The effectiveness of information communication technology (ICT) used in second language (L2) classrooms: A meta-analysis

by

Songtao Wang

B.A., Tianjin University of Finance and Economics, 2013

A Thesis Submitted in Partial Fulfillment of the Requirements of the Degree of

MASTER OF ARTS

in the Department of Curriculum and Instruction

University of Victoria

The effectiveness of information communication technology (ICT) used in second language (L2) classrooms:

A meta-analysis

by

Songtao Wang

B.A., Tianjin University of Finance and Economics, 2013

Supervisory Committee

Dr. Tim Pelton, Co-Supervisor

Department of Curriculum and Instruction

Dr. Todd Milford, Co-Supervisor

Department of Curriculum and Instruction

Abstract

Objectives:

A number of primary empirical studies report strong and positive relationships between information and communication technology (ICT) and second language (L2) learning outcomes. However, these strong relationships were not observed in recent meta-analysis studies associated with ICT and L2 learning outcomes. This meta-analysis study aims to iterate the evaluation of effectiveness of ICT integrated L2 classrooms.

Method:

Initially, a study search and screening of the digital database were conducted to include quantitative studies exploring the use of ICT for L2 skills in classroom contexts with an experimental design. Then essential statistics from included studies were retrieved and coded. Standardized mean difference of comparison groups across included studies were calculated as the unit of analysis. Finally, a random effects model was used to pool mean effects sizes calculated from included studies. Heterogeneity test, subgroup analysis and publication bias were assessed for implication of study variation and reliability.

Results:

After a systematic study search and screening, 30 experimental studies were included for this meta-analysis, which yielded 43 effect sizes based on comparison groups across included studies. Results showed a large effect size in favor of technology-integrated instruction with substantial between-groups heterogeneity. Additionally, nine

moderating variables were identified, covering substantial, methodological and reporting features. Results also updated the current information on the state of methodological practice in the L2 research domain and suggested that the reliability of instruments used for measurements were overlooked and might lead to an overestimation of effect size.

Conclusion:

The current meta-analysis provided an updated evaluation of the ICT integrated L2 classrooms, supporting the effectiveness of integrating technologies into L2 instruction (with the exception of L2 pragmatics). Both theoretical and methodological maturation were observed in the research domain, as well as methodological limitations which might impact the precision of effect size. More research is needed to explore more specific topics, as well as methodological imperfections in ICT integrated L2 classrooms.

Keywords: meta-analysis, information and communication technology, second language classroom

Table of Contents

List of Tables

List of Figures

List of Acronyms

APA	American Psychology Association
CAI	Computer assisted instruction
CALL	Computer assisted language learning
CMC	Computer mediated communication
CI	Confidence interval
DDL	Data driven learning
d.f.	Degree of freedom
DGBL	Digital game-based learning
ESL	English as second language
EFL	English as foreign language
ES	Effect size
ICALL	Intelligent computer assisted language learning
ICT	Information and communication technology
L1	First language
L2	Second language
MALL	Mobile assisted language learning
MARS	Meta-Analysis Reporting Standards
NLP	Natural language processing

SLA	Second language acquisition
PI	Predictive interval
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ	Research question
SCMC	Synchronous computer mediated communication
SD	Standard deviation
SMD	Standardized mean difference
SLA	Second language acquisition
SSS	Systematic secondary synthesis

Chapter 1. Introduction

E-learning is a multi-faceted phenomenon that has become a methodological and pedagogical focus in educational research as the integration of information and communication technologies (ICT) has permeated everyday life (Haythornthwaite & Andrews, 2011). Compared to traditional face-to-face interventions, key benefits of ICT integrated instruction include flexibility and efficiency both in providing and receiving assisting cues for learners. For example, posting questions for students with different locations and learning schedules through an asynchronous online chatroom, or using mobile devices to track students learning patterns and outcomes to evaluate their learning progress and then provide appropriate feedback.

ICT tools are also commonly found in all types of education classrooms, including second language (L2) classrooms. Over the past 20 years, research has focused on documenting the effects of using ICT in L2 classrooms (Otto, 2017). Such practices include integrating digital technology, communication tools, and networks into classroom learning for learners and educators to access, manage, integrate, evaluate, and create learning materials. Given the strength and convenience of accessing learning resources using ICT tools, theories of second language acquisition (SLA) have changed and developed with research to accommodate technological integration.

SLA theories are at the root of development and practice of ICT integrated L2 classrooms (Smith, 2017). Major SLA theories and frameworks that guide L2 teaching and learning can be grouped into three overlapping types: the cognitive approach, the interaction approach, and the constructivist approach. Regardless of the approach, modern

SLA theories reflect changes in L2 classrooms in response to technological developments. More specifically, pedagogies in ICT integrated L2 classrooms have transformed through three main stages, namely "Behavioristic", "Communicative" and "Integrative" (Warschauer & Healey, 1998, p. 57), to an endpoint of being "normalized" where ICT is "invisible, embedded in everyday practice" (Bax, 2003, p. 23). To better utilize ICT in the L2 classroom, it is necessary to experiment with and document designs, contexts and outcomes of L2 learning integrated with ICT.

Researchers have contributed to the body of valid evidence on using ICT in L2 classrooms both via theoretical and empirical studies. First, from a theoretical point of view, learning behaviors are argued to be affected by digital technology. According to cognitive load theories, increased distractions while using digital materials have the potential to impair learning outcomes due to cognitive overload (Sweller, 1988). For example, it is widely acknowledged that mobile phones enable users to obtain information anytime anywhere, allowing learning to be spontaneous, authentic, informal, situational, and personalized (Thornton & Houser, 2005). However, mobile devices also can create external distractions to learners and thus hinder the achievement of learning outcomes. Yang (2017) recruited English as second language (ESL) learners at the university level to test the effects of attention cueing on phonetic awareness performed on mobile phones. Yang found that both visual cueing and verbal tests could be beneficial for learners to develop a greater conceptual understanding if the context continued to attract learners' attention during the intervention. This result indicates that simply using ICT would not promote better learning outcomes if the application of technological features was not in alignment with SLA theories.

Second, primary studies report strong and positive relationships between ICT and L2 learning outcomes (Cerezo et al., 2014). There was already a level of maturity and efficacy in research results with respect to the contribution of ICT to SLA (Plonsky & Zielger, 2016), via a range of substantive questions. In the review on the effectiveness of using computer-assisted language learning (CALL) instructions, Plonsky and Zielger (2016) reported an overall positive and medium effect size covering 408 primary studies. Their findings indicated that there is a significant benefit for learners participating in CALL instruction in terms of L2 learning outcomes.

However, strong relationships were not always found in later studies. Despite the strong magnitudes found favoring the use of ICT in L2 classrooms, Plonsky and Ziegler (2016) also reported features of primary studies in the research domain - such as evolving and emerging ICT tools and a variety of L2 targeted skills - which need further investigation. Recent meta-analysis studies have also outlined several measurement issues in the domain, including multiple reports, instruments reliability and range restrictions of psychometric tools, which would endanger the precision and trustworthiness of previously reported effects (Plonsky & Oswald, 2014). Many earlier studies have not reported clear effect sizes for the relationships between ICT integrated L2 classrooms and L2 outcome measures.

In order to improve the understanding of e-learning phenomenon in the context of L2 learning, researchers need to evaluate the effects of technological developments in their application supporting language education. The purpose of this study is to provide updated evidence on the validity of using ICT for L2 instruction. We identified four general issues in the research domain as:

- What are the trends of substantive, methodological and reporting features in ICT integrated L2 classrooms research?
- What is the current effectiveness of ICT integrated L2 classrooms?
- How do moderator variables in current studies moderate learning outcomes in ICT integrated L2 classrooms?
- Is there publication bias present in the research domain?

Meta-analysis was chosen as the method to systematically collect and analyze secondary data from primary studies for three reasons. First, meta-analysis is one of the systematic secondary synthesis (SSS) approaches which allows researchers to examine the statistical inferences across available evidence. It allows researchers to draw inferences across research contexts, designs and topics from an interdisciplinary point of view.

Second, procedures of meta-analysis are systematic and standardized (see Appelbaum et al., 2018) which would lend support to better consistency in reporting of inferences drawn from primary studies. Meta-analysis researchers need to formulate a meta-analytical topic based on a thorough literature review followed by a systematic study search to collect available evidence. Then, essential information of included studies, such as study contexts, study designs and results, are collected to summarize and evaluate the current research trends of ICT integrated L2 classrooms, as well as apply statistical procedures to further examine the directions and magnitudes of the effects. An analysis of heterogeneity and moderating factors also contribute to the examination of the variability of L2 learning outcomes. Reporting features and publication bias is discussed to illuminate the research practice and future directions in the research domain.

Third, meta-analysis can potentially illuminate the inferential gaps between ICT integrated L2 classrooms and methodological features in the SLA research domain, and thus contribute to the inferential strength of results from primary studies. In addition, issues that might be overlooked in the individual-study-level, such as types of ICT tools, and types of educational settings, might be synthesized across studies. Contradicted results from primary studies, such as the moderating effects of learners' proficiency level are also be examined with updated evidence.

In summary, this chapter provides an introduction of e-learning as a phenomenon and then delves deeper into the use of ICT in L2 classrooms. Theoretical and practical evidence supporting the efficacy of using ICT for SLA were provided along with reasons for updated evaluation research to further contribute the evidence of validity of ICT integration into L2 classrooms. Lastly, three reasons were given as advantages of meta-analysis to support the current study. In the next chapter, a thorough literature review related to ICT integrated L2 classrooms are provided along with a review of theoretical frameworks for SLA research. Research gaps and research questions of the current study are also described in detail.

Chapter 2. Literature review

This chapter is composed by four major sections: implementation of information and communication technology (ICT) used in second language (L2) classrooms, second language acquisition (SLA) theoretical frameworks, rationale for evaluation of ICT used in L2 classrooms, and previous meta-analysis of ICT integrated L2 classrooms. The literature review begins with a review of the development and implementation of ICT used in L2 classrooms. Then an argument for promoting evaluation of ICT integrated L2 classrooms is provided. Critical SLA theoretical frameworks that guide SLA research are reviewed. Lastly, previous meta-analysis studies of ICT integrated L2 classrooms are reviewed in three parts: general inquiry of ICT used in L2 classrooms, types of ICT used in L2 classrooms, and targeted L2 skills as L2 learning outcomes. Research gaps identified from the literature review are addressed in detail. This chapter concludes with a series of research questions (RQs) regarding conducting a meta-analysis on the effectiveness of ICT used in L2 classrooms.

2.1. Definition of key terms

Second language acquisition (SLA). SLA is the shorthand for all the language teaching and learning situations and activities that include both children and adults learning additional or second languages (L2) beyond their first language (L1). This study uses SLA to refer to practice and inquiries of foreign or second language learning in the medium of everyday life, professional contexts, or targeted L2 skills in the second language classroom (L2) classroom (Sauro & Chapelle, 2017).

L2 classrooms. L2 classrooms consist of the pedagogical practice of leaning a second language. It is an effective influence on SLA when compared with naturalistic exposure or no instructions (Leow, 2019). Although not all L2 learning activities would be necessarily carried out in a traditional classroom format (Kukulska-Hulme et al., 2017), such as mobile learning that happens in the fields outside physical classrooms, this study takes a broad definition of classroom and instruction using L2 classrooms to refer to the context for all L2 instructional activities.

L2 skills. L2 teachers and learners often approach L2 learning through multiple linguistic aspects consisted of types of L2 skills. Different types of L2 skills share the same theoretical premises that is any instructional designs should aim for effective changes in learners' L2 proficiency. L2 proficiency could focus on both meta-linguistic cognitions and targeted linguistic aspects (Blake, 2016; Schmidgall & Powers, 2017; Sauro & Chapelle, 2017). This study uses seven L2 skills explain L2 learning outcomes including L2 listening, L2 speaking, L2 reading, L2 writing, L2 vocabulary, L2 grammar, and L2 pragmatics.

Information and communication technology (ICT). ICT belongs to a series of electronically based platforms that use digital data collected or created to augment learning experience and built to simulate certain human qualities of interactions to learners. This study uses ICT as an overarching term to refer to educational technologies used in L2 classrooms, including computer-assisted language learning (CALL), intelligent CALL (ICALL), computer-assisted instructions (CAI), computer-mediated communication (CMC), digital game-based learning (DGBL), and mobile-assisted language learning (MALL).

CALL. CALL refers to computer-assisted language learning which explores language teaching and learning incorporating ICT based on PC (personal computer). As a 'controversial general label' it is (Levy et al., 2015), the term CALL encompasses a broad scope of computational technology when used in practice. This study includes minor strands of CALL tools when referring to specific types of computational technology, such as computer-assisted instruction (CAI), intelligent-CALL (ICALL), computer-mediated communication (CMC), and mobile assisted language learning (MALL) (Sauro & Chapelle, 2017), for further clarification.

CAI. CAI (computer-assisted instruction) and blended learning both represent L2 classrooms that incorporate specific types of multimedia tools based on PC as an add-on to traditional face-to-face L2 classroom. Examples of CAI could be (1) e-textbooks as a multimedia transplant of traditional-formatted knowledge; (2) Web pages and e-portfolios which are open online spaces for learners and practitioners to share and review work in a collaborative form; (3) multimedia gloss, which are online tools with easy access for students to definitions of L2 words.

Blended learning is an extension of CAI that emphasizes the strategy of combining computer-mediated online resources with traditional face-to-face instruction (Picciano, 2019). Blended learning tools include blogs, wikis, electronic workbooks and chats that are also considered as online sourced CAI. Thus, this study treats blended learning and CAI as the same source of ICT interchangeably.

CMC. Computer-mediated communication (CMC) emulates real-time interaction patterns in face-to-face L2 classroom communication through synchronous and asynchronous forms of messages to enhance production comprehension, process and

exchange of L2 information. Although some pedagogical applications of CMC in L2 classrooms, for example, online chats, are in overlap with blended learning (Ziegler, 2015), this study defines ICT that imitate the discussion in L2 classrooms with flexible forms of conversational interactions (e.g. vocal or textual, synchronous or asynchronous) as CMC.

ICALL. Intelligent computer-assisted language learning (ICALL), describes CALL instructions that use novel analytical learning tools and methods used in L2 classrooms. Examples of ICALL include data-driven learning (DDL), automatic feedback and adaptive natural language processing (NLP) that are used in computational linguistics to further analyze human-computer interactions (Amaral & Meurers, 2011; Ziegler et al., 2017; Xu et al., 2013).

