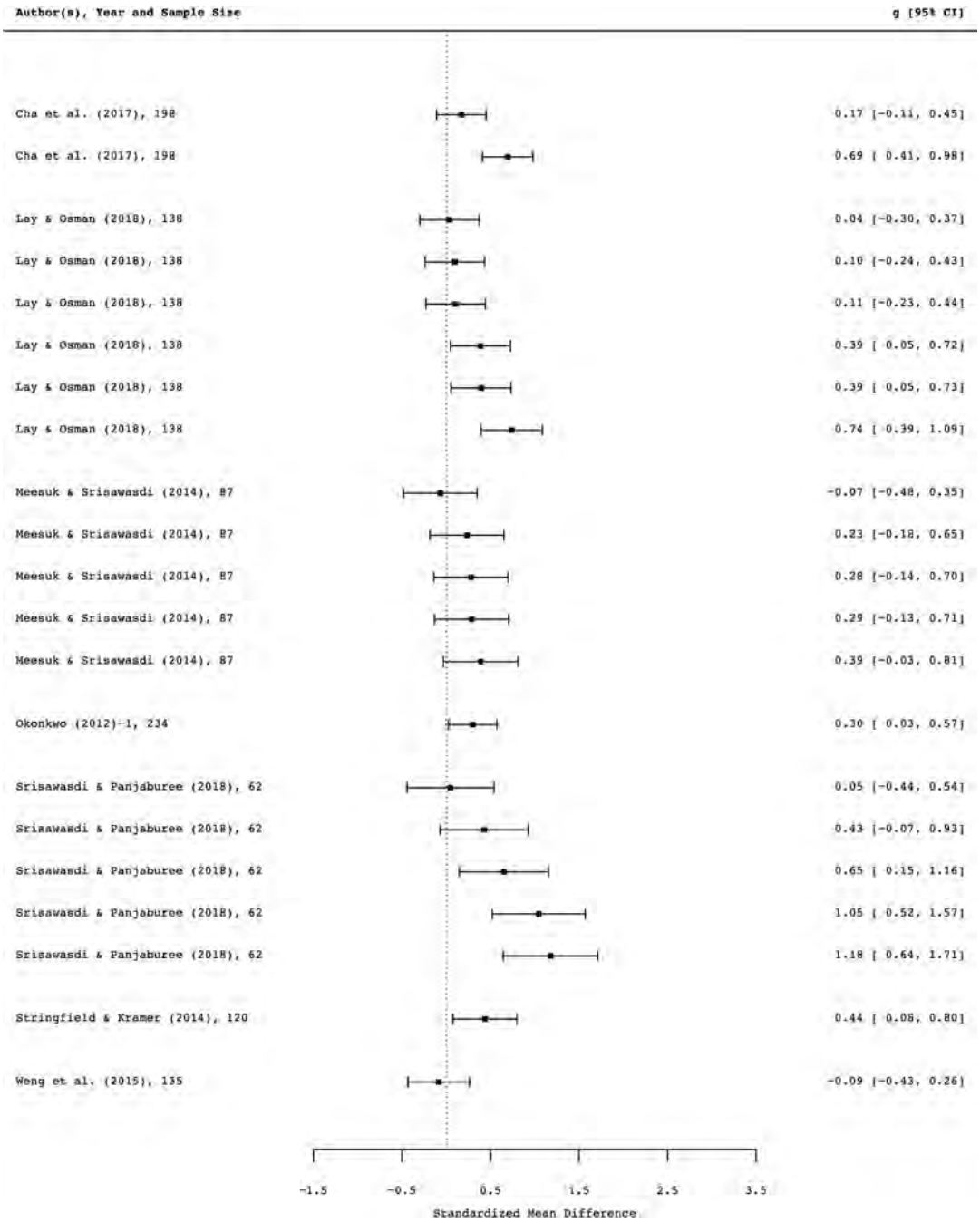

FIGURE 4 Forest plot for motivational outcomes

TABLE 6 Results of random-effects meta-analysis in media comparisons

Variable	*k* (*N*)	*#ES*	*g*	*SE*	95% CI	*Q*	τ^2_{level2}	τ^2_{level3}	I^2_{level2}	I^2_{level3}
Cognition	30 (4155)	57	0.70	0.10	[0.51; 0.89]	415.2*	0.10	0.16	33%	53%
Retention	20 (2860)	31	0.59	0.12	[0.35; 0.83]	240.4*	0.06	0.22	18%	70%
Motivation	7 (974)	21	0.35	0.08	[0.19; 0.50]	49.14*	0.05	0.01	49%	12%

Notes: CI, confidence interval; *#ES*, number of effect sizes; *g*, mean effect size; *k*, number of studies; *N*, total sample size; *Q*, heterogeneity value; *SE*, standard error; τ^2_{level2}, within-study variance; τ^2_{level3}, between-study variance; I^2_{level2}, within-study heterogeneity index (%); I^2_{level3}, between-study heterogeneity index (%).
*$p < .05$.

studies were distributed in the middle and upper parts, but studies in the bottom of the left side were missing (Borenstein et al., 2009). This was supported by funnel plot test ($p = .0017$), Begg's rank correlation test (using variance, $p = .0009$; using sample size, $p = .02$), trim-and-fill method ($L_0^+ = 6$), and the adapted Egger's regression test ($p < .0001$, Figure S2; Fernández-Castilla et al., 2021). The effect free of publication bias can be estimated by the intercept of this model (Knapp et al., 2017; Stanley & Doucouliagos, 2014), that is, 0.09 (95% CI [−0.22; 0.40]), which is not statistically significant and substantially smaller than the original estimate ($g = 0.70$; 95% CI [0.51; 0.89]). Furthermore, moderator analyses on publication source and sample size indicate a statistically significant larger effect in published studies than in gray literature and in small-scale studies than in large-scale studies (small-study effects; Banks et al., 2012; Borenstein et al., 2009; Sterne et al., 2011; Table 7). Overall, we conclude that small-study effects were likely to happen and probably caused by publication bias (Borenstein, 2019; Borenstein et al., 2009).

3.3 | Research question 2: Moderator analysis

As displayed in Table 7, at least one moderator in methodology characteristics exhibits a statistically significant relationship with the effect size ($p < .0001$). Specifically, sample size is negatively related to effect size (estimated coefficient = −0.003, $p < .0001$), and publication source is positively related to effect size (estimated coefficient = 0.62, $p = .001$), while accounting for all the other variables in the model, that is, small-scale studies or published studies are associated with larger effects. Aside from these two variables, no other moderator exhibits a statistically significant relationship with the effect size.

3.4 | Research question 3: Value-added comparison

As displayed in Table 8, only three studies reported six effect sizes for cognition, ranging from −0.11 to 0.46, but none reached statistical significance. The only statistically significant but very small effect is in the study that compares the effect of worked examples on motivation. A meta-analysis was impossible as the comparisons differed strongly, ranging from manipulations of aesthetic, choice, competition, worked example, guiding strategy, and type of AR.

TABLE 7 Results of moderator analysis for cognitive outcomes

Moderator	*k*	*#ES*	Estimate	*SE*	p^a	95% CI	p^b
Instruction characteristics							.76
Intercept	30	57	0.73	0.24	0.01	[0.09; 1.02]	
User grouping							
Multiple (reference)	8	17					
Single	22	40	0.04	0.20	0.86	[−0.36; 0.44]	
No. of game sessions							
Multiple (reference)	8	18					
Single	22	39	0.17	0.22	0.45	[−0.27; 0.60]	
Activity level of control group							.55
Intercept	24	49	0.60	0.17	0.0003	[0.27; 0.92]	
Active (reference)	10	16					
Passive	14	33	0.13	0.22	0.55	[−0.30; 0.57]	
Additional instruction							.38
Intercept	26	51	0.87	0.27	0.001	[0.33; 1.41]	
No (reference)	4	5					
Yes	22	46	−0.26	0.39	0.38	[−0.84; 0.32]	
Methodology characteristics							<.0001
Intercept	30	57	0.49	0.19	0.01	[0.11; 0.87]	
Sample size	30	57	−0.003	0.00	<0.0001	[−0.004; −0.001]	
Publication source							
Gray literature (reference)	6	7					
Published	24	50	0.62	0.19	0.001	[0.25; 1.00]	

(Continues)

TABLE 7 (Continued)

Moderator	*k*	*#ES*	Estimate	*SE*	p^a	95% CI	p^b
Randomization							
Quasi-experiment design (reference)	24	49					
Random controlled trial	6	8	0.14	0.18	0.43	[−0.22; 0.50]	
Assessment type							.63
Intercept	22	48	0.59	0.14	<0.0001	[0.31; 0.87]	
Closed (reference)	14	37					
Mix	3	3	0.25	0.37	0.51	[−0.48; 0.98]	
Non-closed	5	8	0.24	0.30	0.42	[−0.34; 0.82]	

Notes: CI, confidence interval; *#ES*, number of effect sizes; *k*, number of studies; *SE*, standard error; p^a tests. (1) The null hypothesis that the mean effect size in the reference group is zero (for the intercept row); and (2) the null hypothesis that the difference between that subgroup's mean effect and the reference group's mean effect is zero (for the individual subgroup row); p^b tests the null hypothesis that none of the moderators is related to effect size.

TABLE 8 Results of studies with vs. without specific features in value-added comparisons

Study name	Type of feature	Comparison	Outcome	*N*	*g*	*SE*
Martin and Shen (2014)	Game feature	Aesthetic vs. no aesthetic	Cognition	61	0.457	0.256
	Instructional feature	Choice vs. no choice	Cognition	68	0.143	0.240
	Instructional feature	Competition vs. no competition	Cognition	69	−0.112	0.239
Chen and Liao (2015)	Instructional feature	Procedure-guided vs. question-guided strategy	Cognition	76	0.437	0.230
	Game feature	Static-AR vs. dynamic-AR	Cognition	76	0.397	0.230
Chen et al. (2014)	Instructional feature	Worked example vs. no worked example	Cognition	105	0.257	0.195
			Motivation	105	0.144*	0.195

Notes: *g*, effect size; *N*, sample size; *SE*, standard error.
*$p < .05$.

4 | DISCUSSION

This meta-analysis indicates that GBL may address the unique characteristics of chemistry. Essential game design features such as interactivity, challenges, play, and feedback may address the challenges in chemistry education such as low performance, low level of motivation, and the occurrence of negative emotion. Our three-level random-effects model showed that overall, the effect of chemistry GBL on cognitive and motivational outcomes is larger than for non-GBL. Among instruction characteristics (activity level of control group, additional instruction, user grouping, and number of game sessions) and methodology characteristics (randomization, sample size, publication source, and assessment type), publication source and sample size moderate the effect. Evidence for emotional outcomes and game design features and instructional design features that improve chemistry GBL is insufficient.

4.1 | Media comparison

Our first goal was to examine whether chemistry GBL has a larger effect on cognitive, motivational, and emotional outcomes than non-GBL (RQ1). Most studies focused on cognitive outcomes, providing promising evidence that GBL enhances chemistry learning, but did not include motivational outcomes, providing moderate evidence that GBL motivates interest in chemistry. No evidence is available on whether GBL increases positive emotions or decreases negative emotions as no study reported emotional outcomes. Compared with previous GBL meta-analyses, this study is the first meta-analysis that uses a three-level random-effects model to consider the dependency of effect sizes within studies and the first that emphasizes emotion in GBL.

First, this study confirms chemistry GBL is more effective for cognition (Hypothesis 1) and retention (Hypothesis 2) than non-GBL. The mean effects for cognition ($g = 0.70$) and retention

($g = 0.59$) reveal a statistically significant medium ($g > .5$; Cohen, 1988). In other words, the score of the average person in chemistry GBL would be 0.6 SD above non-GBL, exceeding 73% of students in non-GBL (Coe, 2002). The effect for cognition is larger than most previous GBL meta-analyses across all subjects in general and math and science in particular, but equal or smaller than those in English (Table 1). This effect is also comparable with previous meta-analyses particular to chemistry with other educational interventions (cooperative learning: $g = 0.59$, Apugliese & Lewis, 2017; $g = 0.68$, Warfa, 2016; cooperative learning, collaborative learning, problem-based learning, process oriented guided inquiry learning, peer-lead team learning, and flipped instruction: $d = 0.62$, Rahman & Lewis, 2020). Most importantly, the effect for retention is larger than all previous meta-analyses. This implies that 0.59 could be a benchmark for a meaningful effect in chemistry education.

Three reasons could explain the differences in the magnitude of the overall effects between current and prior meta-analyses. One reason is that chemistry has a special relationship to GBL: GBL better align with the key characteristics of chemistry education (see Introduction) than other subjects. If that is the case, policymakers and practitioners should implement GBL in chemistry education. Second, technology development: more sophisticated technologies improve learning. Studies included in this meta-analysis, published from 2006 to 2020, are more recent than those from previous meta-analyses, ranging from 1990 to 2012 (Clark et al., 2016; Wouters et al., 2013). During the past decade, new technologies have emerged (Chen, Wang, et al., 2018). Our included studies applied many sophisticated technologies. For instance, based on voice recognition, eye movement, and brain wave analysis, Natural User Interface is used in gaming consoles (Jagodziński & Wolski, 2015); real-time data capture system is used in blended reality environment (Hodges et al., 2018); different types of automated tutoring based on student performances are combined with interactive dialogs with avatars (Halpern et al., 2012); VR simulates experiential learning (Su & Cheng, 2019); and MR benefits embodied learning (Johnson-Glenberg et al., 2014). These sophisticated technologies may better support chemistry GBL. Furthermore, students now have better access to technologies, leading to less difficulty playing chemistry games. Third, with the development of instructional design, current chemistry GBL may be better embedded in learning theories than older ones. More attention is paid on integrating game design and instructional design when designing chemistry GBL (e.g., Mayer, 2014b; NRC, 2011a; Plass et al., 2015) as most studies are from a later period. Nevertheless, chemistry GBL can enhance cognition, and the effect lasts over time.

This study also suggests chemistry GBL is more motivating than non-GBL (Hypothesis 3). Different from Wouters et al. (2013; $d = 0.26$, $p > .05$), a small but statistically significant effect ($g = 0.35$; $g > 0.2$; Cohen, 1988) for motivation was found. In other words, the motivation score of the average person in chemistry GBL would be 0.4 SD above non-GBL, exceeding 62% of students in non-GBL (Coe, 2002).

This finding seems to refute the critique that GBL may attract students, but higher motivation does not necessarily mean higher learning. Even though students report liking or having interest in the medium (the game), they tend to perceive that it provides an easier path to learning and invest less mental effort and time (Salomon, 1984), resulting in less learning compared with learning without the medium (Clark & Feldon, 2005, 2014). In our case, two included studies confirm this critique: students prefer GBL to study guides (Wood & Donnelly-Hermosillo, 2019) or traditional lectures (Stringfield & Kramer, 2014), but no difference in achievement was found. However, two other studies support our finding that GBL promotes both achievement and motivation to learn chemistry (Cha et al., 2017; Srisawasdi &

Panjaburee, 2019). Nevertheless, given that only seven studies reported motivation, this result should be interpreted with caution.

The cognitive and motivational benefits of chemistry GBL cannot prove a causal relationship between cognition and motivation. In the studies, only five reported both outcomes and their research methods, one-time pre-posttest design or posttest-only design, may not provide the required evidence. Instead, the cross-lagged panel model aims to detect causal or reciprocal relationships between variables, analyzing longitudinal data collected by testing or recording subjects at multiple points over time (Hamaker et al., 2015; Mulder & Hamaker, 2021; Selig & Little, 2012). In our included studies, no such method was used. Thus, whether cognition causes higher motivation in chemistry GBL and whether motivation causes higher cognition remains open questions.

4.2 | Moderator analysis

Our second goal was to examine the possible moderating effects of instruction and methodology characteristics, that is the conditions under which GBL is more effective relative to non-GBL (RQ2). We found some evidence that methodology characteristics moderate the effects, particularly sample size and publication source. Compared with previous GBL meta-analyses, this study uses more advanced methods to detect and correct publication bias and includes a continuous moderator (i.e., sample size).

Larger effects may be associated more with published studies than gray literature and with smaller studies than larger ones. The small-study effects, particularly publication bias, tent to exist. Researchers in chemistry education and GBL should attend to this issue, given that similar findings were also reported by previous meta-analyses particular to chemistry with other educational interventions (e.g., Rahman & Lewis, 2020; Warfa, 2016) and by meta-analyses in GBL (e.g., Lamb et al., 2018; Riopel et al., 2020; Sitzmann, 2011). However, more standardized methods with high statistical power are needed to assess and control how they impact main effects (e.g., the trim-and-fill method imputes adjusted effect size) and other aspects in multilevel meta-analyses (P. Cuijpers, personal communication, April 20, 2020). For instance, should we add sample size or publication source as covariate of the main effect? How and to what extent do small-study effects influence the moderator analysis?

Other moderators did not reveal statistically significant effects. Effect sizes of cognition were equal between non-GBL with active vs. passive instructions, GBL with vs. without additional instructions, GBL with single vs. multiple sessions (Hypothesis 5), GBL individually vs. in groups, RCTs versus QEDs, or with closed question vs. non-closed questions. Given the small number of studies under moderator categories, these results should never be interpreted as evidence that the effects are the same across subgroups or that there is no relation between the effects and included moderators (Borenstein et al., 2009). Instead, further studies are needed for more reliable evidence. Take randomization, for example, it is premature to conclude that larger effects are associated with RCTs than QEDs based on six RCTs versus 24 QEDs. Moreover, it is difficult to explain why we did not find statistically significant results for those moderators due to the limitations of all meta-analyses.