DGBL. Digital game-based learning (DGBL) describes a range of learning activities through computer-based intelligent tutors, games, and simulations. Pedagogical functions of using DGBL in L2 classroom emphasize serving contents, requesting responses and giving feedbacks on structured progressive-instructional treatments depending on the nature of learners' responses (Cope & Kalantzis, 2017). The current study treats ICT such as multi-player games, simulations, online gaming collaborations and other user-driven ICT which would provide novel contest based on learners' progressive performance (Plonsky & Ziegler, 2016) as DGBL.

MALL. Mobile-assisted language learning, MALL, is a term used in SLA research in contrast with CALL. MALL is the adaptation of using portable for ubiquitous learning carried indoors and outdoors (Kukulska-Hulme et al., 2017). Like CALL, MALL is an overarching term including various mobile devices that focus on portability, versatility of feature, connectivity and individuality of L2 classroom (Golonka et al., 2014). Examples

of MALL include using tablet PC, PDA (personal digital assistant) and smartphones in L2 classrooms.

2.2. Implementation of ICT in L2 classrooms

In this section, a review of the implementation of ICT and its roles in L2 classrooms are presented. ICT serves and supports SLA in a progressive way to transform traditional knowledge in a digital approach and create new SLA instructional paradigms (Otto, 2017). The following section explains the technological and theoretical progression of the role of ICT used in L2 classrooms in two parts: ICT as an ancillary tool and ICT as the major generator of instruction in L2 classrooms.

2.2.1. ICT as an ancillary tool in L2 classrooms

ICT first appeared to serve an ancillary role in L2 classrooms that supplemented extensive practice for repetitions of the target language. From the 1960s to the early 1980s, research institutions put out CALL systems to provide students with automated media-enriched interactive practice with immediate feedback and comprehensive learning outcomes in the form of CAI. Pioneering examples of CALL systems include (1) the Programmed Logic for Automated Teaching Operations (PLATO) in University of Illinois at Urbana-Champaign that developed over 50,000 hours of online language instructions per semester on campus; (2) the Tutorial Russian Project at Stanford University which focused on tutorials and drills of pronunciation and speaking for first-year Russian-major students; and (3) the Time-shared Interactive Computer Controlled Information Television (TICCIT) at the University of Texas and Brigham Young University, was a CAI-based

courseware for English literacy education that featured recordkeeping and an advisor function (Suppes, 1968; Anderson, 1976; Hart, 1995). Early ICT in L2 classrooms supplemented classroom interactions in addition to teacher instructions, which reflected the Interactivity aspect from Long's interaction hypothesis (Otto, 2017).

2.2.2. ICT as a major generator of instruction in L2 classrooms

ICT then became a major drive in building connections through virtual learning environments between the 1980s to the 1990s. Afforded by the capability of microcomputers, new contexts of CALL featuring DDL and DGBL emerged in L2 classrooms. Both types of ICT emphasized the authentic language input through interactive tools in contrast to drill-and-practice. ICT broke out of the drill and practice mode to materialize communication skills through authentic input. The shift in ICT's role was also reflected in SLA theories. The focus of SLA moved from eclecticism toward communicative competence defined by five proficiencies: Communication, Cultures, Connections, Comparisons and Communities (National Standards in Foreign Language Education Project, 1996) in the 1990s.

Besides using DDL and DGBL, other types of ICT were also used as a major drive of authentic input in L2 classrooms. Examples include learning management systems and web-based CALL materials stored in local and remote servers. They allowed multiple users to communicate and collaborate in a pre-built virtual community. ICALL also gained attention by SLA researchers for processing natural language input in CALL contexts to mediate human-computer communication. However, this field remained as an under covered field (Ziegler et al., 2017) except for few major implementations for analyzing

conversational texts in the target language, such as the Athena Project developed with a multilingual NLP system.

In summary, ICT's role in L2 classrooms began with the digitalization of SLA materials and has extended through offering adaptive instructional support today. The change of ICT's role in L2 classrooms was not solely due to technical development, but also reflected SLA theories through designs and instructions. In the next section, critical SLA theoretical frameworks are reviewed to provide epistemological contexts for using ICT in L2 classrooms.

2.3. Review of SLA theoretical frameworks

SLA theoretical frameworks reflect the integration of ICT in L2 classrooms (Smith, 2017). This section provides a review of the intersection of critical SLA theoretical frameworks and ICT. First, an argument of identifying and utilizing appropriate SLA theoretical frameworks when using ICT in L2 classrooms is provided. Then, current SLA theoretical frameworks used in SLA research are reviewed including the cognitive approach, the interaction approach, and the constructive approach. Adaptations and constructs incorporating ICT in SLA theoretical frameworks are discussed.

SLA research is a field of academic inquiry about theoretical, empirical and pedagogical issues. It inquiries into the mechanism of SLA under a series of instructional manipulation approaches which enable or facilitate L2 learning (Loewen, 2020). SLA theoretical frameworks have been proposed to help interpret research questions (RQs), variables, and learning outcomes to help understand L2 learning mechanisms. Research about ICT integrated L2 classrooms is considered as a sub-strand of SLA research (Smith,

2017). Given the nexus of such research is to investigate relationships between ICT and L2 learning, study investigating ICT integrated L2 classrooms would need to begin in SLA theoretical frameworks.

An SLA theoretical framework is a collection of ontological, epistemological and methodological stances that SLA researchers and practitioners hold when they inquire, design and argue for effects of L2 instruction. SLA research is interdisciplinary, with an estimated 60 models and hypotheses supporting current SLA theoretical frameworks (Long, 2007). Three foundational SLA theoretical frameworks are used in SLA research to guide research design, assumptions and research constructs (Jia, 2019; Smith, 2017). Next we review three foundational theoretical frameworks including cognitivist approach, interaction approach and constructivist approach.

First, the cognitivist approach denotes efforts in understanding both functional and neurobiological descriptions of the SLA process though the exposure to L2 representations (information processing theories), as well as practical-oriented explanations of L2 development (emergentists). ICT integrated L2 classrooms would fit into the cognitivist construct when students are required to be actively handling information. Examples of constructed variables that reflect cognitivism in ICT integrated L2 classrooms include eliciting task type, production assessment, treatment length (time pressure), and working memory.

Second, the interaction approach has roots in Long's Interaction Hypothesis and focuses on language learners' linguistic input and interaction with interlocutors (the representation of interlanguage that learners use before acquiring the target language).

The interaction approach argues that L2 is acquired when learners notice differences between their own formulations of the target language and the language used by their conversational partners (Long, 1996). Interactionists also emphasize learning behaviors during L2 learning through drill and feedback. Major constructs of ICT integrated L2 classrooms what would fit in interaction approach include L2 input, human-computer interaction, explicit feedback, receptive assessment and implicit task type (e.g. drill practice) (Chapelle, 2006).

Third, the constructivist approach argues that the function of the brain is to explain and construct understanding within an authentic context. Hence, learning occurs through experiencing and self-constructing knowledge in a supportive learning environment. Popular SLA theories that are influenced by constructivism include sociocultural theory (SCT) and situated learning theory (Jia, 2019). SCT argues that learning is essentially a social process developed along with participation in cultural, linguistic, and historically formed settings. Consequently, SCT takes SLA as being facilitated by social interaction within the operation of the zone that learners could comprehend through proper scaffolding, known as the zone of proximal development (ZPD). In addition, situated learning theory extended the concept of social interaction by highlighting situated communication in which learners could engage in the target language (Kaufman, 2004). Pedagogical constructs of constructivism applied within ICT integrated L2 classrooms include L2 settings, institutional context and peer interaction through ICT (Lantolf et al., 2015).

In summary, the use of ICT in L2 classrooms cannot be separated from foundational SLA theoretical frameworks. Assertions, hypothesis and expectations of ICT integrated L2 classroom converge with theoretical and pedagogical constructs of critical SLA theoretical frameworks so that processes and outcomes of ICT integrated L2 classrooms can be seen as reliable and generalizable. That is to say that the effectiveness of using ICT would need to be comparable in efficiency to that in traditional face-to-face classrooms. In the next section, a rationale for examining and evaluating the effectiveness of ICT are discussed.

2.4. A rationale for evaluation of research on ICT integrated L2 classrooms

This section provides a rationale for conducting evaluation research of using ICT in L2 classrooms. Arguments are provided in two aspects: (1) ICT integrated SLA theories; and (2) the fast-paced development of ICT.

First, constructs in SLA theoretical frameworks provide new learning opportunities when integrated with ICT. Hence, L2 teachers inquire whether and how ICT has been effectively used in L2 classrooms. For teachers, it is seemingly an irreconcilable struggle between following conventional SLA theories while exploring ICT as a new element in L2 classrooms. Paradoxically this has also caused fear among teachers who worry that ICT might effectively replace teachers' due to the human-like interactive tendency of ICT (Blake, 2013; Salaberry, 2001). Drawn from pedagogical concerns, it is valuable to evaluate the use of ICT in L2 classrooms comparing with L2 learning in traditional face-to-face classrooms.

Second, SLA researchers wonder whether the pedagogical benefits if ICT have been fully explored with respect to their use in L2 classrooms (Chun, 2016). New types of ICT could yield complex pedagogical designs in L2 classrooms. Older ICT types often overlap with the new and may persist if their instructional media can be still be presented on the current systems. For example, web-based audiovisual materials, local-based CAI tools and digital corpus have been present in L2 classrooms since the 1980s (Otto, 2017), and now they are still being widely researched with new technical features. Hence it is critical to present up-to-date evaluation for the effectiveness of types of ICT used in L2 classrooms.

To address the practical and technical concerns of using ICT, it is necessary to iterate the process of evaluating and summarizing efficiency for using ICT in L2 classrooms. One of the trusted and efficient forces for summarizing results systematically is meta-analysis. In the following section, a review of meta-analysis studies investigating the effectiveness of ICT used in L2 classrooms are provided.

2.5. Meta-analysis studies on ICT integrated L2 classrooms

Researchers constantly refer to meta-analysis studies for evaluation of L2 classrooms because it is a more systematic and replicable approach to validate findings across primary studies (Plonsky & Oswald, 2015). A growing number of meta-analysis studies have also provided accumulated evidence on the effectiveness of ICT integrated L2 classrooms (Liu et al., 2002). In this section, a review of meta-analysis studies investigating the effectiveness of ICT used in L2 classrooms are provided in three parts:

(1) the effectiveness of ICT integrated L2 classrooms; (2) the effectiveness of types of ICT; and (3) targeted L2 skills as ICT integrated L2 outcomes.

2.5.1. The effectiveness of ICT integrated L2 classrooms

Evidence of the overall effectiveness of ICT integrated L2 classrooms has accumulated over the past 20 years. There are three major meta-analysis studies that have documented a broad spectrum of ICT and learning outcomes mainly focusing on ICT integrated L2 classrooms.

Zhao (2003) conducted the first meta-analysis study that examined the effectiveness of ICT in L2 classrooms using CALL as an example. The analysis examined nine empirical studies published from 1997 to 2000 and reported a medium to large mean effect size (ES = 0.81, 95% CI [0.55, 1.07]). This early meta-analysis has been a cornerstone for subsequent CALL meta-analysis studies each indicating that there is a positive effect of using CALL in L2 classrooms. Grgurović et al. (2013) updated Zhao's (2003) results including 37 CALL empirical studies published from 1970 to 2006. this meta-analysis study extended the knowledge of CALL by including research design as a substantive feature in the analysis. The quasi-experimental design group yielded a higher effect size (ES = 0.35, 95% [0.26, 0.44]) comparing with the experimental design group (ES = 0.23, 95% CI [0.14, 0.32]). Although the mean effect sizes appear to have decreased since Zhao's study (2003), both groups indicated a small but significant result favoring CALL used in L2 classroom in contrast to traditional face-to-face classrooms. A more recently published meta-analysis study also investigated the effectiveness of CALL under experimental design (Sharifi et al., 2018). The latest meta-analysis study examined

140 empirical studies using ICT in L2 classrooms published between 1990 and 2016. Results were consistent with the previous ones presenting an overall medium mean effect size (ES = 0.50, CI = [0.43, 0.60]) in favor of CALL.

Results of these three major meta-analysis studies indicated positive effects on using CALL in L2 classrooms although variation exists. Despite of the difference in experimental conditions, a series of heterogeneity tests in two of the meta-analyses showed significant between-studies differences in terms of types of ICT and targeted L2 skills (Grgurović et al., 2013; Sharifi et al., 2018). In the following section, reviews of effectiveness of ICT types and targeted L2 skills in meta-analysis are provided.

2.5.2. The effectiveness of types of ICT

Researchers have also addressed questions regarding diversity and complexity of ICT types used in L2 classrooms (Golonka et al., 2014; Plonsky & Ziegler, 2016). In this section meta-analysis studies that examined types of ICT are reviewed.

Four meta-analysis studies have contributed to the understanding of the effectiveness of CMC used in L2 classrooms. Lin et al. (2013) and Ziegler (2016) reviewed the use of Synchronous-CMC (SCMC) in L2 classrooms. Results from both studies showed a small to insignificant effect sizes favoring SCMC (ES = 0.33, 95% CI [0.18, 0.49]; ES = 0.13. 95% CI [-0.11, 0.38]) including primary studies published between 1990 and 2012. Another researcher examined the effectiveness of CMC used in L2 classrooms and similarly found a small effect size (ES= 0.441, 95 % CI [0.30, 0.59]) (Lin, 2014, 2015). In both studies subgroup analysis indicated L2 setting, task type, L2

proficiency and instructor type are significant moderator variables to L2 learning outcomes (ibid)

Other ICT types were also examined in meta-analysis studies. Chiu et al. (2012) reviewed 14 studies that used DGBL in L2 classrooms. Results similarly showed a small effect size (ES = 0.67, 95% CI [0.55 0.80]). Chang and Lin (2013) examined the effectiveness of web based L2 instruction. Results showed a medium effect size (ES= 0.67, 95% CI [-0.20, 2.82]) in favor of CALL. Boulton and Cobb (2017) examined the effectiveness of DDL used in L2 classrooms. Results showed large effect sizes in favor of using DDL in L2 classroom for both studies with quasi-experimental design (ES = 1.50, 95% CI [1.28, 1.71]) and studies with experimental design (ES = 0.95, 95% CI [0.67, 1.22]). A subgroup analysis including publication variables, design variables, treatment variables and objective variables revealed a void of research examining L2 speaking and reading skills in the language corpus.