Other variables may help explain the potential sources of between-study variance. Unfortunately, the number of studies in total or under each subgroup was too low to conduct a moderator analysis. Instead, we performed an explanatory analysis based on findings from specific studies. One potential moderator is game genre. A specific game genre may suit specific

chemistry content (Wouters et al., 2013). For instance, puzzle games may help build factual knowledge (e.g., nomenclature, $g = 1.8$; Chimeno et al., 2006) through strengthening and weakening associations (reinforcement theory; Skinner, 1938); simulation games may help build conceptual knowledge (e.g., redox reaction, $g = 0.61$; Hodges et al., 2018) through constructing a schema of the cause-and-effect system (schema theory; Paas & Sweller, 2012); simulation games with MR or VR may help build procedural knowledge (e.g., titration, $g = 0.97$; Johnson-Glenberg et al., 2014) through deliberate practice with feedback (automaticity theory; Fitts & Posner, 1967; Mayer, 2014b). However, which genres suit which types of chemistry knowledge for which types of learners and under which contexts remains to be explored.

Another potential moderator is individual difference, such as gender (e.g., Steegh et al., 2021), prior knowledge (e.g., Lou & Jaeggi, 2019), and prior game experience. Among our included studies, compared with non-GBL, (1) girls outperformed boys but were not more motivated in chemistry GBL (Okonkwo, 2012), whereas others found no gender difference (Hodges et al., 2018; Merchant et al., 2013; Weng et al., 2015); (2) students with lower prior knowledge experienced greater learning gains from GBL than those with higher prior knowledge (Merchant et al., 2013; Wood & Donnelly-Hermosillo, 2019), but others found no difference (Sousa Lima et al., 2019); and (3) students with game experience achieved slightly higher learning gains than those without game experience (Merchant et al., 2013).

4.3 | Value-added comparison

Our third goal was to identify the more effective game design and instructional design features for chemistry GBL (RQ3). However, studies that used value-added comparisons of GBL with or without specific features ($k = 3$) are too few to perform a meta-analysis. This lack of studies confirms that the study of effective design features of GBL (value-added research) is often underestimated compared with media comparison research (Boyle et al., 2016; Clark et al., 2016; Young et al., 2012). In line with previous meta-analyses, more evidence from value-added research is required for researchers and practitioners.

First, value-added research may provide design guidelines for chemistry GBL, especially for practitioners such as game developers who create games for learning and teachers who implement GBL (Mayer, 2014b). GBL can be complex and require well-designed guidelines (e.g., Eastwood & Sadler, 2013). There are little evidence-informed guidelines for developers to integrate instructional design with game design features. Most game developers are familiar with game design but not instructional design. However, most teachers can only change the GBL environment by instructional design, not the game environment per se. Second, game researchers must first conduct value-added research to refine GBL environments before comparing GBL with non-GBL. Without optimizing GBL through value-added comparisons, it is unpromising to compare learning with poorly designed games versus other media (Plass et al., 2020).

According to one side of the Clark-Kozma "media-effects" debate, media comparison studies come with two challenges. Conceptually, research may confound media (games) with methods; it is not the medium but the method that causes learning (Clark, 1983, 1991, 1994a, 1994b, 2007; Clark et al., 2008; Kirschner & Hendrick, 2020; Mayer, 2014b). Methodologically, GBL and non-GBL groups may differ in dimensions (e.g., instructions, learning materials) other than the game, making it unclear what makes a difference in learning (Clark, 2007; Clark et al., 2008; Kirschner & Hendrick, 2020; Mayer, 2014b; NRC, 2011a). Therefore, it is

difficult to attribute learning effects to games, instructional methods, or other factors (e.g., Daubenfeld & Zenker, 2015).

One solution is to focus on value-added research within one game. The other side of the Clark-Kozma "media-effects" debate argues that it is unnecessary to separate instructional methods from games, as together they cause learning (Kozma, 1991, 1994a, 1994b; Parker et al., 2008). Instead of separating them, a good GBL design integrates instructional design with game design. Cognitive benefits are not the sole potential of chemistry GBL as games complement learning experiences with other aforementioned unique potentials (see Introduction). To employ these potentials, the focus should be less on media comparisons regarding learning effects, and more on improving GBL via value-added comparisons.

However, this does not mean media comparisons are meaningless and should be abandoned completely; they are still valuable, especially when testing GBL superiority claims (GBL is more effective than learning other media; Mayer, 2014b), justifying the reward, the effort, and cost of developing games for learning, and verifying whether certain instruction methods work specifically for GBL but not for non-GBL. Furthermore, with media comparison research, another solution is to equate GBL and non-GBL in all variables except for the game (Mayer, 2014b). Before that, however, we need high-quality GBL. Again, value-added research comes into play. Since value-added research and media comparison research serve different functions, researchers must first conduct value-added comparisons to create a well-designed GBL and then, if necessary, conduct media comparisons using a rigorous experimental design.

4.4 | Limitations

The following concerns may affect the study findings. First, some studies fail to report background information. For example, because six studies reported limited or no information about additional instructions and activity levels of control groups (e.g., Fatokun et al., 2016; Rastegarpour & Marashi, 2012), their moderating effect is unknown. Although the considered moderators captured part of high heterogeneity, there is clearly unexplained variance. Missing information prevents us from including other potential moderators, such as game experience, educational research experience, or duration of intervention (Wouters et al., 2013; Wouters & van Oostendorp, 2013). Missing information affects the selection, coding, and/or analysis of moderators. Furthermore, missing information might affect our assessment regarding study quality. For instance, GBL adopts debriefing, whereas this information is missing in non-GBL, or two groups may use different ways to present the learning content. Thus, research may be contaminated (Kirschner & Hendrick, 2020) as there are more differences between the comparison groups other than just the game (Clark, 1983). Again, we cannot include or control this influence because of the missing information.

Unfortunately, we had to exclude many studies because essential information was missing. Out of 842 screened articles, only 34 met our criteria, indicating that many chemistry GBL have been developed but are not well-reported and/or well-tested. The increasing popularity of games stimulated a flood of publication (Hwang & Wu, 2012; Tobias et al., 2011), but most of the excluded studies only describe the game without testing its effectiveness on learning outcomes (Tobias & Fletcher, 2012); report students' subjective opinions, satisfaction, or conceptions of the game (i.e., the usability test) without an objective assessment; or measure learning outcomes without a control group. Although post hoc power analysis is not recommended, it indicates we have sufficient power for cognition, retention, and motivation benefits of chemistry GBL.

Moreover, our broad definition of cognitive and motivational outcomes may bias the main effects. Studies vary regarding outcome measures (Cooper, 2015), and the limited number of studies made it impossible to categorize them further into different constructs (e.g., interest, self-efficacy). As all studies are different, the focus is less on sameness and more on difference: what makes effect sizes varied (Hattie, 2013). In this sense, for motivation, the meta-analysis indicates the effects of the interventions on motivation did not vary across type of motivation (Lazowski & Hulleman, 2016). For cognition, previous meta-analyses on GBL also imply the effects did not vary across type of cognitive outcomes: knowledge vs. skills (Wouters et al., 2013), declarative vs. procedural knowledge (Riopel et al., 2020; Sitzmann, 2011), or intrapersonal vs. cognitive learning outcomes (Clark et al., 2016; see Table 1).

4.5 | Implications

We make the following recommendations regarding the practices and theories of chemistry GBL. For practitioners, the positive effect of chemistry GBL on cognition, retention, and motivation suggests implementing GBL in chemistry education. The overall effect size provides the benchmark of chemistry GBL interventions for further research, and the distribution of effect sizes may help researchers anticipate the effects before their study.

For researchers, we agree with previous meta-analyses that it is time to move beyond "whether or not chemistry GBL works" (media comparison) to "what works for chemistry GBL and why it enhances learning and others do not" (value-added comparison; Chen et al., 2020; Young et al., 2012) and conduct more research to provide design guidelines for implementing chemistry GBL. This meta-analysis suggests that GBL may address the unique characteristics of chemistry. To confirm this, more GBL meta-analyses on subjects other than chemistry are needed.

Regarding learning outcomes, more considerations are needed: (1) for cognitive outcomes, more delayed tests measuring retention (Mayer, 2014b; Wood & Donnelly-Hermosillo, 2019) to avoid the novelty effect (Clark, 1983); (2) more research in motivation to learn chemistry, which could be a common but questionable appeal of chemistry GBL (Clark & Feldon, 2005, 2014); (3) more research into emotions (e.g., Raker et al., 2019) since the most desirable instruction is that learners learn most from what they enjoy most (Clark, 1982); (4) further research regarding which game genre is best for which type of learning outcome; and (5) more research on the relationships between cognition, motivation, and emotion (e.g., Gibbons & Raker, 2019) in GBL.