Lastly, emerging ICT such as MALL was also under examination in L2 classrooms (Chun, 2016). As with the first twenty years of CALL, meta-analysis studies of MALL were supported by objective evidence of the pedagogical implications (Burston, 2015), focusing on qualitative review to provide an understanding of the research trends regarding the type of applications, interaction forms, and learning outcomes. For example, Burston (2015) examined 291 primary studies with MALL implemented projects and found that MALL studies only began showing interesting results in 2011. Meta-analysis studies in recent years also concluded promising results of the effectiveness of MALL in L2 classrooms. Medium to large effect sizes were reported

from three meta-analysis studies including a total of 77 primary studies published between 2008 and 2017 with a tendency that smaller effect size were reported in later published meta-analysis (e.g., Sung et al., 2015; Taj et al., 2016; Cho et al., 2018). In addition, later published study included subgroup analysis indicating between-studies variations in addition to ICT types ($Q = 103.27$, $p < 0.05$) (Cho et al., 2018).

2.5.3. The effectiveness of using ICT for L2 skills

Another substantive feature when evaluating ICT integrated classrooms is targeted L2 skills described and measured as L2 learning outcomes. Meta-analysis studies mainly focused on topics in L2 listening (e.g., Kang, 2019; Perez et al., 2013), L2 vocabulary (e.g., Abraham, 2008; Chen et al., 2018; Chiu, 2013; Mahdi, 2018; Perez et al., 2013; Tsai & Tsai, 2018; Yun, 2011), L2 speaking (e.g., Lin, 2015), L2 reading (e.g., Abraham, 2008; Taylor, 2009; 2013a; 2013b), and L2 writing (e.g., Xu et al., 2019).

L2 vocabulary learning has been widely investigated in ICT integrated L2 classrooms using ICT including CAI, DGBL and MALL. CAI, such as computer-mediated glosses (Abraham, 2008), visual-assisted glosses (Yun, 2011), captioned videos (Prez et al., 2013), and a mix of traditional CAI tools (Chiu, 2013) were examined in previous meta-analysis studies. Overall use of CAI for vocabulary learning showed a range of effect sizes (ES = 1.40, 95% CI [0.83, 1.98]; ES = 0.37, 95% CI [0.22, 0.51]; ES = 0.866, 95% CI [0.58, 1.15]; ES = 0.75, 95% CI [0.34, 1.15]) from primary studies published from 1994 to 2011. DGBL and MALL also showed medium to large effect sizes assisting L2 vocabulary learning (ES = 0.97, 95 % CI [0.58, 1.36]; ES = 1.02, 95% CI [0.51, 1.55]; ES = 0.97, 95 % CI [0.58, 1.36]; ES = 1.02, 95% CI [0.51, 1.55]) (Chen

et al., 2018; Lin & Lin, 2019; Mahdi, 2018; Tsai & Tsai, 2018). Each of these four meta-analysis studies found significant between-studies heterogeneity ($Q = 26.43$; $Q = 47.73$; $Q = 75.08$; $Q = 274.73$, $p < 0.05$) which indicated uncovered factors might moderate L2 learning outcomes.

Meta-analysis studies that examined other L2 skills (reading, speaking, listening, and writing) also demonstrated positive learning results. L2 reading appears to benefit from using ICT - reporting large effect sizes (ES = 0.73, 95% CI [0.22, 1.25]; ES = 1.09, 95% CI [0.78, 1.41]) (Abraham, 2008; Taylor, 2006; Taylor, 2010). Moderator variable such as assessment type was reported with significant within-groups heterogeneity ($Q = 78.52$, $p < 0.001$). Perez et al. (2013) and Kang (2019) examined the effectiveness of using CAI for L2 listening learning. Results showed medium effect sizes for captioned videos (ES = 0.99, 95% CI [0.60, 1.40] and multimedia tools (d = 0.69, 95% CI [0.67, 0.72]). Lin (2014) investigated using CMC for improving L2 speaking. Results showed a medium effect size (ES = 0.40, 95% CI [0.15, 0.65]) favoring CMC. Lastly, Xu et al. (2019) examined using ICT for L2 writing instructions. Results showed medium to large effect sizes favoring using ICT to improve L2 writing quality ($ES_{(d)} = 0.93$, 95% CI [0.80, 1.06]; $ES_{(g)} = 1.28$, 95% CI [0.85, 1.70]). Subgroup analysis indicated that type of ICT, genre of writing, and curriculum intensity are significant moderator variables in ICT integrated L2 classrooms.

In summary, the overall effectiveness of ICT integrated L2 classrooms has been summarized with meta-analysis studies examining more than 200 primary studies published between 1997 and 2016. In addition to the promising results found for using

ICT in L2 classrooms, the analysis also reported significant between-studies heterogeneity indicating unrevealed variations contributed by different types of ICT and targeted L2 skills. On one hand, emerging ICT types such as ICALL and MALL did not indicate consistent results on the effectiveness due to small sample sizes, thus, there is a lack of convincing evidence. Emerging ICT might also take time to be adapted into L2 classroom teaching before sufficient evidence is available to be reviewed (Otto, 2017). On the other hand, the dearth of homogeneous implications across studies regarding the effectiveness of using different ICT types and targeted L2 skills endangered the implication taken from meta-analysis studies. Meta-analysis studies focused on different moderator variables of ICT integrated L2 classrooms might even produce contradictory results. However, there is still a void of systematic and updated evidence of ICT used in L2 classrooms. In the following section, research gaps identified from the review of meta-analysis studies are discussed.

2.6. Research gaps

Meta-analysis studies have provided promising results regarding the effectiveness of ICT integrated L2 classrooms. However, results are not free from criticism. In this section, research gaps from previous meta-analysis studies including missing emerging ICT and inconsistent findings across studies are provided. Finally, research questions (RQ) are proposed for the current study.

2.6.1. Meta analyses missing emerging ICT

So far, more than 30 meta-analysis studies on ICTs used through L2 classroom have been published , but few have compared various types of ICT used in L2 classrooms (e.g., Chun, 2016; Golonka et al., 2014; Grgurović et al., 2013; Plonsky & Ziegler, 2016; Sharifi et al., 2018). Lacking systematic comparisons caused concerns regarding the level of integration and effectiveness using ICT. For example, Plonsky and Zeigler (2016) found that CMC is widely investigated in a total of 83 primary studies while DGBL and MALL were far from being fully investigated. The difference in research efforts of ICT might cause biased interpretation for the effectiveness of ICT in L2 classrooms.

In addition, the number of primary studies examining emerging ICT is ascending from 1990 through 2016 and later published articles appear to have provided superior evidence. Comparing with early published studies, recently published ones reported smaller effect size (ES = 0.13) with more primary studies included (ES = 0.33) (Lin et al., 2013; Ziegler, 2016). Variation of effect size might be due to the adaptation of emerging ICT in L2 classrooms. For example, it was not until 2015 when the effectiveness of MALL was examined in the form of meta-analysis (Burston, 2015; Chun, 2016). To the point when this review was written, there had been 7 out of 30 meta-analysis studies that dedicated to MALL. Hence, an updated meta-analysis comparing the effectiveness across multiple types of ICT would help further clarify the effectiveness of ICT integrated L2 classrooms.

2.6.2. Inconsistent findings in meta-analyses

Another concern raised from the review of meta-analysis studies is inconsistent findings across studies. More specifically, it consists of three issues (1) high variety of coded variables; (2) contradictory findings of the same variable; and (3) inconsistent outcomes.

Firstly, variables coded in previous meta-analysis studies were extremely variable from study to study due to observed complexity and variety of the nature of L2 classrooms (Norris & Ortega, 2000). While meta-analysis studies provide the power to synthesize the cumulative evidence of the effectiveness of L2 classrooms under various conditions from primary studies (Norris & Ortega, 2006), it is a high-inference and subjective procedure to interpret moderating factors from previous meta-analysis studies. Firstly, reporting of results varies between reviews. For example, L2 proficiency was widely investigated as an independent variable in meta-analysis studies and indicated inconsistent results. Abraham (2008) and Yun (2011) both investigated the effectiveness of multimedia CAI tools used in L2 vocabulary learning while only the previous study reported L2 proficiency as a significant moderator variable. Moderating effects of L2 proficiency was also inconsistent for different L2 skills, such as L2 listening learning (non-significant) and L2 vocabulary learning (significant), as well as when using captioned video (Prez et al, 2013).

Secondly, previous researchers utilized differing terminology in describing types of research contexts as independent variables, which might confuse practitioners and future researchers. For example, student age and educational level were often found

reported in previous meta-analysis studies to indicate the maturity of learners. Although these two variables were measured on different scales, typically adult learners would be found in undergraduate or college level participants and younger learners in the primary or secondary level. When different scales are adopted in different studies, results might be conflicted based on the researcher's interpretation of the experiment. For example, learner age was found to be a significant moderator variable when using MALL tools in L2 classrooms (Mahdi, 2017) while non-significant findings were reported with educational level variables (Cho et al, 2018).

Lastly, previous meta-analysis researchers disagreed on attributes when assessing L2 outcome measures as a dependent variable. For example, Yun (2011) and Lin & Lin (2019) both reported assessment type as a significant moderator variable to the effectiveness of using ICT for L2 vocabulary learning. However, the two studies reported different assessment attributes using vocabulary test types and receptive/productive assessments. Difficulties in finding consistent implications from previous studies increase when studies investigating the same L2 skills disagreed on attributes of outcome measures. For example, Lin (2014) reported "elicit assessment" type as an independent variable in the meta-analysis. While in the later mete-analysis done by Sharifi et al., (2018), they only reported "whether included assessment" as an independent variable.

A broad range of substantive and moderator variables have been reported from the previous meta-analysis studies on the effectiveness of ICT integrated L2 classrooms. It is important to retrieve useful and consistent information across previous studies to reflect a current scenario of features and effects of ICT used in L2 classrooms. However, results

from meta-analysis studies were found inconsistent in terms of ICT inclusion and moderator variables. In the next section, an updated meta-analysis study proposed to address these issues.

2.7. Research rationale and research questions

The current meta-analysis study aims to address the research gaps by proposing a meta-analysis study of the effectiveness of ICT integrated L2 classrooms. Previous meta-analysis studies on this topic have provided abundant evidence on both the effectiveness of CALL and other types of ICT used for L2 learning. However, diversity in contexts and overlaps of results from previous studies require future studies to iterate a similar evaluation of ICT integrated L2 with a trustworthy and consistent inclusion and coding strategy. In this section, research questions of the current study are proposed based on research gaps identified throughout the literature review.

The effectiveness of ICT integrated L2 classrooms has been reported in previous meta-analysis studies. However, it is confusing and sometimes intimidating for future teachers and researchers to decide what to use, and how to use ICT in the L2 classrooms. There are two practical issues that need to be addressed in the domain. First, incorporating emerging ICT demands L2 teachers adapt teaching paradigms and materials in their practice. Second, SLA researchers strive to provide systematic evidence of the effects of emerging ICT integrated into L2 classrooms.

The evaluation of various types of ICT integrated into L2 classrooms and the evidence of their effectiveness are still rather rare and inconsistent (Liou & Lin, 2017).

Issues regarding the current ICT integrated L2 classrooms are identified as: (1) an up-to-date effects of ICT integrated L2 classrooms; (2) multiple outcomes reported as targeted L2 skills; (3) inconsistent moderator variables reported from primary research; and (4) trustworthiness and quality of primary studies.

Accordingly, the current study aims to answer the following research questions (RQs):

1. What is the direction and magnitude of the pooled effect size of ICT integrated L2 classrooms?

Three further RQs are proposed regarding moderator variables to ICT integrated L2 classrooms:

2. How do different types of ICT moderate the effectiveness of ICT integrated L2 classrooms?
3. How do different targeted L2 skills as outcome measures moderate the effectiveness of ICT integrated L2 classrooms?
4. What are the moderator variables and their impacts on the effectiveness of ICT integrated L2 classrooms?

RQs 1 to 4 aim to provide an up-to-date summary of what research to date has shown about the effectiveness of general ICT and types of ICTs used through L2 classroom. RQs 1 to 4 are addressed by synthesizing the relevant empirical studies published between 2016 and 2020.

Lastly, the current study also address potential issues regarding reporting transparency and publication bias. Hence this study also asks:

5. To what extent have included studies represented the research field?

RQ 5 is addressed through an assessment of publication bias of studies examined in the current study.

In summary, the current study aims to contribute to the current knowledge about fundamental questions of ICT integrated L2 and illuminate potential variables that might moderate the effectiveness of ICTs usage. Results of quantitative analysis would also be established based on SLA theories and L2-instruction models for pedagogical interpretations. The assessment of publication bias would also provide insights into areas in need of further empirical attention.

2.8. Chapter Summary

This chapter started with clarification and definitions of terms in SLA, L2 classroom, L2 skills and ICT which are used throughout the current meta-analysis. Following the definition of key terms, the literature review first introduced critical SLA theoretical frameworks to evaluate research and current practices in ICT integrated L2 classrooms. Then a chronological review of ICT development and integration in the L2 classroom indicated the need for meta-analysis studies in the research field. A literature review of previous meta-analysis studies on the subject matter was provided under three themes: ICTs integrated L2 classrooms, ICTs types, and targeted L2 skills. The review indicated that an updated meta-analysis including a broad spectrum of ICTs used in L2 classroom might assist in closing current research gaps in the field. This chapter also

provided a systematic synthesis of contradictory moderators from previous meta-analysis in the field. This study proposed 5 RQs to further illuminate the effectiveness of ICT integrated L2 classrooms. The next chapter explains the rationale and method of meta-analysis in response to identified RQs.

Chapter 3. Method

The objective of this chapter is to describe the meta-analysis method with transparency and limitations in mind to further illuminate the effectiveness of ICT integrated L2 classrooms. This chapter is composed of two sections. First, a review and critique of the meta-analysis - the method used to answer proposed RQs on ICT integrated L2 classrooms - are provided. Advantages and challenges of the meta-analysis are discussed. Second, the series of meta-analytical procedures used in the current study are described. Procedures of topic formulation, study search, study eligibility, coding, and data analysis are described in detail. Calculation steps and computational packages are provided for transparency and replication purposes.

3.1. Rationale of a meta-analysis study

The rationale of a meta-analysis study is typically based on arguments for conducting a systematic secondary synthesis (SSS) which provides a synthesis of available evidence on the RQs through collecting, reviewing and analyzing data from primary studies. SSS research pool evidence from primary studies with rigorous and replicable procedures to produce verifiable findings (Norris & Ortega, 2006). It is particularly popular in SLA research because of its sensitivity to answering unclear or contradicted issues in primary studies. In the following section, basic concepts of a meta-analysis and its advantage compared to other SSS methods are provided.