Regarding methodology, more high-quality intervention research in chemistry GBL is required to identify what works, for whom, and under which conditions. Considering the small-study effects, we suggest researchers to use power analysis (e.g., G*Power; Faul et al., 2007) to estimate the minimum number of participants needed (Ellis, 2010) when planning primary studies as statistical power and effect size depend on sample size (Simpson, 2017). Considering the contextual factors, mixed methods are promising to evaluate the effectiveness of chemistry GBL; tests or questionnaires should be combined with observations, interviews, and/or log data (e.g., Hodges et al., 2018; Wood & Donnelly-Hermosillo, 2019). Regarding assessments, all assessments of the included studies were taken postintervention using separate tests, such as self-reports on motivation after gameplay when motivation might decrease (Wouters et al., 2013). We advocate more embedded tests (e.g., stealth assessment for adaptivity; Shute et al., 2017) or real-time overt assessments (e.g., eye tracking, physiological monitoring, heart rate, blood pressure; Mayer, 2020; Wouters et al., 2013) focusing on learning processes.

On the theoretical aspect, more evidence is needed regarding how people learn chemistry with games (learning mechanics) and how to support chemistry GBL (instructional support). First, we lack studies on how chemistry games affect learning processes and outcomes; how they affect motivation or emotions; what the roles of motivation and emotions are; and how motivation, cognition, and emotions interact with each other. Second, we lack studies on how people learn better with chemistry games. As GBL is part of multimedia learning, game researchers can refer to multimedia principles from CTML (Mayer, 2014a). In certain subjects, some principles enhance GBL (e.g., modality), whereas some do not (e.g., redundancy; Mayer, 2020). Future studies should explore the learning effects of these multimedia principles in chemistry GBL.

5 | CONCLUSION

This meta-analysis suggests that GBL may address the unique characteristics of a single subject. For example, essential game design features such as interactivity, challenges, play, and feedback may address the challenges in chemistry education such as low performance, motivation, and emotion. More GBL meta-analyses on subjects other than chemistry are needed. We systematically reviewed 34 studies on the cognitive, motivational, and emotional effects of GBL in chemistry. Compared with previous GBL meta-analyses, this study is the first meta-analysis that uses a three-level random-effects model to consider the dependency of effect sizes within studies. Generally, we found chemistry GBL is more effective not only for cognition and retention but also motivation than non-GBL. Publication source and sample size possibly moderate this effect. The substantial heterogeneity between studies underscores how chemistry GBL is implemented, particularly sample size and publication source. This study used more advanced methods to detect and correct publication bias and is the first GBL meta-analysis that includes sample size as a continuous moderator. We found that there may be the small-study effects, particularly publication bias. Furthermore, this study is also the first meta-analysis that emphasizes emotions in GBL. Unfortunately, studies assessing learner's emotions in chemistry GBL are absent. More robust research is required to provide a clear understanding of their true effects. Similarly, more value-added research is needed to identify more effective game design features and instructional design features and provide design guidelines for chemistry GBL. We advocate conducting well-developed value-added research to optimize GBL before comparing it with non-GBL.

In closing, GBL has good chemistry with chemistry education in media comparison research—chemistry GBL holds the right formula for improved learning and motivation; they need more value-added research before getting married; and design is the key in this relationship.

ACKNOWLEDGMENTS

We would like to thank Caspar van Lissa, Wolfgang Viechtbauer, David van Alten, Renée Jansen, Guido Schwarzer, Michael Borenstein, Mathias Harrer, Pim Cuijpers for sharing their statistical expertise; Judith M. Conijn, Belén Fernández-Castilla, Cheryl Bodnar, Kim Chwee Daniel Tan, Kian Seh Low, Kaoru Sumi, Kermin Joel Martínez-Hernández, Jason M. Harley, José Nunes da Silva Júnior, Mike Coffey, and Su Cai for providing additional information on their studies.

ORCID

Yuanyuan Hu https://orcid.org/0000-0002-2314-546X

REFERENCES

References marked with an asterisk indicate studies included in the meta-analysis.

Acquah, E. O., & Katz, H. T. (2020). Digital game-based L2 learning outcomes for primary through high-school students: A systematic literature review. *Computers & Education*, *143*, 103667. https://doi.org/10.1016/j.compedu.2019.103667

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

Ainsworth, S. E. (2014). The multiple representations principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 464–486). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.024

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1922/CDH_2120VandenBroucke08

*Akkuzu, N., & Uyulgan, M. A. (2016). How to improve students' comprehension concerning the major terms of functional groups?—In the experiment of OrCheTaboo game. *International Journal of Higher Education*, *5*(2), 196–212. https://doi.org/10.5430/ijhe.v5n2p196

American Chemical Society [ACS]. (2015a). *ACS guidelines for Bachelor's degree programs*. The American Chemical Society.

American Chemical Society [ACS]. (2015b). *ACS guidelines for chemistry in two-year college programs*. The American Chemical Society.

American Chemical Society [ACS]. (2018). *ACS guidelines and recommendations for teaching middle and high school chemistry*. The American Chemical Society.

Apugliese, A., & Lewis, S. E. (2017). Impact of instructional decisions on the effectiveness of cooperative learning in chemistry through meta-analysis. *Chemistry Education Research and Practice*, *18*, 271–278. https://doi.org/10.1039/C6RP00195E

Artino, A. R., & Jones, K. D. (2012). Exploring the complex relations between achievement emotions and self-regulated learning behaviors in online learning. *The Internet and Higher Education*, *15*(3), 170–175. https://doi.org/10.1016/j.iheduc.2012.01.006

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis*, *34*(3), 259–277. https://doi.org/10.3102/0162373712446144

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. https://doi.org/10.2307/2533446

Bellou, I., Papachristos, N. M., & Mikropoulos, T. A. (2018). Digital learning technologies in chemistry education: A review. In D. Sampson, D. Ifenthaler, J. M. Spector, & P. Isaías (Eds.), *Digital technologies: Sustainable innovations for improving teaching and learning* (pp. 57–80). Springer International Publishing.

Bennett, A. G., & Rebello, N. S. (2012). Retention and learning. In *Encyclopedia of the sciences of learning* (pp. 2856–2859). Springer. https://doi.org/10.1007/978-1-4419-1428-6_664

Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat, Incorporated.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive meta-analysis version 3*. Biostat, Incorporated.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education*, *94*, 178–192. https://doi.org/10.1016/j.compedu.2015.11.003

Brown, T. L., Eugene LeMay, H., Bursten, B. E., Murphy, C. J., Woodward, P. M., Stoltzfus, M. W., & Lufaso, M. W. (2018). *Chemistry: The central science* (14th ed.). Pearson Education.

Byun, J., & Joung, E. (2018). Digital game-based learning for K–12 mathematics education: A meta-analysis. *School Science and Mathematics*, *118*(3–4), 113–126. https://doi.org/10.1111/ssm.12271

*Cahyana, U., Paristiowati, M., Savitri, D. A., & Hasyrin, S. N. (2017). Developing and application of mobile game based learning (M-GBL) for high school students performance in chemistry. *EURASIA Journal of Mathematics, Science and Technology Education*, *13*(10), 7037–7047. https://doi.org/10.12973/ejmste/78728

*Cha, J., Kan, S. Y., Wahab, N. H. A., Aziz, A. N., & Chia, P. W. (2017). Incorporation of brainteaser game in basic organic chemistry course to enhance students' attitude and academic achievement. *Journal of the Korean Chemical Society*, *61*(4), 218–222. https://doi.org/10.5012/jkcs.2017.61.4.218

Checa, D., & Bustillo, A. (2019). A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications*, *79*, 1–27. https://doi.org/10.1007/s11042-019-08348-9

Chee, Y. S., & Tan, D. K. C. (2012). Becoming chemists through game-based inquiry learning: The case of legends of Alkhimia. *Electronic Journal of e-Learning*, *10*(2), 185–198.