3.1.1. Rationale for synthesis

A growing number of empirical studies support the positive effects of ICT integrated L2 classrooms (Norris & Ortega, 2000; Otto, 2017). Researchers raised concerns about the complexity and trustworthiness of claims made in primary studies. For example, idiosyncrasies in research design, and limitations in sampling from individual studies might cause biased results due to chance variability (Cooper, 1998; Light & Pillemer, 1984; Taveggia, 1974). In addition, it might further lead to biased interpretations and expectations of research results due to a high degree of objectiveness, or even limitations, from the readers' point of view, and thus hinder the development of a research domain (Norris & Ortega, 2000).

SSS can be used to aggregate primary results when evaluating the effectiveness of the use of ICT (Littell et al., 2008; Means et al., 2013). Previous SSS research has examined the application of different types of ICT and various research contexts in which ICT is being used in L2 classrooms (e.g., Plonsky & Zielger, 2016; Sharifi et al., 2018). Thus, SSS is an appropriate approach to examine cumulative knowledge and summarize previous knowledge and indicate future research directions regarding the effectiveness of using ICT in L2 classrooms.

3.1.2. Types of synthesis

SSS research can be used to minimize some sources of potential publication biases as well as maximize data by collecting all the relevant, available evidence on the subject matter (Littell et al., 2008). There are two major types of SSS research: systematic review and meta-analysis to assess the effectiveness of interventions.

Although some secondary researchers have used systematic review and meta-analysis interchangeably, they represent different approaches. A systematic review aims to search, synthesize and use available information across studies, providing a clear narrative of what the current best practice should be on a given topic. Meta-analysis combines quantitative results mathematically from multiple studies and focuses on identifying directions and magnitudes of intervention effects through a set of statistical methods.

Although both synthesis types provide insights on the state of research efforts and findings, a systematic review (either a narrative approach or a vote-counting method) has serious limitations. First, conclusions drawn from narrative reviews are often subjective and might be inconsistent from other separate reviews. The evaluating criteria based on the chosen scope of a narrative reviewer might cause incomplete or biased sampling of relevant primary studies, and thus causing a lack of trustworthiness in the results. In addition, the narrative approach suffers from critiques of its ad-hoc nature, meaning conclusions of reviews are based on conclusions of primary studies and might have little to do with empirical data (Rosenthal, 1991). Second, a systematic review might misinterpret potential statistical relationships in primary studies. For example, the vote-counting method – a common systematic review approach - utilizes empirical data in the form of statistical significance to "cast a vote" as a key component of to mitigate the subjectivity of the narrative approach. However, it still may result in inaccurate interpretations of the accumulative evidence because statistical significance is sensitive to sample size within primary studies (Borenstein et al., 2011). Third, systematic reviews are limited in providing the magnitude of effects observed within a group of primary

studies, and thus might fail to prove statistical trustworthiness of the overall findings (Cooper, 1998; Hedges & Olkin, 2014).

In contrast to systematic review, meta-analysis enables an estimation of the magnitude of observed findings or effects on a common scale, which could be combined and standardized as a mean effect size. It provides an accurate summary of actual retrieved effects from primary studies to evaluate statistical trustworthiness (e.g., standard errors and confidence intervals). Interpretations of meta-analytical results are also less sensitive to small sample size studies (Shadish & Haddock, 1994). Hence, meta-analysis is the method of choice for the current study to provide an updated summary of cumulative knowledge on the effectiveness of ICT integrated L2 classrooms.

3.2. Challenges of a meta-analysis study

Meta-analysis, a quantitative research synthesis method, is one of the SSS methods adopted from psychology to evaluate the effectiveness of instructions across various research contexts (Lin et al., 2018). It produces a summary of empirical knowledge on the subject matter for future research reference (Glass et al., 1981). However, meta-analysis is not free from challenges as interpretation and presentation of knowledge is situated and perspectival. In this section, three challenges regarding the statistical method, study heterogeneity and study quality are discussed along with detailed solutions.

3.2.1. Statistical method: Rationale for a random-effects model

Meta-analysis quantifies the directions and magnitudes of effectiveness in terms of effect sizes. Calculation of effect sizes in meta-analysis is based upon the assumptions that the researchers hold for the true effect size across synthesized studies (Borenstein et al., 2011). There are two statistical models to cope with the variation of effect sizes: the fixed-effect model and the random-effects model. The fixed-effect model requires the assumption that the calculated effect size in all studies is homogeneous. This means that participants and experiments are considered similar at the baseline before instructions and experiments. Thus, the fixed-effect model assumes that experimental conditions are the only reason to explain any within-group variance (Hedges, 1994). In contrast, the random-effects model assumes that not only do experimental conditions explain between-groups heterogeneity but other sources of between-groups differences might also contribute heterogeneity. It treats calculated effect sizes separately from individual studies allowing other variables to potentially moderate results of experiments. Thus, when pooling the mean effect size, the random-effects model would provide an analysis of between-groups variance through a test of heterogeneity (Pigott, 2012).

The choice of model mainly depends on the type of data retrieved from the studies. The current researcher aims to provide an overview of a range of studies using ICT tools to facilitate L2 learning. Identified studies used in this meta-analysis would be performed by individual researchers independently, and the subjects, contexts, and interventions would be different from one to another. Therefore, we assume calculated effect sizes on the topic vary across studies. Hence, a random-effects model was adopted for the research design.

3.2.2. Study heterogeneity: On the apples-and-oranges problems

Another challenge in meta-analysis is between-studies variance. More precisely, the challenge is whether included studies would be "similar enough" to be included in a meta-analysis and to be statistically combined. This challenge is commonly expressed in meta-analysis studies as the *apples and oranges problem* indicating low inferential strength due to high variabilities (or similarity) within included studies. Research of ICT integrated L2 classrooms is complex in substantive features and methodological conditions (Norris & Ortega, 2006). The current researcher assumed that between-groups heterogeneity is not limited to differences in substantive variables, theoretical and methodological conditions would define a population and contribute to between-groups heterogeneity as well. However, the assumption of a heterogeneous sample population at the meta-analysis level might lead to high study variability, and thus reduce the strength of meta-analysis results. We consider two *apples and oranges* problems in the following part: experimental/quasi-experimental designs and research diversity.

Experimental or quasi-experimental designs. A great deal of attention with ICT integrated L2 classroom research has been given to research design. Conventional research designs include experimental and quasi-experimental design (Hudson & Llosa, 2015). An experimental design is crafted with adequate controls for potential sources of variability between treatment and control groups. To ensure the control of variability, the key requirement is the randomness in selection and group assignment. Thus the variability within participants would be randomly distributed and controlled between two groups (Cook & Campbell, 1979). For studies using an experimental design, effect size

would be calculated based on the standardized mean difference between treatment groups and a control group at the post-test.

On the other hand, due to the limitation of resources and convenience of sampling, a quasi-experimental design is also commonly used in SLA research (Hudson & Llosa, 2015; Ziegler et al., 2017). A quasi-experimental design assumes that random assignment is not achieved between treatment and control groups in the experiment. The standardized mean differences would be statistically biased due to uncontrolled sources of variability, thus not inferentially comparable with between-groups estimates. Therefore, quasi-experimental studies should be treated separately from standardized mean difference when pooling effect sizes (Lipsey & Wilson, 2001).

The experimental design is considered to be more reliable from previous meta-analysis studies on the effectiveness of ICTs integrated L2 classroom (e.g., Grgurović et al., 2013; Sharifi et al., 2018). Results of previous meta-analysis further argued that the experimental design could provide more rigorous inferences regarding the effect of instruction. Hence, the current researcher treated experimental design as a critical criterion when screening primary studies.

Research diversity. In addition to research designs, other between-study differences might also contribute to unexplained heterogeneity. The challenge of a meta-analysis is the diversity in research conducted within ICTs integrated L2 classroom because researchers may pursue similar questions from distinct epistemological stances. The divergence in research contexts and topics make the synthesis process unclear by weaving all into a single meta-narrative. To provide an accurate depiction of the results

contained in the actual available research, this meta-analysis also coded and analyzed differences between included studies in terms of subgroups to investigate potential moderator variables that might influence learning outcomes. In addition to substantive features of included studies, subgroups were combined and contrasted based on methodological and reporting features. The mean effect sizes obtained from subgroups of studies were calculated using the same approach as for the overall analysis.

In summary, the researcher adopted two approaches to ensure study validity: experimental design and subgroup analysis. A subgroup analysis across included studies as well as a heterogeneity test of between-studies differences were provided to explain potential variation within the pooled mean effect size.

3.2.3. Study Quality: On the garbage-in-garbage-out problem

The third challenge of a meta-analysis study is the quality of included studies, known as the garbage-in-garbage-out problem. If a meta-analysis were based on flawed or biased empirical data, it would yield erroneous interpretations of cumulative knowledge (Norris & Ortega, 2006). The garbage-in-garbage-out problem is particularly prevalent in SLA research as it is an interdisciplinary field with multiple epistemological and ontological approaches (Smith, 2017). Hence, it is critical to set a threshold of methodological quality when evaluating and synthesizing empirical evidence. Classical normal distribution meta-analysis assumes methodological quality as *a posteriori* question, meaning that it is an all-inclusive approach when synthesizing results from primary studies (Glass et al. 1981). Compelling arguments of including relevant studies then coding them for methodological features are well documented (e.g., Iyengar &

Greenhouse, 2009; Lipsey & Wilson, 1993; Norris & Ortega, 2006). This study followed suggestions by Lipsey and Wilson (1993) to assess the impact of primary study quality as part of the meta-analysis.

Publication bias is the primary concern of included study quality in a meta-analysis study. Journal article publication status is considered as a guarantee of study quality while unpublished studies are generally considered to be of lower quality. However, this standard might cause synthesized evidence to be biased in favour of statistically positive results as studies with positive results were more likely to be published. Exclusion of unpublished studies would also inflate the pooled mean effect size from included studies therefore threatening the trustworthiness of a meta-analysis study. Hence, this study adopted *a posteriori* protocol and included unpublished dissertations in the literature searching results to tackle potential publication bias.

In summary, critical challenges of the current meta-analysis are statistical model, study heterogeneity, and publication bias. The current researcher adopted a random effect size model to synthesize studies with an experimental design. Subgroup analysis is also included to further explore potential and uncovered variables which may influence learning outcomes in L2 classrooms when using ICT. Finally, publication bias of included studies is assessed to clarify the trustworthiness.

3.3. Meta-analysis Procedures

The current study followed suggestions from Meta-Analysis Reporting Standards (MARS) established by the American Psychological Association (APA) Publications and Communications Board for meta-analysis procedures (Appelbaum et al., 2018). To ensure organized, transparent, and replicable meta-analysis procedures (Cooper et al., 2019), procedures of this meta-analysis include: topic formulation, study search, coding, and data analysis.

3.3.1. Topic formulation

The first step of this meta-analysis is to identify a set of specific, answerable research questions (RQs). As discussed in the literature review, the current meta-analysis focused on the characterization and effectiveness of ICT integrated L2 classrooms. The researcher first reviewed previous meta-analyses on the effectiveness of using ICT in L2 classrooms, then identified the RQs as:

- What is the direction and magnitude of the pooled effect size of ICT integrated L2 classrooms?
- How do different types of ICT moderate the effectiveness of ICT integrated L2 classrooms?
- How do different targeted L2 skills as outcome measures moderate the effectiveness of ICT integrated L2 classrooms?
- What are the moderator variables and their impacts on the effectiveness of ICT integrated L2 classrooms?

- To what extent have included studies represent the research field?

3.3.2. Study search

The goal of the study search is to undercover a group of primary studies that would answer the RQs through a systematic search of related literature (Littell et al., 2008). The researcher searched for current available studies (2016 - 2020) to provide updated evidence on the effectiveness of ICT integrated L2 classrooms in three steps.

First, keywords that would identify relevant studies were proposed and grouped into three categories:

- ICT related keywords (e.g., informational and communication technology (ICT), e-learning, educational technology, computer-assisted language learning (CALL), computer-assisted instruction (CAI), blended learning, e-book, data-driven learning (DDL), computer-mediated communication (CMC), digital game-based language learning (DGBL), intelligent CALL (ICALL), intelligent tutor, natural language processing (NLP), learning management system (LMS), mobile-assisted language learning (MALL);
- Learning-related keywords (e.g., teaching, learning, training, intervention, instruction, instructional treatment, classroom, program, implementation);
- Language-related keywords (e.g., language learning, second language (L2), second language acquisition (SLA), English as foreign language (EFL), English as second language (ESL);

Second, based on keywords identified in relevant studies, the researcher developed search strings using Boolean operators to identify relevant studies in an electronic database according to title of the studies and abstracts (see Appendix C).

Third, the researcher conducted an electronic database search using the search string to identify related studies in digital databases including Academic Search Premier, ERIC (EBSCOhost), PsycINFO, JSTOR, Social Sciences Abstracts, Google Scholar, ProQuest Academic Complete, ScienceDirect, Web of Science, Dissertation and Theses (ProQuest), and Dissertation and Theses @UVic.

3.3.3. Study eligibility

After the literature search, the researchers proceeded to formulate an explicit study eligibility criterion (see Appendix A). A clearly demonstrated inclusion criteria at the outset is requested in the method section of a meta-analysis according to MARS (Appelbaum et al., 2018). It is critical to capture a complete universe of studies focusing on ICT integrated L2 classrooms. The researcher developed a written protocol based on a review of previous meta-analysis studies as well as theoretical works. The protocol stated below specifies explicit inclusion and exclusion criteria which reflected the study RQs to indicate which studies would be included in the data analysis.

To be included in the current meta-analysis, a study had to meet all the following inclusion criteria:

1. The study used a strict experimental design implying random selection and assignment of participants.

2. The study included adequate reports of independent variables including (a) at least one type of ICT tool in L2 intervention; (b) targeted at specific language skills; (c) included treatment and control groups.
3. The study included adequate reports of dependent variables, measured participants' performance on objective language testing instruments using the same quantitative tests for treatment and control groups.
4. The study included primary studies published between 2016 and 2020. Studies published after May 2020 were not available in published form at the time of searching.

Searched studies would be excluded from the data collection if any of the exclusion criteria were met, including

1. The study did not provide sufficient (missing any of the statistics including sample size, mean of the measured outcome, and corresponding standard deviations) statistical information to support effect size calculation.
2. The study was primarily descriptive, focusing on the implementation of ICT in L2 classrooms.
3. The study did not incorporate a placebo/control group (in experimental design studies).
4. Dependent variables in the study did not measure L2 skills as part of the learning outcomes of instructional treatments.

5. The study was a conference proceeding publication or technical report that lacks of rigorous methodology, formal theoretical design or peer review.

3.3.4. Coding

Following study eligibility screening, included studies were coded to summarize study features. Included study features are summarized as substantive features, methodological features and reporting features.

Substantive features. Substantive features reflect theoretical motivations and answer RQs related to the overall effectiveness of types of ICT used in different L2 classrooms. Previously validated variables and potential moderator variables of instructional effects were coded and grouped under substantive features. All included studies were coded under substantive features including:

1. Type of ICT. This is a categorical independent variable coded for contextual variables based on the type of ICT used for the treatment group. Categories were derived from a literature review of previous meta-analysis studies (see e.g., Golonka et al., 2014; Grgurović et al., 2013; Plonsky & Ziegler, 2016), including (a) “CAI”; (b) “DDL”; (c) “CMC”; (d) “DGBL”; (e) “ICALL”; (f) “MALL”.
2. Targeted L2 skills. This is a categorical independent variable coded for contextual variables based on types of L2 skills being addressed in ICT integrated L2 classrooms. Categories were consistent with suggestions in previous theoretical work on L2 skills moderated through ICTs integrated L2

instruction (Sauro & Chapelle, 2017), including (a) L2 grammar; (b) L2 vocabulary; (c) L2 reading; (d) L2 writing, (e) L2 speaking; (f) L2 pragmatics; (g) mixed.

3. L2 context. This is a categorical independent variable that coded for contextual variables based L2 classroom environment when the instructional treatments applied. Categories were retrieved by researchers based on included studies, and summarized as (a) ESL (English as a second language); (b) EFL, (English as a foreign language); (c) German; (d) Chinese; (e) Dutch.
4. Educational level. This is a categorical independent variable that was coded for participants variables based on educational situations at the time of receiving instructional treatments. Categories were consistent with previous meta-analysis studies (see e.g. Sharifi et al., 2018) on this topic, including (a) post-secondary; (b) secondary.
5. L2 language proficiency. This is a categorical independent variable that coded for participants variables based on participants' L2 proficiency before instructional treatments. Data were either self-reported by participants from individual study or indicated through a pre-test provided in individual study. Categories were consistent with previous meta-analysis studies (e.g., Grgurović et al., 2013), including (a) beginner level; (b) intermediate level; (c) advanced level; (d) missing.
6. L1 background. This is a categorical variable that coded for participants features based on their first language background (L1) self-reported by

participants from individual study. Categories were retrieved by researchers based on included studies, and summarized as (a) Mixed; (b) Arabic; (c) English; (d) Chinese; (e) Korean; (f) Persian; (g) Spanish; (h) Portuguese; (i) Norwegian; (j) Turkish.

Methodological features. All included studies were also coded for methodological features to provide an updated overview of methods and study characteristics. Methodological features from included studies would also indicate potential moderating effects of non-instructional variables, hence, potentially indicate desirable characteristics for future studies (Lipsey, 1994). Methodological features coded in the current meta-analysis include:

(1) Treatment duration. This is an interval value that is converted to a categorical variable for analysis purposes. Conversion and categorical decisions were based on distribution observed in retrieved data from included studies, including: (a) short (less than 1.25 week); (b) medium (1.25 week to 13.5 weeks); (c) long (more than 13.5 weeks); (d) missing.

(2) Measured outcomes. This is a categorical dependent variable that coded for the type of assessment individual study reported for learning outcomes. Categories were consistent with Norris and Ortega's (2000) suggestion, including (a) meta-linguistic judgement; (b) selected-response; (c) constrained-constricted-response; (d) free-constructed-response; (e) mixed.

Reporting features. Lastly, the researcher coded reporting features of included studies that are of the direct interest of primary studies but meta-linguistic judgement

superimposed across included studies. These features adopted from previous practical considerations and transparency concerns regarding ICTs related to SLA research. Categories are grounded in the actual features as reported, including

(1) Publication year. This is a categorical variable. Categories of publication year include (a) 2016; (b) 2017; (c) 2018; (d) 2019; (e) 2020.

(2) Sample size. This is an interval value that is converted to a categorical variable for analysis purposes. Conversion and categorical were based on distribution of observed samples in retrieved data from included studies, including (a) small (less than 45 participants); (b) medium (45 – 80 participants); large (more than 80 participants).

3.3.5. Data analysis

The fundamental goal of data analysis in the current meta-analysis is to investigate the magnitude of effects and relationships of ICT used in L2 classrooms across synthesized empirical data. Steps of data analysis consist of (1) data extraction; (2) effect size calculation; (3) bias assessment; (4) subgroup analysis.

Data extraction. There are four critical features in calculating and understanding effect size estimates as they are extracted from included studies. First, the post-test values from one dependent variable were collected to calculate effect size estimates (Hedges' *g* in this meta-analysis). If more than one outcome were reported in the individual study based on different groups of participants, they would contribute to separate effect size estimates. If outcomes reported in the individual study were based on the same group of

participants, then the calculated effect size is the mean of all effect sizes. Second, the mean difference (served as the numerator in effect size equation) was collected based on the difference between a treatment group and a control group from each study. Third, the sample size of treatment groups, the control group and the total participants were extracted for effect size calculation as well as coding for study characteristics. Fourth, the standard deviation was collected from each experimental group and control group (served as the denominator in effect size equation) to standardize the mean difference across included studies.

Effect size calculation. Effect size is a measure of strength and direction of a relationship between variables, which could be influenced by the research design of the individual study and the format of reported data. Different effect size indices are used for dichotomous, continuous or correlational data based on the nature of the variables. Common effect size indices are Cohen's *d*, Hedges'*g*, Pearson's *r* among others (Borenstein et al., 2011). Different effect size estimates make sense when interpreting the kinds of relationship observed from included studies, although most of the statistics can be transformed to allow comparison (e.g., Rosnow et al., 2000).

Included studies on the topic of ICT integrated L2 classrooms commonly focused on differences between treatment groups and the control group, which would be considered as continuous data (Borenstein et al., 2009). There are two steps of effect size calculation including individual effect size as well as adjusting and aggregating effect sizes. The researcher adopted *Hedges' g* as the common form of the calculation of pooled effect size. The actual analysis was conducted in R environment using R package *meta,*

metafor, dmetar, and *esc* (Schwarzer, 2020; Viechtbauer, 2020; Harrer *et al.*, 2019; Lüdecke, 2019).

Individual effect size. First, we calculated the common effect size Cohen's *d* to explain the difference between comparisons using equations introduced by Borenstein et al. (2011):

$$d = \frac{\bar{x}_1 - \overline{x_2}}{S_{within}} \quad (1)$$

where *Swithin* is the standard deviation of the difference between two comparing groups given by the following:

$$S_{within} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} \quad (2)$$

The variance of *d* is given by

$$V_d = \frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)} \quad (3)$$

and the standard error of *d* is

$$SE_d = \sqrt{V_d} \quad (4)$$

To correct for small sample bias, researcher calculated Hedge's correction factor *J* using

$$J = 1 - \frac{3}{4df-1} \quad (5)$$

Second, we calculate Hedge's *g* based on the common effect size for the purpose of small sample size adjustments, given by

$$g = J \times d \tag{6}$$

Results of Hedge's *g* were used to explain the standardized mean difference and represent the individual effect size. Values calculated from procedures above are reported and also used for subsequent meta-analysis and subgroups analysis.

Adjusting and aggregating effect sizes. The current researcher also aims to provide an estimation of a pooled effect size from included studies on the same topic while obtaining better parameter estimates. Hence, after data extraction and calculation, the current researcher calculated a pooled effect size by combining individual effect sizes.

It is common to adjust individual effect sizes when effect sizes were extracted from primary studies with differing sample sizes (Norris & Ortega, 2006). A well-recognized approach of adjusting effect sizes is to mathematically assign heavier "weight" to effect sizes that come from larger sample studies (Borenstein et al., 2009). Weighted effect size from an individual study is given by

$$W_i^* = \frac{1}{V_{Y_i}^*} \tag{7}$$

where

$$V_{Y_i}^* = V_{Y_i} + T^2$$

(T^2 is a measure of heterogeneity)

$$V_{Y_i} = V_d \tag{8}$$

As discussed in the previous section, the researcher assumed that the smaple population from included studies is heterogeneous in experimental conditions, and thus used a random-effects model to estimate the weighted mean effect size from included studies. To report the weighted mean effect size, the researcher used generic M* to represent mean effect sizes using Hedge's *g*, given as

$$g^* = \mathrm{M}^* \tag{9}$$

$$\mathrm{M}^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*} \tag{10}$$

where *i* represents for the case of individual weighted mean effect size.

Testing heterogeneity. The test of heterogeneity explains the variation of weighted mean effect size which allows the researcher to judge whether the calculated effect is based on a single population due to sampling errors. In addition, it is suggested in MARS that included studies in a meta-analysis should have a degree of homogeneity so that it is legitimate to synthesize effect sizes (Appelbaum et al., 2018). The current researcher adopted three steps to report the test heterogeneity of pooled effect sizes:

(1) the current researcher computed the ratio of observed variation between observed effect sizes and the fixed effect model estimate, Q, given as

$$Q = \sum_{i=1}^{k} W_i (Y_i - M)^2 \tag{11}$$

(2) then the between-groups variance estimator T^2 was computed, given by

$$T^2 = \frac{Q - df}{C} \tag{12}$$

where

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i} \tag{13}$$

in order to estimate the "true" variance of effect sizes, bringing the ratio on the same scale as the effect size metric;

(3) and lastly the current researcher reported the proportion of dispersion between studies, given by

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100\% \tag{14}$$

Subgroup analysis. In addition to exploring between-groups heterogeneity, the researcher inquired upon potential moderator variables in the domain of studies. Thus, a subgroup analysis was conducted by combining and contrasting coded study features suggested by previous meta-analysis findings. The number of subgroups is limited to 10 according to the coding protocol developed during data collection phase. Standardized mean differences were calculated according to 10 subgroups to investigate the relationship of effect sizes between subgroups and potential impacts of moderator variables (coded variables) to learning outcomes. Calculation procedures of standardized mean differences are identical to those above for all included studies.

Publication bias. Publication bias refers to the issue that included studies in a meta-analysis are a biased sample of relevant studies, thus the reported mean effect would also reflect the bias. It is a vital threat to the validity of a meta-analysis specifically when studies with statistically significant results are more likely to be published, cited and included in a meta-analysis than those are not statistically significant (Dickersin, 2006). The researcher followed two suggestions (see e.g. Littell et al., 2008) to cope with potential bias. First, grey literature was considered (e.g., unpublished dissertation) in the selection of publication for meta-analysis. Second, a funnel plot (Light & Pillemer, 1984; Littell et al., 2008) is reported to visualize potential publication bias. Finally, Egger's test (Egger et al., 1997) and Duval and Tweedie's trim-and-fill procedure (Duval & Tweedie, 2000) are also used to quantify potential publication bias.

In summary, this meta-analysis inquires into the effectiveness of using types of ICT in L2 classrooms. First, an exhaustive bibliography search was conducted in multiple electronic database to identify experimental studies using types of ICT to assist targeted L2 skills. Then study eligibility screening and coding were conducted to select and categorize included studies based on substantive features, methodological features and reporting features. Data analysis was conducted in R environment using random-effects model. Test of heterogeneity and a subgroups analysis were also conducted. Lastly, potential publication bias was assessed through the funnel plot, Egger's test, and Duval and Tweedie's trim-and-fill procedures.

3.4. Chapter summary

This chapter introduced meta-analysis as a research method to explore the effectiveness of ICT integrated L2 classrooms. Beginning with a general review of systematic secondary synthesis methods, the first part of this chapter provided arguments, challenges and designs for using meta-analysis to synthesize cumulative knowledge. After stating rationales and challenges, this chapter proceeded to state a detailed description of meta-analytical procedures including topic formulation, study search, study eligibility, coding, and data analysis. The next chapter reports the findings of all meta-analysis procedures.

Chapter 4. Findings

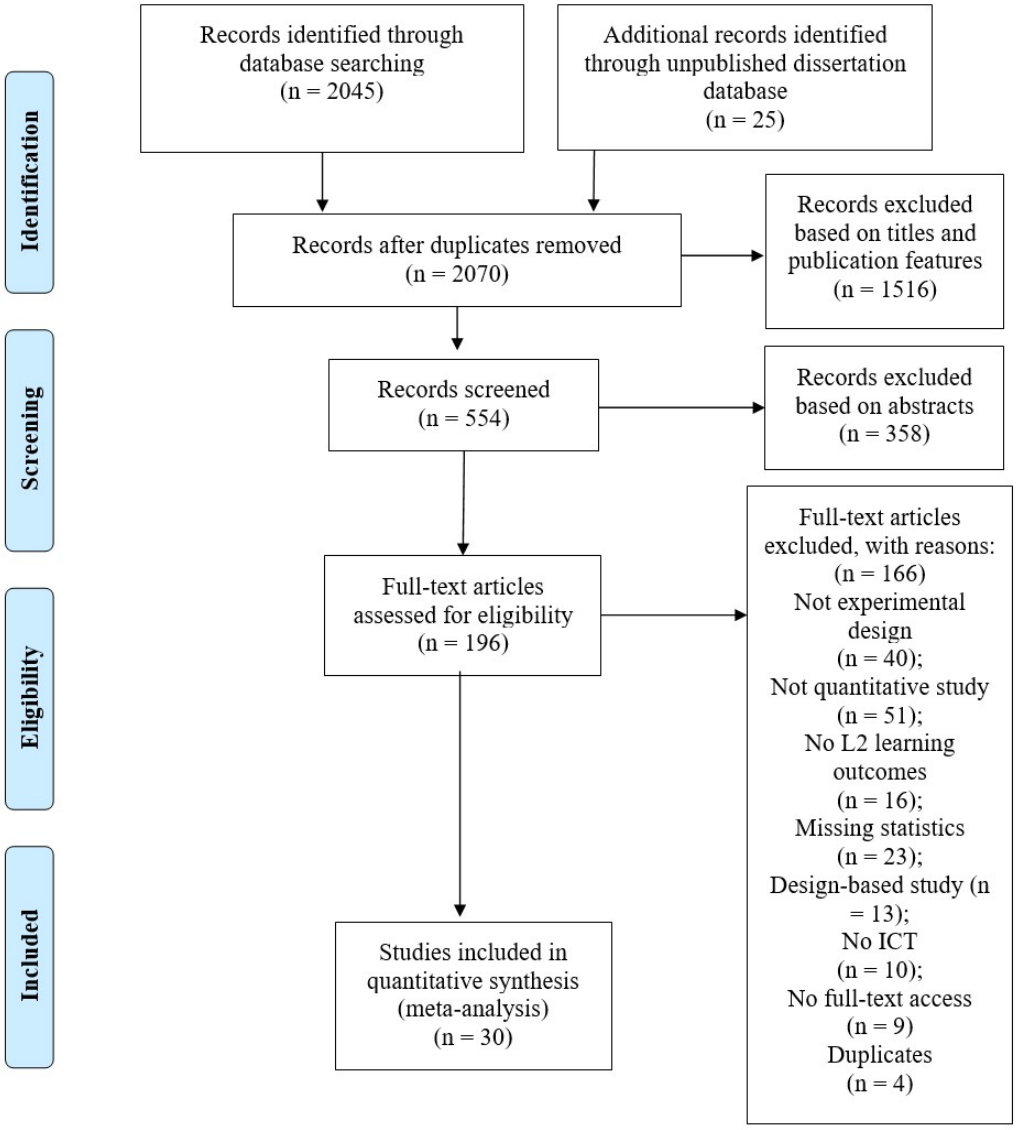
Findings of the current meta-analysis are provided in this chapter in four sections: study selection, study characteristics, results of meta-analysis, and publication bias. Study selection provides procedures and results of study searching and screening based on the searching strategy and inclusion criteria described in Chapter 3. Then, in the study characteristics section, characteristics of independent variables, dependent variables and moderator variables of included studies are reported. Study characteristics are grouped and described as substantive features, methodological features and reporting features based on the coding strategy described in Chapter 3. After describing coding results of study characteristics, weighted mean effect sizes are reported in terms of comparison groups. Results of a random-effects model analysis including directions and magnitudes of effects, confidence intervals, and estimates of between-groups heterogeneity are reported and discussed. Results of subgroup analysis are described and discussed. Lastly, publication bias of the included studies are described and assessed.