Chen, C.-H., Shih, C.-C., & Law, V. (2020). The effects of competition in digital game-based learning (DGBL): A meta-analysis. *Educational Technology Research and Development*, *68*(4), 1855–1873. https://doi.org/10.1007/s11423-020-09794-1

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

Chen, J., Wang, M., Kirschner, P. A., & Tsai, C.-C. (2018). The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis. *Review of Educational Research*, *88*(6), 799–843. https://doi.org/10.3102/0034654318791584

*Chen, M.-P., & Liao, B.-C. (2015). Augmented reality laboratory for high school electrochemistry course. In *2015 IEEE 15th International Conference on Advanced Learning Technologies* (pp. 132–136). IEEE. https://doi.org/10.1109/ICALT.2015.105

*Chen, M. P., Wong, Y. T., & Wang, L. C. (2014). Effects of type of exploratory strategy and prior knowledge on middle school students' learning of chemical formulas from a 3D role-playing game. *Educational Technology Research and Development*, *62*(2), 163–185. https://doi.org/10.1007/s11423-013-9324-3

Cheng, M. T., Chen, J. H., & Chu, S. J. (2015). The use of serious games in science education: a review of selected empirical research from 2002 to 2013. *Journal of Computer Education*, *2*, 353–375. https://doi.org/10.1007/s40692-015-0039-9

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods*, *19*(2), 211. https://doi.org/10.1037/a0032968

Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, *29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

*Chimeno, J. S., Wulfsberg, G. P., Sanger, M. J., & Melton, T. J. (2006). The rainbow wheel and rainbow matrix: Two effective tools for learning ionic nomenclature. *Journal of Chemical Education*, *83*(4), 651. https://doi.org/10.1021/ed083p651

Chiu, M. H., & Wu, H. K. (2009). The roles of multimedia in the teaching and learning of the triplet relationship in chemistry. In J. K. Gilbert & D. Treagust (Eds.), *Multiple representations in chemical education* (pp. 251–283). Springer.

Clark, D., Nelson, B., Sengupta, P., & D'Angelo, C. (2009). Rethinking science learning through digital games and simulations: Genres, examples, and evidence. In *Proceedings of The National Academies Board on Science Education Workshop on Learning Science: Computer Games, Simulations, and Education* (pp. 36–41). National Academy of Sciences.

Clark, D., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86*(1), 79–122. https://doi.org/10.3102/0034654315582065

Clark, R. E. (1982). Antagonism between achievement and enjoyment in ATI studies. *Educational Psychologist*, *17*(2), 92–101. https://doi.org/10.1080/00461528209529247

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, *53*(4), 445–459. https://doi.org/10.3102/00346543053004445

Clark, R. E. (1991). When researchers swim upstream: Reflections on an unpopular argument about learning from media. *Educational Technology*, *31*(2), 34–40.

Clark, R. E. (1994a). Media and method. *Educational Technology Research and Development*, *42*(3), 7–10. https://doi.org/10.1007/BF02298090

Clark, R. E. (1994b). Media will never influence learning. *Educational Technology Research and Development*, *42*(2), 21–29.

Clark, R. E. (2007). Learning from serious games? Arguments, evidence, and research suggestions. *Educational Technology*, *May-June*(3), 56–59 https://www.jstor.org/stable/44429512

Clark, R. E., Bissell, A., Burrus, C. S., & Breck, J. (2008). What evidence would change your mind about the learning benefits of serious games? A reply to Parker, Becker, and Sawyer. *Educational Technology*, *48*(3), 56–58 https://www.jstor.org/stable/44429581

Clark, R. E., & Feldon, D. F. (2005). Five common but questionable principles of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (1st ed., pp. 97–116). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.007

Clark, R. E., & Feldon, D. F. (2014). Ten common but questionable principles of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 151–173). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.009

Coe, R. (2002). It's the effect size, stupid—What effect size is and why it is important. In *Annual Conference of the British Education Research Association*. University of Exeter. http://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, *59*(2), 661–686. https://doi.org/10.1016/j.compedu.2012.03.004

Cook, D. A., & Artino, A. R. (2016). Motivation to learn: An overview of contemporary theories. *Medical Education*, *50*(10), 997–1014. https://doi.org/10.1111/medu.13074

Cooper, H. (2015). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Sage https://us.sagepub.com/en-us/nam/research-synthesis-and-meta-analysis/book241775

Cooper, M. M., & Stowe, R. L. (2018). Chemistry education research—From personal empiricism to evidence, theory, and informed practice. *Chemical Reviews*, *118*(12), 6053–6087. https://doi.org/10.1021/acs.chemrev.8b00020

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety: The experience of work and play in games*. Jossey Bass.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper and Row.

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, *19*(2), 272–309. https://doi.org/10.1080/15213269.2015.1015740

*da Silva Júnior, J. N., Nobre, D. J., do Nascimento, R. S., Torres, G. S., Leite, A. J. M., Monteiro, A. J., Alexandre, F. S. O., Rodríguez, M. T., & Rojo, M. J. (2018). Interactive computer game that egages students in reviewing organic compound nomenclature. *Journal of Chemical Education*, *95*(5), 899–902. https://doi.org/10.1021/acs.jchemed.7b00793

*Daubenfeld, T., & Zenker, D. (2015). A game-based approach to an entire physical chemistry course. *Journal of Chemical Education*, *92*(2), 269–277. https://doi.org/10.1021/ed5001697

de Brabander, C. J., & Martens, R. L. (2014). Towards a unified theory of task-specific motivation. *Educational Research Review*, *11*, 27–44. https://doi.org/10.1016/j.edurev.2013.11.001

de Jong, O., & Taber, K. S. (2007). Teaching and learning the many faces of chemistry. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 631–652). Routledge.

de Jong, K., Conijn, J. M., Gallagher, R. A., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, 102002. https://doi.org/10.1016/j.cpr.2021.102002

Deci, R. M., & Ryan, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78 10.1037110003-066X.55.1.68.

Di Natale, A. F., Repetto, C., Riva, G., & Villani, D. (2020). Immersive virtual reality in K-12 and higher education: A 10-year systematic review of empirical research. *British Journal of Educational Technology*, *51*(6), 2006–2033. https://doi.org/10.1111/bjet.13030

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics-Physics Education Research*, *10*(2), 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Duit, R., Niedderer, H., & Schecker, H. (2007). Teaching physics. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 599–629). Routledge.

Duval, S., & Tweedie, R. (2000a). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Duval, S., & Tweedie, R. (2000b). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98. https://doi.org/10.1080/01621459.2000.10473905

Eastwood, J. L., & Sadler, T. D. (2013). Teachers' implementation of a game-based biotechnology curriculum. *Computers & Education*, *66*, 11–24. https://doi.org/10.1016/j.compedu.2013.02.003

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology*, *103*(3), 632–648. https://doi.org/10.1037/a0023952

Ellis, P. D. (2010). *The essential guide to effect sizes* (Vol. 136). Cambridge University Press. https://doi.org/10.1017/CBO9780511761676

European Commission. (2015). *Science education for responsible citizenship*. Publications Office of the EU https://op.europa.eu/en/publication-detail/-/publication/a1d14fa0-8dbe-11e5-b8b7-01aa75ed71a1

*Fatokun, K. V. F., Egya, S. O., & Uzoechi, B. C. (2016). Effect of game instructional approach on chemistry students' achievement and retention in periodicity. *European Journal of Research and Reflection in Educational Sciences*, *4*(7), 29–40. www.idpublications.org

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *Journal of Experimental Education.*, *89*(1), 125–144. https://doi.org/10.1080/00220973.2019.1582470

Fiedler, K. & Beier, S. (2014). Affect and cognitive processes in educational contexts. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), (pp. 36–55). International handbook of emotions in education. Routledge. https://doi.org/10.4324/9780203148211.ch3

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.

Forsthuber, B., Motiejunaite, A., de Almeida Coutinho, A. S., Baïdak, N., & Horvath, A. (2011). *Science education in Europe: National policies, practices and research*. Education, Audiovisual and Culture Executive Agency, European Commission.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation and Gaming*, *33*(4), 441–467. https://doi.org/10.1177/1046878102238607

Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review*, *27*, 244–260. https://doi.org/10.1016/j.edurev.2019.04.001

Garzón, J., Pavón, J., & Baldiris, S. (2019). Systematic review and meta-analysis of augmented reality in educational settings. *Virtual Reality*, *23*(4), 447–459. https://doi.org/10.1007/s10055-019-00379-9

Gibbons, R. E., & Raker, J. R. (2019). Self-beliefs in organic chemistry: Evaluation of a reciprocal causation, cross-lagged model. *Journal of Research in Science Teaching*, *56*(5), 598–618. https://doi.org/10.1002/tea.21515

Gilbert, J. K., & Treagust, D. (2009). *Multiple representations in chemical education*. Springer. https://doi.org/10.1007/978-1-4020-8872-8

Girard, C., Ecalle, J., & Magnan, A. (2013). Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, *29*(3), 207–219. https://doi.org/10.1111/j.1365-2729.2012.00489.x

*Gupta, T. (2019). Game-based learning in chemistry: A game for chemical nomenclature. In *Technology integration in chemistry education and research (TICER)* (pp. 65–79). American Chemical Society. https://doi.org/10.1021/bk-2019-1318.ch005

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google scholar in evidence reviews and its applicability to grey literature searching. *PLoS One*, *10*(9), e0138237. https://doi.org/10.1371/journal.pone.0138237

*Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity*, *7*(2), 93–100. https://doi.org/10.1016/j.tsc.2012.03.006

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Harrer, M., Cuijpers, P., Furukawa, T. . A., & Ebert, D. D. (2019). *Doing meta-analysis in R: A hands-on guide*. Chapman and Hall/CRC. https://doi.org/10.5281/zenodo.2551803

Hattie, J. (2013). *Visible learning and the science of how we learn*. Routledge. https://doi.org/10.4324/9781315885025

Hwang, G.-J., & Wu, P.-H. (2012). Advancements and trends in digital game-based learning research: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology*, *43*(1), E6–E10. https://doi.org/10.1111/j.1467-8535.2011.01242.x

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hidi, S., & Renninger, K. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

*Hodges, G. W., Wang, L., Lee, J., Cohen, A., & Jang, Y. (2018). An exploratory study of blending the virtual world and the laboratory experience in secondary chemistry classrooms. *Computers & Education*, *122*, 179–193. https://doi.org/10.1016/j.compedu.2018.03.003

Homer, B. D., Raffaele, C., & Henderson, H. (2020). Games as playful learning: Implications of developmental theory for game-based learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 25–52). MIT Press.