4.1. Study selection

In this section, a detailed report of procedures and decisions of the literature search and screening are provided. Numbers of studies searched at each decision step as well as reasons for exclusion are reported. Figure 1 describes the results of the literature search and records decisions made in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Liberati et al., 2009).

Figure 1

PRISMA Flow Diagram of Study Selection Results



The literature search was conducted in May 2020, in the Summon database accessed through the Library at the University of Victoria. As shown in Figure 1, initial searching identified 2070 articles including journal articles (n = 2045) and dissertations (n = 25). These initial searching results were refined through publication features,

including “full-text”, “scholarly peer review” and published in “education” and “language & literacies” domains for quality control in the Summon database. Finally, the searching process identified 554 primary studies.

After the searching process, identified studies then were reviewed and screened for study eligibility. Studies at this stage were included for meta-analysis only if they used an experimental design to explore the use of ICT and reported L2 skills as learning outcomes. The screening process began with abstract elimination and proceeded to full-text elimination. Starting with 554 screened studies, 342 journal articles and 16 dissertations were excluded based on abstract elimination. A total of 196 studies were then moved through full-text eligibility screening. Appendix A provides the inclusion criterion that searched studies needed to meet to be considered for meta-analysis. Full-text eligibility screening followed the exclusion criteria described in Chapter 3. Searched studies were excluded from the meta-analysis if they did not report essential experimental details and statistics, or if full text was not available. Out of 196 studies identified for full-text screening, 166 studies were excluded for not meeting requirements (see Figure 1). After screening was complete, there were 30 studies coded and included for meta-analysis (see Appendix B).

In summary, there were 2070 studies identified during the literature identification. After duplication removal, title elimination and abstract elimination, there were 196 studies identified as searched studies. Lastly, 30 studies were identified as the included studies for meta-analysis through full-text eligibility screening.

4.2. Study characteristics

A cohort of 30 included studies published between January 2016 and December 2019 provided a total sample size of 2059 participants (mean: 68.6, SD: 36.8, sample size range: [10, 160]). Included studies also reported participant data in terms of comparison groups. There were 43 comparison groups reported across the included studies (sample mean: 52.51, SD: 27.99, sample size range: [10, 160]). In the following section, results of coding are reported to describe included studies characteristics. Following the coding strategies described in Chapter 3, study characteristics of the independent, dependent and moderator variables were grouped and explored across the included studies in three categories: substantive features, methodological features and reporting features.

The coding results of substantive features (see Appendix D1) included important characteristics of the participant and contextual variables across included studies (see Table 1).

Table 1

Coding Results of Substantive Features

Participant variables	Participants	Contextual variables	Participants
Educational level:		L2 context:	
Post-secondary	1801	EFL	1780
Secondary	258	Chinese	162
		ESL	50

L2 proficiency level:		German	36
Beginning	421	Dutch	31
Intermediate	1404		
Advanced	112	ICT type:	
Missing	122	MALL	788
		CAI	444
L1 background:		ICALL	426
Chinese	563	CMC	306
Arabic	292	DGBL	59
Spanish	293		
Persian	210	Targeted L2 skills:	
Korean	204	L2 reading	65
Mixed	145	L2 listening	160
English	134	L2 grammar	582
Turkish	88	L2 writing	362
Portuguese	72	L2 vocabulary	369
Norwegian	58	L2 speaking	301
		Mixed	180
		Pragmatics	40

In terms of participant variables, 1801 participants were reported as adult learners with post-secondary educational level and 258 young adult participants with secondary school educational level. The most common L2 proficiency level across studies is the intermediate level (n = 1404) followed by beginning level (n = 421). Advanced level participants were the least investigated across included studies (n = 112). Two included studies were coded as missing because they did not report learners' L2 proficiency level (n = 122). In addition to L2 proficiency, participants L1 backgrounds were also reported across included studies. Chinese language as L1 included the most participants (n = 563) followed by Spanish language (n = 293) and Arabic language (n = 292). Other types of L1 reported across included studies include Persian language (n = 210), Korean language (n = 204), English language (n = 134),Turkish language (n = 88), Portuguese language (n = 72), and Norwegian language (n = 58). Four included studies reported participants with mixed L1 backgrounds (n = 145).

Contextual variables include L2 contexts, ICT types and targeted L2 skills. EFL context were reported to be the most investigated L2 context (n = 1780) across included studies followed by Chinese as L2 context (n = 162). Other types of L2 contexts reported from individual studies include ESL (n = 50), German (n = 36), Dutch (n = 31). There were 5 types of ICT coded across included studies. MALL was found with the most participants across studies (n = 788) followed by CAI (n = 444), ICALL (n = 426), and CMC (n = 306). DGBL was only reported in one included study with 59 participants recruited. DDL was not reported in any of the included studies, and thus was dropped. In terms of targeted L2 skills, L2 grammar skill was reported to be the most widely investigated (n = 582) followed by L2 writing skill (n = 362) and L2 vocabulary skill (n =

369). Other types of L2 skills were also explored across studies including L2 speaking (n = 301), L2 listening (n = 160), L2 reading (n = 65), pragmatics (n = 40) and mixed skills (n = 180).

Methodological features (see Appendix D2) including dependent variables, such as outcome measures, and independent variables, such as treatment duration are presented in Table 2.

Table 2

Coding Results of Methodological Features

Dependent variable	Participants	Independent variable	Participants
Outcome measure:		Treatment duration:	
Meta-linguistic judgement	191	Short	464
Selected response	275	Medium	1102
Constrained constructed response	255	Long	433
Free constructed response	704		
Mixed	634		

Outcome measure that contained the most participants was free-constructed response measures (n = 704) followed by mixed outcome measures (n = 634). Other types of outcome measure reported across studies include selected-response (n = 275), constrained-constructed response (n = 255) and meta-linguistic judgement (n = 191).

Treatment duration ranges from less than 1 hour to 48 weeks with a median of treatment duration as 4 weeks. The cutoff points for short, medium and long treatments follows the first quartile and the third quartile of sample distribution from included studies. Medium length treatments (1.25 to 13.5 weeks) were investigated with the most participants across studies (n = 1102) followed by short length treatments (less than 1.25 weeks) (n = 464) and long length treatments (more than 13.5 weeks) (n = 433).

Table 3

Coding results of reporting features

Sample size	Participants	Studies	Publication year	Participants	Studies
Small	233	8	2016	571	9
Medium	730	12	2017	983	12
Large	1096	10	2018	137	2
			2019	368	7

Lastly, reporting features (see Appendix D3) including publication years and sample size were coded and explored across studies. Cut-off points of the small, medium and large groups were based on the first quartile and third quartile of the sample distribution. Table 3 presents the coding results of reporting features. 12 included studies were coded with a medium sample size (45 - 80 participants) for a total of 730 participants and 10 included studies were coded with a large sample size (with more than 80 participants each study) containing 1096 participants in total. The remaining 8

included studies were coded with a small sample size (less than 45 participants) with a total of 233 participants. Out of 30 included studies, 12 studies were published in 2017 reporting 983 participants, followed by 9 included studies published in 2016 reporting 571 participants. Although this meta-analysis searched for published studies before May 2020, none of the searched results qualified for inclusion criteria was published in 2020.

4.3. Results of Meta-analysis

This section reports the results of a meta-analysis including results of primary analysis, results of heterogeneity test and results of subgroup analysis. First, adjustments based on complex data structures are reported as well as decisions to pool multiple reports across studies. Then results of a random-effects model analysis are reported with a forest plot. Between-groups heterogeneity is then reported to assess the variation between comparison groups. Lastly, subgroup analysis based on substantive features, methodological features and reporting features are reported.

There are 30 primary studies included in the meta-analysis (see Appendix B). All primary studies included in the analysis are published, peer-reviewed, journal articles. Out of the 30 included studies, 7 studies reported results with more than one comparison group. Comparison groups reported from these studies range from 2 to 4 pairs. Thus, there were 43 comparisons coded based on reported experimental conditions. Each comparison was between the use of ICT and traditional face-to-face instruction in L2 classrooms. Among 43 comparison groups, 21 of them also reported multiple mean scores as L2 learning outcomes. The number of mean scores drawn from each comparison group ranged from 1 to 8 scores.

Some studies had complex data structures, therefore the number of effect sizes extracted from the included studies ranged from 1 to 24 effect sizes. Included studies reported findings in a complex data structure including multiple comparison groups and multiple reports for each comparison group. In such cases, the effect sizes from the same group of participants were averaged to create a mean effect size. For studies reporting with multiple treatment groups based on independent samples, effect sizes were not combined but treated separately and coded with individual sample IDs (Lipsey & Wilson, 2001; Plonsky & Oswald, 2015). In summary, the current meta-analysis included 43 comparison groups which yielded 106 effect sizes.

4.3.1. Results of primary analysis.

The pooled effect size and related statistics were calculated in terms of the comparison groups and presented in Table 4.

Table 4

Results of Weighted Mean Effect Size

Model	k	Standardized mean difference (SMD)	95% CI	z	P-value	95% PI (Prediction Interval)
All comparison groups combined	43	1.03	[0.84; 1.22]	10.63	p < 0.0001	[-0.23; 2.30]

Results of meta-analysis showed that the overall effectiveness of using ICT is 1.03, with 95% CI [0.84, 1.22]. The pooled effect size is also significant ($n = 43$, $z = 10.63$, $p < 0.0001$). Similar to the CI, PI (prediction interval) was also calculated based on the effect size scale to quantify the distribution of true effect sizes (Borenstein et al., 2011). The 95% PI in the random-effects model is [-0.23; 2.30], which is broad and stretched below zero, indicating that in 95% of cases the true effect size in a new study will fall inside this range. Thus, the positive overall effect should be interpreted with caution as negative effects are possible in future studies.

4.3.2. Test of Heterogeneity

Given the broad prediction interval, between-group heterogeneity was tested and presented in Table 5.

Table 5

Results of Test of Heterogeneity

Model	Q	d. f.	P-value	Tau-square	I-square
All comparison groups combined	2300.93	42	$p < 0.0001$	0.38	98.2%

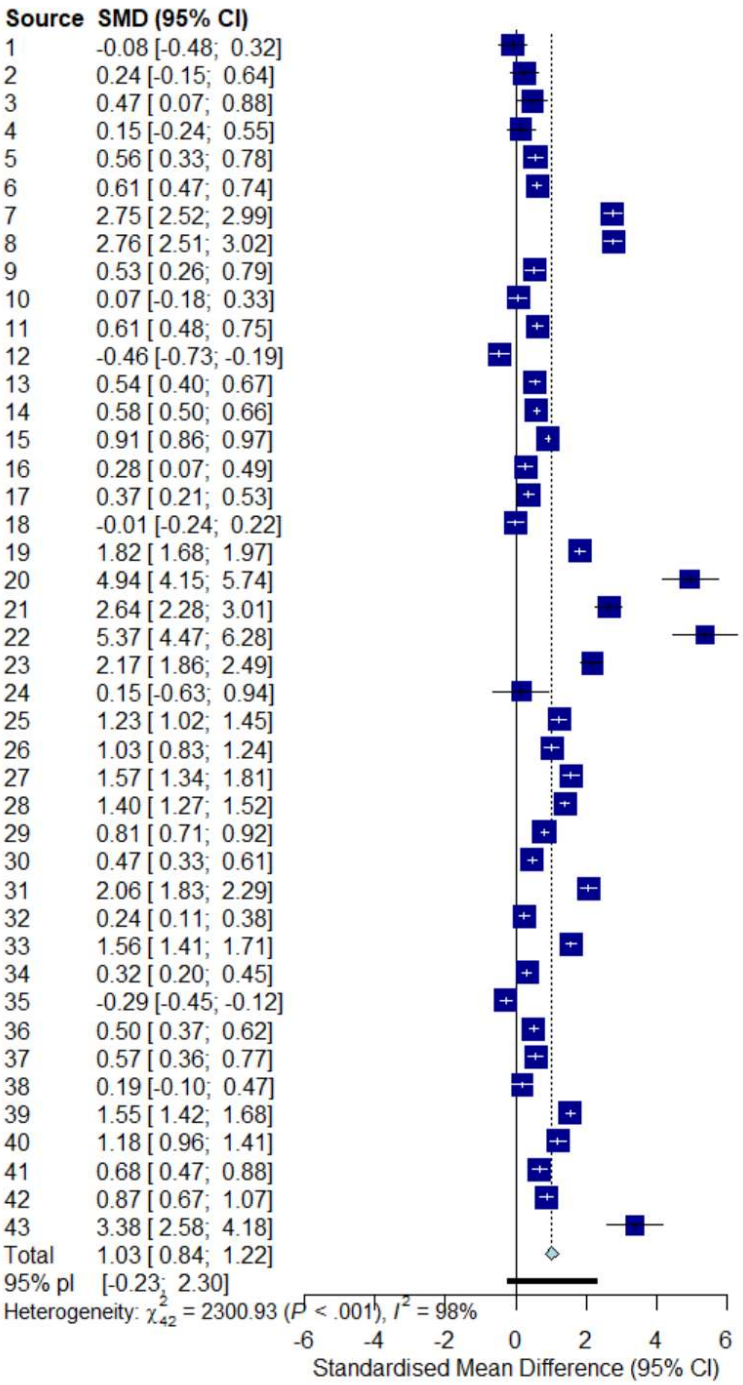
The results of this test showed substantive heterogeneity ($Q = 2300.93$, $p < 0.0001$) indicating significant between-groups heterogeneity among 43 comparison groups. More specifically, the between-group variance of the random-effects model is

0.38, with 95% CI [0.20, 0.77] in terms of the effect size scale. To understand the amount of variance on a relative scale, we also chose to report I-square, (98.2%, 95% CI [97.9%, 98.4%]), which reflects what proportion of the observed variance is real. The resuts indicated that the true between-group heterogeneity explained by sampling error is 1.8% and that 98.2% of the total variance observed across effect size estimates are the result of true heterogeneity.

To further investigate between-groups heterogeneity, a forest plot was presented in Figure 2. A forest plot is a graphical display of individual effect sizes as well as the pooled effect size depicted as point estimates bounded by confidence intervals.

Figure 2

Forest Plot of All Comparison Groups



Between-group heterogeneity could be caused by extreme effect sizes in the random-effects model which are considered as outliers. For example, as shown in Figure 2, Group ID 7, 8 ,12, 20, 21, 22, 43 are outside of the 95% PI indicating that they may potentially have a large impact on the weighted mean effect size. The researcher defined outliers as those for which either the upper bound of 95% CI is lower than the lower bound of the pooled effect size's 95% CI, or the lower bound of the 95% CI is higher than the upper bound of the pooled effect size's 95% CI. An adjusted model with outliers (data points exceeding the 95% CI of the pooled effect size) removed was provided to assess the impact of extreme effects. Table 6 presents results of the adjusted random-effects model with outliers excluded.

Table 6

Adjusted Results with Outliers Removed

Instruments	k	Standardized mean difference (SMD)	95% CI	z	p-value	Prediction interval (PI)
Comparison groups combined without outliers	9	0.90	[0.79; 1.02]	15.03	$p < 0.0001$	[0.54; 1.26]

If outliers are excluded, the adjusted pooled effect size is 0.90 with 95% CI [0.79; 1.02], which is still significant ($z = 15.03$, $p < 0.0001$). The prediction interval of the adjusted random-effects model is also narrower when outliers are excluded as [0.54,

1.26]. Comparing results from the adjusted model and the original model, it can be argued that outliers from included studies had a small impact on the overall effects. Hence, the researcher reports statistics of the original model as results and includes all data points for the following analysis.

4.3.3. Subgroup analysis

To further investigate heterogeneity within included studies, a subgroup analysis of substantive features across included studies is presented in Table 7.

Table 7

Results of Substantive Features' Subgroup Analysis

Substantive features	Q	d. f.	P-value	k	Standardized mean difference (SMD)	95% CI
ICT type	25.52	4	p < 0.05			
ICALL				12	0.74	[0.15; 1.33]
CAI				9	0.92	[0.61; 1.23]
CMC				7	2.16	[1.27; 3.06]
MALL				14	0.94	[0.69; 1.20]
DGBL				1	0.47	[0.33; 0.61]

Targeted L2 skill	77.24	7	p < 0.05			
L2 grammar				13	1.70	[1.11; 2.28]
L2 writing				7	1.05	[0.41; 1.69]
L2 speaking				7	0.75	[0.27; 1.22]
L2 listening				1	0.91	[0.86; 0.97]
L2 vocabulary				9	0.99	[0.66; 1.32]
L2 reading				1	0.50	[0.37; 0.62]
Pragmatics				2	-0.27	[-0.83; 0.29]
Mixed				3	0.51	[0.36; 0.67]
L2 Context	49.43	4	p < 0.05			
ESL				4	0.20	[-0.03; 0.42]
EFL				30	1.29	[1.06; 1.51]
German				1	0.56	[0.33; 0.79]
Dutch				2	0.30	[-0.14; 0.74]
Chinese				6	0.60	[-0.05; 1.26]
Educational Level	0.20	1	0.62			

Post-secondary				38	0.99	[0.81; 1.17]
Secondary				5	1.27	[0.05; 2.49]
L2 Proficiency	14.48	3	p < 0.05			
Beginner				13	0.76	[0.25; 1.28]
Intermediate				26	1.20	[0.99; 1.41]
Advanced				2	1.33	[-0.08; 2.74]
Missing				2	0.02	[-0.58; 0.62]
L1 background	128.39	9	p < 0.05			
Mixed				8	0.11	[-0.13; 0.34]
Arabic				5	1.01	[0.52; 1.52]
English				5	0.99	[0.60; 1.38]
Chinese				8	1.05	[0.72; 1.38]
Korean				3	1.96	[0.22; 3.70]
Persian				5	3.08	[1.61; 4.56]
Spanish				4	0.77	[0.12; 1.412
Portuguese				1	0.28	[0.07; 0.49]
Norwegian				1	-0.29	[-0.45; -0.12]

Turkish	3	0.91	[0.63; 1.19]

For contextual variables, CMC group showed the highest effect size (k = 7, ES = 2.16, 95% CI [1.27, 3.05]) among other types of ICT followed by MALL group (k = 14, ES = 0.94, 95% CI [0.69; 1.20]) and CAI group (k = 9, ES = 0.91, 95% CI [0.61; 1.23]). Tests of subgroup difference under random effects model showed significant between-groups heterogeneity (Q = 25.52, p < 0.05). Subgroup analysis on targeted L2 skills also showed significant between-groups heterogeneity (Q = 77.24, p < 0.05). L2 grammar showed the highest effect size (k = 13, ES = 1.69, 95% CI [1.11, 2.28]) followed by L2 writing (k = 7, ES = 1.05, 95% CI [0.41; 1.69]), L2 vocabulary (k = 9, ES = 0.99, 95% CI [0.66, 1.32]) and L2 listening (k =1, ES = 0.91, 95% CI [0.86; 0.97]. Pragmatics group showed harmful effects (k = 2, ES = -0.27, 95% CI [-0.83, 0.29]). Lastly, L2 context variables showed significant heterogeneity (Q = 49.43, p < 0.05) between subgroups. EFL context showed the highest effect size (k = 30, ES = 1.29, 95% CI [1.06; 1.51]) followed by Chinese group (k = 6, ES = 0.60, 95% CI [-0.05; 1.26]) and German group (k = 1, ES = 0.56, 95% CI [0.33; 0.79]).

For participants variables, there is no significant heterogeneity (Q = 0.20, p = 0.6) found between the post-secondary group (k = 38, ES = 0.99, 95% CI [0.81; 1.17]) and the secondary group (k = 5, ES = 1.27, 95% CI [0.05; 2.49]) while the secondary group showed higher mean effect size. L2 proficiency showed significant heterogeneity (Q = 14.48, p <0.05). Intermediate group showed the highest effect size (k = 26, ES = 1.1981, 95% CI [0.99; 1.41]) followed by beginner group (k = 13, ES = 0.76, 95% CI [0.25;

1.28]). Groups coded with missing information for L2 proficiency showed harmful effects (k = 2, ES = 0.02, 95% CI [-0.58; 0.62]). For learners L1 backgrounds, Persian group showed the highest effect size (k = 5, ES = 3.08, 95% CI [1.61; 4.56]) followed by Korean group (k = 3, ES = 1.96, 95% CI [0.21; 3.70]) and Chinese group (k = 8, ES = 1.05, 95% CI[0.72; 1.38]). Norwegian group showed harmful effects (k = 1, ES = -0.29, 95% CI [-0.45; -0.12]). Test of heterogeneity for subgroup differences is significant (Q = 122.19, p < 0.05).

In addition to examining substantive features above, subgroups coded based on methodological features were also explored and results are presented in Table 8.

Table 8

Results of Methodological Features' Subgroup Analysis

Methodological features	Q	d. f.	P-value	k	Standardized mean difference (SMD)	95% CI
Outcome measure	15.56	4	p < 0.05			
Meta-linguistic judgement				3	0.51	[-0.01; 1.04]
Selected response				6	2.58	[1.68; 3.49]

Constrained-constructed response				3	0.87	[0.28; 1.47]
Free-constructed response				13	0.93	[0.53; 1.32]
Mixed				18	0.82	[0.46; 1.17]
Treatment duration	31.22	3	$p < 0.05$			
Short				10	0.40	[0.13; 0.66]
Medium				25	1.30	[1.02; 1.58]
Long				7	1.17	[0.68; 1.65]
Missing				1	0.54	[0.40; 0;67]

For treatment duration, medium length treatment duration (2 – 14 weeks) showed the highest effect size (k = 25, ES = 1.30, 95% CI [1.02; 1.58]) followed by long treatment duration group (more than 14 weeks) (k = 7, ES = 1.17, 95% CI [0.69; 1.65]). Test of heterogeneity for subgroup differences is significant (Q = 31.22, $p < 0.05$). Lastly, measurement type showed significant heterogeneity (Q = 15.56, $p < 0.05$) for subgroup differences. Selected response test showed the highest effect size (k = 6, ES = 2.58, 95% CI [1.68; 3.49]) followed by constrained-constructed response test (k = 3, ES = 0.87, 95% CI [0.27; 1.47]), free-constructed response test (k = 13, ES = 0.87, 95% CI [0.28; 1.467]) and mixed measurement outcomes (k = 18, ES = 0.82, 95% CI [[0.46; 1.170]. Meta-

linguistic judgment test only showed a medium effect size (k= 3, ES = 0.51, 95% CI [-0.01; 1.04].

Lastly, reporting features including sample size and publication years are presented in Table 9.

Table 9

Results of Reporting Features' Subgroup Analysis

Reporting features	Q	d. f.	P-value	k	Standardized mean difference (SMD)	95% CI
Sample size	33.82	2	p < 0.05			
Small				9	0.45	[0.09; 0.81]
Medium				19	0.72	[0.44; 1.00]
Large				15	1.78	[1.44; 2.11]
Publication year	30.53	3	p < 0.05			
2016				13	2.08	[1.51; 2.64]
2017				18	0.81	[0.55; 1.06]

2018	2	0.41	[0.20; 0.62]
2019	10	0.44	[0.13; 0.74]

Comparison groups coded with large sample size based on samples size reported from included studies showed the highest effect size (k = 15, ES = 1.78, 95% CI [1.44; 2.11]) followed by medium group (k = 19, ES = 0.72, 95% CI [0.44; 01.00]. Test of heterogeneity for subgroup differences is significant ($Q = 33.82$, $p < 0.05$). Publication year also showed significant between-group heterogeneity ($Q = 30.53$, $p < 0.05$). Comparison groups from included studies published in 2016 showed the highest effect size (k = 13, ES = 2.08, 95% CI [1.51; 2.64]) followed by 2017 group (k = 18, ES = 0.81, 95% CI [0.55; 1.06]).

In summary, this section reported three parts of the results of meta-analysis: results of primary analysis, a test of heterogeneity and subgroups analysis. 43 effect sizes were calculated and combined. Results of primary analysis indicated a large pooled effect size in favor of using ICT in L2 classrooms. A test of heterogeneity indicated high variance between comparison groups that were not explained by sampling error nor by outlier data points. Subgroups analysis indicated that CMC group, EFL groups and L2 grammar groups presented better effects. Learners with intermediate L2 proficiency or speaking Persian as L1 presented better effects. Comparison groups adopted selected response and medium treatment length presented better effects. Lastly, comparison

groups retrieved from included studies published in 2016 and recruited a large sample size presented better effects.

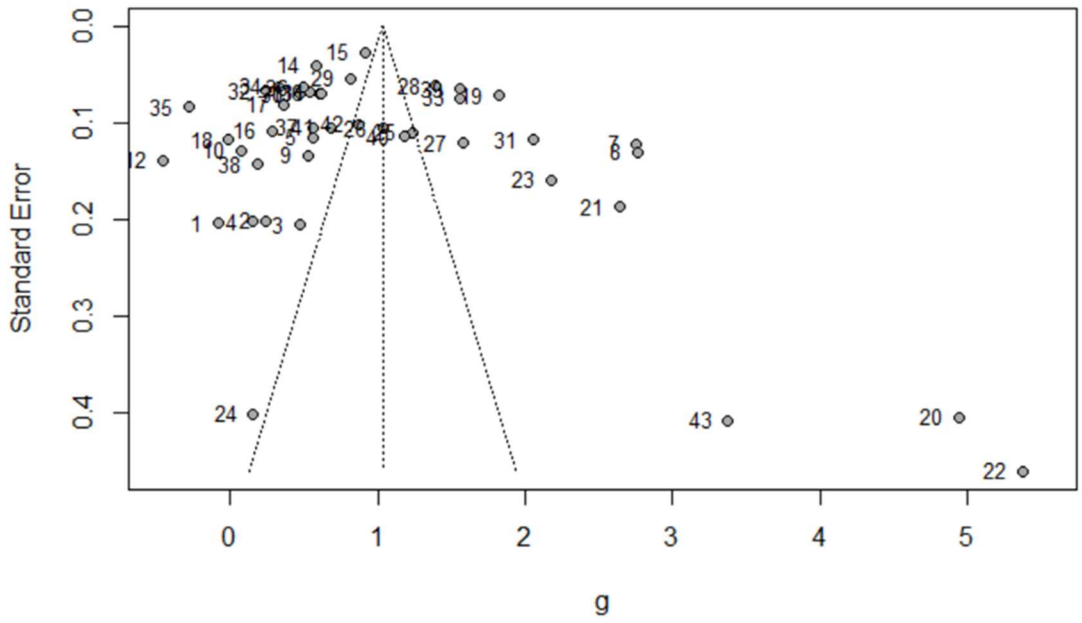
4.3.4. Publication bias

Results of meta-analysis and subgroup analysis of included studies have demonstrated large effects of using ICT in L2 classrooms. However, the large effects might be caused by publication bias, which is stated as the *file drawer* problem, meaning that studies with large and positive effect size were more likely to be published and included in meta-analysis, and thus bias the corpus of results available. This section presents an assessment of potential publication bias of included studies. A funnel plot is presented to visualize potential publication bias. Egger's test and Duval and Tweedie's trim-and-fill procedure are also reported to statistically quantify potential publication bias (Duval & Tweedie, 2000; Egger et al., 1997).