Hung, H. T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C. C. (2018). A scoping review of research on digital game-based language learning. *Computers & Education*, *126*, 89–104. https://doi.org/10.1016/j.compedu.2018.07.001

*Jagodziński, P., & Wolski, R. (2015). Assessment of application technology of natural user interfaces in the creation of a virtual chemical laboratory. *Journal of Science Education and Technology*, *24*(1), 16–28. https://doi.org/10.1007/s10956-014-9517-5

Jaber, L. Z., & Hammer, D. (2016). Learning to feel like a scientist. *Science Education*, *100*(2), 189–220. https://doi.org/10.1002/sce.21202

Jansen, R. S., van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (2019). Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. *Educational Research Review*, *28*, 100292. https://doi.org/10.1016/j.edurev.2019.100292

*Joag, S. D. (2014). An effective method of introducing the periodic table as a crossword puzzle at the high school level. *Journal of Chemical Education*, *91*(6), 864–867. https://doi.org/10.1021/ed400091w

Johnson, C. I., Bailey, S. K., & Van Buskirk, W. L. (2017). Designing effective feedback messages in serious games and simulations: A research review. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 119–140). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_7

*Johnson-Glenberg, M. C., Birchfield, D. A., Tolentino, L., & Koziupa, T. (2014). Collaborative embodied learning in mixed reality motion-capture environments: Two science studies. *Journal of Educational Psychology*, *106*(1), 86–104. https://doi.org/10.1037/a0034008

Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning*, *7*, 75–83.

Johnstone, A. H. (2000). Teaching of chemistry-logical or psychological? *Chemistry Education Research and Practice*, *1*(1), 9–15.

Karakoç, B., Eryılmaz, K., Özpolat, E. T., & Yıldırım, İ. (2020). The effect of game-based learning on student achievement: A meta-analysis study. *Technology, Knowledge and Learning*, *1–16*, 207–222. https://doi.org/10.1007/s10758-020-09471-5

*Kavak, N. (2012). ChemOkey: A game to reinforce nomenclature. *Journal of Chemical Education*, *89*(8), 1047–1049. https://doi.org/10.1021/ed3000556

Ke, F. (2016). Designing and integrating purposeful learning in game play: A systematic review. *Educational Technology Research and Development*, *64*(2), 219–244. https://doi.org/10.1007/s11423-015-9418-1

King, D., Ritchie, S. M., Sandhu, M., Henderson, S., & Boland, B. (2017). Temporality of emotion: Antecedent and successive variants of frustration when learning chemistry. *Science Education*, *101*(4), 639–672. https://doi.org/10.1002/sce.21277

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review*, *21*(1), 31–42. https://doi.org/10.1007/s10648-008-9095-2

Kirschner, F., Paas, F., & Kirschner, P. A. (2011). Task complexity as a driver for collaborative learning efficiency: The collective working-memory effect. *Applied Cognitive Psychology*, *25*(4), 615–624. https://doi.org/10.1002/acp.1730

Kirschner, P. A., & Hendrick, C. (2020). *How learning happens* (1st ed.). Routledge.

Klopfer, E., & Thompson, M. (2020). Game-based learning in science, technology, engineering, and mathematics. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 387–408). MIT Press.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Knapp, F., Viechtbauer, W., Leonhart, R., Nitschke, K., & Kaller, C. P. (2017). Planning performance in schizophrenia patients: A meta-analysis of the influence of task difficulty and clinical and sociodemographic variables. *Psychological Medicine*, *47*(11), 2002–2016. https://doi.org/10.1017/S0033291717000459

Kozma, R. B. (1991). Learning with media. *Review of Educational Research*, *61*(2), 179–211.

Kozma, R. B. (1994a). A reply: Media and methods. *Educational Technology Research and Development*, *42*(3), 11–14. https://doi.org/10.1007/BF02298091

Kozma, R. B. (1994b). Will media influence learning? Reframing the debate. *Educational Technology Research and Development*, *42*(2), 7–19. https://doi.org/10.1007/BF02299087

Kromrey, J., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement*, *66*(3), 357–373. https://doi.org/10.1177/0013164405278585

Laffey, J. M., Sadler, T. D., Goggins, S. P., Griffin, J., & Babiuch, R. N. (2019). Mission HydroSci: Distance learning through game-based 3D virtual learning environments. In *Virtual Reality in education: Breakthroughs in research and practice* (pp. 623–643). IGI Global.

Lamb, R. L., Annetta, L., Firestone, J., & Etopio, E. (2018). A meta-analysis with examination of moderators of student cognition, affect, and learning outcomes while using serious educational games, serious games, and simulations. *Computers in Human Behavior*, *80*, 158–167. https://doi.org/10.1016/j.chb.2017.10.040

*Lay, A.-N., & Osman, K. (2018). Developing 21st century chemistry learning through designing digital games. *Journal of Education in Science, Environment and Health (JESEH)*, *4*(1), 81–92. https://doi.org/10.21891/jeseh.387499

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, *86*(2), 602–640. https://doi.org/10.3102/0034654315617832

*le Maire, N. V., Verpoorten, D. P., Fauconnier, M.-L. S., & Colaux-Castillo, C. G. (2018). Clash of chemists: A gamified blog to master the concept of limiting reagent stoichiometry. *Journal of Chemical Education*, *95*(3), 410–415. https://doi.org/10.1021/acs.jchemed.7b00256

Lee, K. M. (2004). Presence, explicated. *Communication Theory*, *14*(1), 27–50. https://doi.org/10.1111/j.1468-2885.2004.tb00302.x

Li, M. C., & Tsai, C. C. (2013). Game-based learning in science education: A review of relevant research. *Journal of Science Education and Technology*, *22*(6), 877–898. https://doi.org/10.1007/s10956-013-9436-x

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research (NSCER 2013–3000).

Loderer, K., Pekrun, R., & Lester, J. C. (2018). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction*, *70*, 101162. https://doi.org/10.1016/j.learninstruc.2018.08.002

Loderer, K., Pekrun, R., & Plass, J. L. (2020). Emotional foundations of game-based learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 111–151). MIT Press.

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 30–48. https://doi.org/10.1111/bmsp.12002

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. (2018). Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods*, *9*(3), 336–351. https://doi.org/10.1002/jrsm.1310

Lou, A. J., & Jaeggi, S. M. (2020). Reducing the prior-knowledge achievement gap by using technology-assisted guided learning in an undergraduate chemistry course. *Journal of Research in Science Teaching*, *57*(3), 368–392. https://doi.org/10.1002/tea.21596

*Low, K. S. (2010). *Effect of the card game SynTactic on the learning of organic chemistry reactions* (Master's thesis). http://hdl.handle.net/10497/4594

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*(4), 641–654. https://doi.org/10.1002/sim.698

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, *5*(4), 333–369. https://doi.org/10.1016/S0364-0213(81)80017-1

Maria, F., Dos Santos, T., & Mortimer, E. F. (2003). How emotions shape the relationship between a chemistry teacher and her high school students. *International Journal of Science Education*, *25*, 1095–1110. https://doi.org/10.1080/0950069032000052216

*Martin, M. W., & Shen, Y. (2014). The effects of game design on learning outcomes. *Computers in the Schools*, *31*(1–2), 23–42. https://doi.org/10.1080/07380569.2014.879684

Martinez-Garza, M., Clark, D., & Nelson, B. C. (2013). Digital games and the US National Research Council's science proficiency goals. *Studies in Science Education*, *49*(2), 170–208. https://doi.org/10.1080/03057267.2013.839372

*Martinez-Hernandez, K. J. (2010). *Development and assessment of a chemistry-based computer video game as a learning tool* (Doctoral dissertation). ProQuest Dissertations Publishing.

Mayer, R. E. (2014a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005

Mayer, R. E. (2014b). *Computer games for learning: An evidence-based approach*. MIT Press.

Mayer, R. E. (2019). Computer games in education. *Annual Review of Psychology*, *70*, 531–549. https://doi.org/10.1146/annurev-psych-010418-102744

Mayer, R. E. (2020). Cognitive foundations of game-based learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 83–110). MIT Press.

*Meesuk, K., & Srisawasdi, N. (2014). Implementation of student-associated game-based open inquiry in chemistry education: Results on students' perception and motivation. In *Proceedings of the 22nd International Conference on Computers in Education* (pp. 219–226). Asia-Pacific Society for Computers in Education.

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education*, *70*, 29–40. https://doi.org/10.1016/j.compedu.2013.07.033

*Merchant, Z., Goetz, E. T., Keeney-Kennicutt, W., Cifuentes, L., Kwok, O., & Davis, T. J. (2013). Exploring 3-D virtual reality technology for spatial ability and chemistry achievement. *Journal of Computer Assisted Learning*, *29*(6), 579–590. https://doi.org/10.1111/jcal.12018

Merrill, D. M. (2012). *First principles of instruction*. John Wiley & Sons.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group* (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269.

Moreno, R., & Mayer, R. E. (2002). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology*, *94*(3), 598–610. https://doi.org/10.1037//0022-0663.94.3.598

Moreno, R., & Mayer, R. E. (2004). Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology*, *96*(1), 165–173. https://doi.org/10.1037/0022-0663.96.1.165

Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, *97*(1), 117–128. https://doi.org/10.1037/0022-0663.97.1.117

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, *19*(3), 309–326. https://doi.org/10.1007/s10648-007-9047-2

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Mulder, J. D., & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, *1-11*, 638–648. https://doi.org/10.1080/10705511.2020.1784738

National Research Council [NRC]. (2009). *A new biology for the 21st century*. The National Academic Press. https://doi.org/10.17226/12764

National Research Council [NRC] (2011a). In M. A. Honey & M. L. Hilton (Eds.), *Learning science through computer games and simulations*. The National Academies Press.

National Research Council [NRC]. (2011b). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. The National Academies Press. https://doi.org/10.17226/13158

National Research Council [NRC]. (2012a). *Discipline-based education research: understanding and improving learning in undergraduate science and engineering*. The National Academies Press. https://doi.org/10.17226/13362

National Research Council [NRC]. (2012b). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://doi.org/10.17226/13165

National Research Council [NRC]. (2012c). *Challenges in chemistry graduate education: A workshop summary*. The National Academies Press. https://doi.org/10.17226/13407

National Research Council [NRC]. (2014). *Undergraduate Chemistry Education: A Workshop Summary*. The National Academies Press. https://doi.org/10.17226/18555

National Research Council [NRC]. (2015). *Reaching students: What research says about effective instruction in undergraduate science and engineering*. The National Academies Press. https://doi.org/10.17226/18687

National Research Council [NRC]. (2016). *Science Literacy: Concepts, Contexts, and Consequences*. The National Academies Press. https://doi.org/10.17226/23595

Neelen, M., & Kirschner, P. A. (2020). *Evidence-informed learning design: Creating training to improve performance*. Kogan Page Publishers.

*Okonkwo, I. G. A. (2012). *Effects of concept mapping and simulation-game teaching strategies on students' achievement and interest in enviromental concepts in chemistry* (Doctoral dissertation). University of Nigeria. http://www.unn.edu.ng/publications/files/images/MRSOKONKWOFINALPROJECTPh.D.pdf

Opfermann, M., Schmeck, A., & Fischer, H. E. (2017). Multiple representations in physics and science education–why should we use them? In D. F. Treagust, R. Duit, & H. F. Fischer (Eds.), *Multiple representations in physics education* (pp. 1–22). Springer.

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, *24*(1), 27–45. https://doi.org/10.1007/s10648-011-9179-2

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 27–42). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.004

Park, B., Knörzer, L., Plass, J. L., & Brünken, R. (2015). Emotional design and positive emotions in multimedia learning: An eyetracking study on the use of anthropomorphisms. *Computers & Education*, *86*, 30–42. https://doi.org/10.1016/j.compedu.2015.02.016

Parker, J., Becker, K., & Sawyer, B. (2008). Re-reconsidering research on learning from media: Comments on Richard E. Clark's point of view column on serious games. *Educational Technology: The Magazine for Managers of Change in Education*, *48*(1), 39–43 https://www.jstor.org/stable/44429544

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, *110*(6), 785–797. https://doi.org/10.1037/edu0000241

Parong, J., & Mayer, R. E. (2021). Learning about history in immersive virtual reality: Does immersion facilitate learning? *Educational Technology Research and Development*, *1-19*, 1433–1451. https://doi.org/10.1007/s11423-021-09999-y

Pekrun, R., & Linnenbrink-Garcia, L. (2014). *International Handbook of Emotions in Education*. Routledge.

Pekrun, R., & Perry, R. P. (2014). Control-value theory of achievement emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 120–141). Routledge. https://doi.org/10.4324/9780203148211

Pellas, N., Fotaris, P., Kazanidis, I., & Wells, D. (2018). Augmenting the learning experience in primary and secondary school education: A systematic review of recent trends in augmented reality game-based learning. *Virtual Reality*, *23*(4), 329–346. https://doi.org/10.1007/s10055-018-0347-2

Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D., & Um, E. (2014). Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction*, *29*, 128–140. https://doi.org/10.1016/j.learninstruc.2013.02.006

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Plass, J. L., Homer, B. D., MacNamara, A., Ober, T., Rose, M. C., Pawar, S., Hovey, C. M., & Olsen, A. (2019). Emotional design for digital games for learning: The effect of expression, color, shape, and dimensionality on the affective quality of game characters. *Learning and Instruction*, *70*, 101194. https://doi.org/10.1016/j.learninstruc.2019.01.005

Plass, J. L., & Kaplan, U. (2015). Emotional design in digital media for learning. In S. Y. Tettegah & M. Gartmeier (Eds.), *Emotions, technology, design, and learning* (pp. 131–161). Academic Press. https://doi.org/10.1016/b978-0-12-801856-9.00007-4

Plass, J. L., Mayer, R. E., & Homer, B. D. (2020). Theoretical foundations of game-based learning and playful learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 3–24). MIT Press.

Prensky, M. (2001). *Digital game-based learning*. McGraw-Hill.

Rahman, T., & Lewis, S. E. (2020). Evaluating the evidence base for evidence-based instructional practices in chemistry through meta-analysis. *Journal of Research in Science Teaching*, *57*(5), 765–793. https://doi.org/10.1002/tea.21610

Raker, J. R., Gibbons, R. E., & Cruz-Ramírez de Arellano, D. (2019). Development and evaluation of the organic chemistry-specific achievement emotions questionnaire (AEQ-OCHEM). *Journal of Research in Science Teaching*, *56*(2), 163–183. https://doi.org/10.1002/tea.21474

*Rastegarpour, H., & Marashi, P. (2012). The effect of card games and computer games on learning of chemistry concepts. *Procedia— Social and Behavioral Sciences*, *31*, 597–601. https://doi.org/10.1016/j.sbspro.2011.12.111

*Renner, J. C. (2014). *Does augmented reality affect high school student's learning outcomes in chemistry?* (Doctoral dissertation). ProQuest LLC.

Riopel, M., Nenciovici, L., Potvin, P., Chastenay, P., Charland, P., Sarrasin, J. B., & Masson, S. (2020). Impact of serious games on science learning achievement compared with more conventional instruction: An overview and a meta-analysis. *Studies in Science Education*, *55*(2), 169–214. https://doi.org/10.1080/03057267.2019.1722420

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, *58*, 136–153. https://doi.org/10.1016/j.compedu.2011.07.017

Ryan, R. M., & Rigby, C. S. (2020). Motivational foundations of game-based learning. In J. L. Plass, R. E. Mayer & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 25–52). MIT Press.

Sabourin, J. L., & Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing*, *5*(1), 45–56. https://doi.org/10.1109/T-AFFC.2013.27

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. MIT Press.

Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology*, *76*(4), 647–658. https://doi.org/10.1037/0022-0663.76.4.647

Schiefele, U. (2009). Situational and individual interest. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of motivation at school* (1st ed., pp. 197–222). Routledge.

Schola Europaea. (2019). Chemistry Syllabus – S4-S5. https://www.eursc.eu/Syllabuses/2019-01-D-46-en-2.pdf

Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 265–278). Psycnet https://psycnet.apa.org/record/2012-07988-016

Setiawan, H., & Phillipson, S. (2019). The effectiveness of game-based science learning (GBSL) to improve students' academic achievement: A meta-analysis of current research from 2010 to 2017. *REiD (Research and Evaluation in Education)*, *5*(2), 152–168. https://doi.org/10.21831/reid.v5i2.28073

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, et al. (Eds.), *Assessment in game-based learning: foundations, innovations, and perspectives* (pp. 43–58). Springer. https://doi.org/10.1007/978-1-4614-3546-4_4

Sicart, M. (2014). *Play matters*. MIT Press.

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*(4), 450–466. https://doi.org/10.1080/02680939.2017.1280183

Sinatra, G. M., Broughton, S. H., & Lombardi, D. O. U. G. (2014). Emotions in science education. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International Handbook of Emotions in Education* (pp. 425–446). Routledge.

Sirhan, G. (2007). Learning difficulties in chemistry: An overview. *Journal of Turkish science education*, *4*(2), 2.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, *64*(2), 489–528. https://doi.org/10.1111/j.1744-6570.2011.01190.x

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century-Crofts.

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14. https://doi.org/10.3102/0013189X08314117

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. https://doi.org/10.3102/0162373709352369

*Sousa Lima, M. A., Monteiro, Á. C., Melo Leite Junior, A. J., de Andrade Matos, I. S., Alexandre, F. S. O., Nobre, D. J., Monteiro, A. J., & da Silva Júnior, J. N. (2019). Game-based application for helping students review chemical nomenclature in a fun way. *Journal of Chemical Education*, *96*(4), 801–805. https://doi.org/10.1021/acs.jchemed.8b00540

*Srisawasdi, N., & Panjaburee, P. (2019). Implementation of game-transformed inquiry-based learning to promote the understanding of and motivation to learn chemistry. *Journal of Science Education and Technology*, *28*(2), 152–164. https://doi.org/10.1007/s10956-018-9754-0

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*, 60–78. https://doi.org/10.1002/jrsm.1095

Steegh, A., Höffler, T., Höft, L., & Parchmann, I. (2021). First steps toward gender equity in the chemistry Olympiad: Understanding the role of implicit gender-science stereotypes. *Journal of Research in Science Teaching*, *58*(1), 40–68. https://doi.org/10.1002/tea.21645

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002. https://doi.org/10.1136/bmj.d4002

*Stringfield, T. W., & Kramer, E. F. (2014). Benefits of a game-based review module in chemistry courses for non-majors. *Journal of Chemical Education*, *91*(1), 56–58. https://doi.org/10.1021/ed300678f

*Su, C.-H., & Cheng, T.-W. (2019). A sustainability innovation experiential learning model for virtual reality chemistry laboratory: An empirical study with PLS-SEM and IPMA. *Sustainability*, *11*(4), 1027. https://doi.org/10.3390/su11041027

*Sugiyarto, K. H., Ikhsan, J., & Lukman, I. R. (2018). The use of an android–based-game in the team assisted individualization to improve students' creativity and cognitive achievement in chemistry. *Journal of Physics: Conference Series*, *1022*(1), 012037. https://doi.org/10.1088/1742-6596/1022/1/012037

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Taber, K. S. (2009). Learning at the symbolic level. In J. K. Gilbert & D. Treagust (Eds.), *Multiple representations in chemical education* (pp. 75–105). Springer. https://doi.org/10.1007/978-1-4020-8872-85

Taber, K. S. (2013). Revisiting the chemistry triplet: Drawing upon the nature of chemical knowledge and the psychology of learning to inform chemistry education. *Chemistry Education Research and Practice*, *14*(2), 156–168. https://doi.org/10.1039/c3rp00012e

Talanquer, V. (2011). Macro, submicro, and symbolic: The many faces of the chemistry "triplet". *International Journal of Science Education*, *33*(2), 179–195. https://doi.org/10.1080/09500690903386435

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences*, *61*, 136–150. https://doi.org/10.1016/j.lindif.2017.11.015

Thompson, C. G., & von Gillern, S. (2020). Video-game based instruction for vocabulary acquisition with English language learners: A Bayesian meta-analysis. *Educational Research Review*, *30*, 100332. https://doi.org/10.1016/j.edurev.2020.100332

Tobias, S., & Fletcher, J. D. (2012). Reflections on "a review of trends in serious gaming". *Review of Educational Research*, *82*(2), 233–237. https://doi.org/10.3102/0034654312450190

Tobias, S., Fletcher, J. D., Dai, D. Y., & Wind, A. P. (2011). Review of research on computer games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 127–222). Information Age.

Tokac, U., Novak, E., & Thompson, C. G. (2019). Effects of game-based learning on students' mathematics achievement: A meta-analysis. *Journal of Computer Assisted Learning*, *35*(3), 407–420. https://doi.org/10.1111/jcal.12347

Towns, M., & Kraft, A. (2011). Review and synthesis of research in chemical education from 2000–2010. In *Paper presented at the Second Committee Meeting on the Status, Contributions, and Future Directions of Discipline-Based Education Research*. National Academies https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072594.pdf

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

Tsai, Y. L., & Tsai, C. C. (2020). A meta-analysis of research on digital game-based science learning. *Journal of Computer Assisted Learning*, *36*(3), 280–294. https://doi.org/10.1111/jcal.12430

Tsui, C. Y., & Treagust, D. F. (2013). Introduction to multiple representations: Their importance in biology and biological education. In D. F. Treagust & C. Y. Tsui (Eds.), *Multiple representations in biological education* (pp. 3–18). Springer.

Um, E. R., Plass, J. L., Hayward, E. O., & Homer, B. D. (2012). Emotional design in multimedia learning. *Journal of Educational Psychology*, *104*(2), 485–498. https://doi.org/10.1037/a0026609

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3

van Alten, D. C. D., Phielix, C., Janssen, J., & Kester, L. (2019). Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis. *Educational Research Review*, *28*, 100281. https://doi.org/10.1016/j.edurev.2019.05.003

van Merriënboer, J. J. G., & Kester, L. (2014). The four-component instructional design model: Multimedia principles in environments for complex learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 104–148). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.007

van Merriënboer, J. J., & Kirschner, P. A. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design*. Routledge.

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. (2017). *Metafor package: bias and sensitivity diagnostics*. Retrieved from https://stats.stackexchange.com/questions/155693/metafor-package-bias -and-sensitivity-diagnostics

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*(3), 360–374. https://doi.org/10.1037/met0000023

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, *34*(3), 229–243. https://doi.org/10.2190/FLHV-K4WA-WPVQ-H0YM

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

Wandersee, J. H., Fischer, K. M., & Moody, D. E. (2000). The Nature of Biology Knowledge. In K. M. Fisher, J. H. Wandersee, & D. E. Moody (Eds.), *Mapping biology knowledge* (Vol. 11, pp. 25–38). Springer.

Warfa, A.-R. M. (2016). Using cooperative learning to teach chemistry: A meta-analytic review. *Journal of Chemical Education*, *93*, 248–255. https://doi.org/10.1021/acs.jchemed.5b00608

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

*Weng, H. J., Chang, D. F., & Shyu, H. Y. (2015). Testing the effect of mnemonic strategy embedded in digital game. *ICIC Express Letters*, *9*(3), 827–833.

What Works Clearinghouse [WWC]. (2019). *Standards handbook (Version 4.1)*. Institute of Education Sciences.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

*Wood, J., & Donnelly-Hermosillo, D. F. (2019). Learning chemistry nomenclature: Comparing the use of an electronic game versus a study guide approach. *Computers & Education*, *141*, 103615. https://doi.org/10.1016/j.compedu.2019.103615

Wu, H. K., & Shah, P. (2004). Exploring visuospatial thinking in chemistry learning. *Science Education*, *88*(3), 465–492. https://doi.org/10.1002/sce.10126

Wouters, P., Paas, F., & van Merriënboer, J. J. G. (2008). How to optimize learning from animated models: A review of guidelines based on cognitive load. *Review of Educational Research*, *78*(3), 645–675. https://doi.org/10.3102/0034654308320320

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*(2), 249–265. https://doi.org/10.1037/a0031311

Wouters, P., & van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education*, *60*(1), 412–425. https://doi.org/10.1016/j.compedu.2012.07.018

Wouters, P., & van Oostendorp, H. (2017). Overview of instructional techniques to facilitate learning and motivation of serious games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 1–16). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_1

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., Simeoni, Z., Tran, M., & Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, *82*(1), 61–89. https://doi.org/10.3102/0034654312436980

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Hu, Y., Gallagher, T., Wouters, P., van der Schaaf, M., & Kester, L. (2022). Game-based learning has good chemistry with chemistry education: A three-level meta-analysis. *Journal of Research in Science Teaching*, 1–45. https://doi.org/10.1002/tea.21765