A funnel plot, which shows a visually symmetric pattern of included studies with effect size on the X-axis and the variance on the Y-axis, was created and presented in Figure 3 to assess and visualize the presence of publication bias. The funnel plot should be distributed symmetrically around the pooled effect size when publication bias is absent (Borenstein et al., 2011).

Figure 3

Funnel Plot of All Comparison Groups



As shown in figure 3, comparison groups were plotted on a scatter plot showing the relationship between effect size and standard error shaped as a funnel around the pooled effect size situated in the middle. From the distribution of data points, it could be observed that the funnel plot is relatively symmetric except for the three data points at the bottom left of the plot with large effect sizes. Hence it could be argued that there is little to no publication bias in this meta-analysis. To quantitively assess this argument, an Egger's test of the intercept (Egger et al., 1997) an Duval and Tweedie's trim-and-fill procedure (Duval & Tweedie, 2000) were conducted. When there is no evidence of asymmetry in funnel plot, the intercepts from Egger's test would not significantly differ from zero. Results showed that there is no significant asymmetry in the funnel plot (intercept = 2.4, p = 0.31). In addition, Duval and Tweedie's procedure imputes small

effect size studies in the funnel plot and estimates an adjusted effect size with the simulated data points. Results of this trim-and-fill adjusted model (ES = 0.96, 95% CI [0.77; 1.15]) were relatively similar to the original model. Based on the insignificant asymmetry and results of trim-and-fill procedure, it could be concluded that there is no publication bias observed from included studies.

4.4. Chapter summary

In this chapter, findings of the current meta-analysis were reported as results of study selection, coded study characteristics, results of meta-analysis and publication bias. In total 2070 studies were identified in the electronic database then screened based on inclusion and exclusion criteria. 30 studies met the study eligibility and included for meta-analysis. 43 effect sizes were combined to assess the magnitude of the relationship between ICT and L2 learning outcomes. Results of the meta-analysis indicated a large pooled effect size in favor of using ICT in L2 classrooms. A test of heterogeneity was also significant indicating that between-groups variance was present. Subgroups analysis of independent variables, dependent variables and moderator variables also showed significant within-group heterogeneity for coded subgroups except for educational level variables. Assessments of bias showed that there was no significant publication bias observed across included studies. In the next chapter, interpretations and indications of findings are discussed.

Chapter 5. Discussion

This chapter discusses indications and limitations of the effectiveness of information and communication technology (ICT) integrated second language (L2) classrooms found in the current meta-analysis. Indications are provided based on descriptive features of included studies and statistical inferences derived from 43 comparison groups. Limitations are discussed in three aspects: sampling features, theoretical maturation, and instrument reliability and validity. Finally, future research directions in ICT integrated L2 classrooms are discussed as well.

5.1. Descriptive features

In this section, indications of the current meta-analysis are described based on excluded and included studies. The current study provided a meta-analysis of published studies between 2016 and 2019, examining the effectiveness of ICT integrated L2 classrooms. A vast body of primary studies was searched and screened. The current research characteristics and trends of ICT integrated L2 classrooms are discussed.

Second language acquisition (SLA) research is an interdisciplinary domain which consists of multiple topics, including qualitative narratives, computational designs and psychometrics (Liu et al., 2002). The study screening results reflected the diverse and interdisciplinary features of the research domain. Results showed that empirical evidence supported by rigorous research designs in the domain were still rare.

5.1.1. Features of excluded studies

One of the characteristics observed in current research domain was that the actual percentage of studies that focused on ICT integrated L2 classrooms remains low in SLA research. The current meta-analysis examined 30 primary studies after an exhaustive study search and eligibility screening. Out of 2070 primary studies searched in the electronic database, less than 2 % (n = 30) of searched studies qualified for the meta-analysis. Excluded studies were disqualified mainly due to methodological or substantive inadequacy. From a methodological point of view, excluded studies either showed a lack of experimental study design or an absence of quantitative research methods. Although results from some searched studies might be informative to answer RQs, they were not included in the meta-analysis due to limitation in statistical inferences. Substantive inadequacies included not reporting essential statistics, not reporting L2 learning outcomes, not using ICT for instructions, or reporting only technological design-based results.

5.1.2. Features of included studies

Another characteristic observed in current research domain based on included studies were theoretical and methodological maturation in the research domain. In this section, research trends and characteristics of included studies are summarized in terms of reporting features, substantive features and methodological features. Results of subgroup analysis are discussed as well.

Reporting features. Reporting features (i.e., sample size and publication years) showed significant dispersion across included studies. On one hand, samples size was

variable with several extreme cases observed. The minimum and the maximum number of participants across included studies ranged from 10 to 160 participants respectively and half of the total participants were contributed by large sample size studies (n = 1096). On the other hand it was observed that 75% of the sample size (n = 1554) were collected from studies published in 2016 and 2017. It might be argued that included studies of the current meta-analysis reflected characteristics and trends of studies in the research domain with large sample size published in 2016 and 2017.

Substantive features. A variety of substantive features including participant variables and contextual variables were found across included studies. First, participant variables showed different trends on ICT integrated L2 classrooms. Post-secondary educational level was found consistent with results in previous meta-analysis, being the most frequently researched among others (Grgurović et al., 2013; Zhao, 2003). While post-secondary interventions appeared to be more effective using ICT in previous research (Chiu, 2013; Cho et al., 2018; Grgurović et al., 2013; Lin et al., 2015; Sharifi et al., 2018) , the current meta-analysis found higher effect size in the secondary group (ES = 1.2697) than the post-secondary (ES = 0.98) in subgroup analysis.

Similarly, advanced and intermediate L2 proficiency levels was observed to be the most frequently researched in the current meta-analysis, and it was found to have a higher effect size (ES = 1.20) than previous meta-analysis studies (Grgurović et al., 2013; Sharifi et al., 2018). However, results in the previous meta-analyses showed contradictions. Two similar meta-analyses exploring digital game based learning (DGBL)

and computer assisted instruction (CAI) used for vocabulary and reading reported the greater positive effect size for beginning learners (Abraham, 2008; Tsai & Tsai, 2018).

In terms of L1 backgrounds, participants with less commonly reported L1 backgrounds - Persian, Turkish, Portuguese and Norwegian - were included in the current meta-analysis. Other L1 languages such as Chinese, Arabic and Spanish were more widely observed in the current meta-analysis - consisting of 56% of the sample size (n= 1148). Subgroup analysis showed that Persian, Chinese and Arabic participants obtained medium to large effect size (ES = 3.084; 1.0489; 1.0188). In addition, among observed L1 backgrounds, non-native English speakers were generally using ICT to learn English as a foreign language. It could be argued that the three L1 groups with higher effect sizes would be considered as distant to their target languages. Thus, the greater linguistic distance may result in greater benefit using ICT in L2 classrooms.

Second, contextual variables including L2 contexts, ICT types and targeted L2 skills showed consistent effects with newly emerged trends comparing to previous meta-analyses. L2 contexts reported across included studies were consistent with previous meta-analysis studies (Grgurović et al., 2013; Sharifi et al., 2018; Zhao, 2003). EFL was the most frequently researched L2 context, taking up 86% (n = 1780) of the total sample size with the highest effect size. A growing number of less commonly taught languages - such as Chinese, German and Dutch - were also reported from 9 of comparison groups but only observed with small to medium effect size.

Innovative ICT and newly emerged L2 skills in the L2 classroom were observed across included studies. 5 types of ICT tools were reported, in which MALL was found to

be the most frequently researched ICT type. While Sharifi and her team (Sharifi et al, 2018) included only web-based CAI in their examination of innovative ICT, results of the current meta-analysis found various types of innovative ICT with substantive effect sizes such as MALL (ES = 0.9417) and IiCALLCALL (ES = 0.7430). This result differs from Cho et al. (2018)'s analysis where the effect size for MALL was much lower (ES = 0.51). Effects of different ICT types also indicated significant heterogeneity. Thus, it could be argued that recently published studies applied more innovative ICT types and found greater learning effects than traditional CALL in similar meta-analyses (Grgurović et al., 2013; Sharifi et al., 2018; Zhao, 2003).

Similar to patterns of ICT type, 8 types of targeted L2 skills were observed across included studies with significant between-groups heterogeneity. L2 grammar was found to be the most frequently researched L2 skill and had the largest effect size (ES = 1.69) among others. L2 pragmatics was found in two comparison groups and appears to have had a negative effect (ES = -0.27). Instead of reporting L2 grammar and L2 pragmatics, previous meta-analyses reported large effects in L2 vocabulary (ES = 0.73) and L2 mixed skills (S = 0.81) (Cho et al., 2018; Sung et al., 2015). Negative effects were found for L2 speaking skills (measuring accuracy and fluency) when using CMC (Lin, 2015), while the current meta-analysis obtained a medium positive effect size of using ICT for L2 speaking skills using innovative ICT. Thus, it could be argued that the current meta-analysis contributed new information in the research domain, suggesting that innovative ICT integrated L2 classrooms may have a greater positive effect on L2 grammar and possibly a negative effect on would be L2 pragmatic learning.

Methodological features. Methodological features including treatment duration and outcome measurements showed different results than previous studies in the current meta-analysis. First, medium treatment length was observed to be the most researched across included studies and medium to long treatment duration resulted in large effect size (ES = 1.30; 1.17). Treatment duration reported from similar meta-analyses often used different matrices and ranges (e.g., Grgurović et al., 2013; Sharifi et al., 2018). The current meta-analysis coded treatment length between 1.25 and 13.5 weeks as medium lengths based on the 25th and the 75th quartiles from the distribution of treatment length retrieved from included studies. Results of the analysis are consistent with similar meta-analyses, indicating that longer treatments would produce better learning effects.

Similarly, outcome measurements reported from the previous meta-analyses also lack a common practice in the description of outcome measures. Measurement types reported across included studies also showed different results. Norris and Ortega (2006) pointed out a reasonable practice of outcome measurements including five types of measurements: meta-linguistic judgement, selected response, constrained-constructed response and free-constructed response. Results indicated that mixed measurements were the most frequently researched while selected response (ES = 2.59) and free-constructed response (ES = 0.93) obtained the largest effect sizes.

In summary, research on ICT integrated L2 classrooms reflected features of diversity and interdisciplinary in the research domain. A growing number of studies using ICT as design-based tools or reporting psychometrics in the affective domain were found. However, quantitative studies with rigorous experimental design are still rare and the

effectiveness of ICT integrated L2 classrooms has yet to be fully investigated. Out of 2070 studies identified, 30 studies were included in the current meta-analysis and provided updated features and trends in the domain, including larger sample size and clusters of publication between 2016 and 2017. In terms of participant features, post-secondary level, intermediate L2 proficiency level, and far linguistic distance were more researched variables in the domain and were found to have larger effect sizes. Contextual variables including ICT type, targeted L2 skills and L2 contexts showed new research trends in the domain. It appears that innovative ICT type, such as mobile assisted language learning (MALL) and intelligent computer assisted language learning (ICALL), may be more beneficial for ICT integrated L2 classrooms. Newly researched skills, including L2 grammar and pragmatics, were included in the selected studies. L2 grammar appears to have greater benefits in ICT integrated L2 classrooms while pragmatics showed negative effects. EFL was still the most frequently researched variable and yielded the highest effect size. Lastly, strong experimental manipulation and methodological improvement were observed across included studies with more studies including longer treatments and complex measurements.

5.2. Statistical inferences

In this section, research questions regarding the direction, magnitude and potential impact of moderating factors of the effectiveness of ICT integrated L2 classrooms are answered. A total of 43 effect sizes were combined in the unit of comparison groups based on 106 standardized mean score differences retrieved and calculated as effect sizes from 30 included studies. A pooled effect size was calculated to evaluate the

effectiveness of ICT integrated L2 classrooms. In this section, indications of individual effect sizes and the pooled effect size are discussed in two parts: the direction of effects and the magnitude of effects. Between-groups heterogeneity across included studies are also discussed.

5.2.1. The direction of effect sizes

The current meta-analysis examined the direction of the effects in ICT integrated L2 classrooms as either positive or negative. The direction of effects is discussed both in the unit of mean effect sizes and in the unit of the pooled effect size. In the unit of mean effect sizes from individual study level, both negative and positive effect sizes were found in comparison groups. Out of 43 mean effect sizes, there were 4 negative effect sizes and 39 positive effect sizes. The largest effect size (ES = 5.37, 95% CI [3.97; 6.76]) was contributed by Alibakhshi and Mohammadi's (2016) study exploring the use of graphic assistance to support a synchronous computer mediated communication (CMC) tool for Iranian EFL learners in learning L2 grammar. These researchers argued that synchronized CMC tools raised the awareness and opportunities for learners to negotiate for meaningful conversation. In addition, graphics in synchronous CMC might trigger additional cueing and attention in L2 classrooms which matches the cognitive skill of receptive tests (multiple choice recognition) reported in the study. Thus the majority of included studies indicated a level of the convergence of internal validity between ICT design and the outcome measurement, which might contribute to the large pooled effect size in the current meta-analysis.

The 4 negative effects were contributed by individual studies exploring the use of ICALL, CMC, MALL tools for L2 grammar, writing, speaking and pragmatics. Ziegler et al. (2017) investigated using ICALL to assist L2 grammar learning and found that it would require learners to interact with input materials to obtain positive effects for computerized input enhancement tools. Mørch et al. (2017) further refer to such negative effects on the transferability of learning skills in ICT integrated L2 classrooms. They found negative effects using a web-based semantic analysis system to provide L2 English writing feedback for Norwegian secondary students. Researchers argued that the negative effect of the ICALL tool might be due to younger participants' lack of certainty about how to progress, as well as inability to transfer ideas to their essay writing.

The negative transferability between ICT integrated L2 classrooms and learning outcomes were also investigated for learning Chinese as L2. Lan and Lin (2016) explored the use of a mobile application assisting Chinese L2 speaking. In the mixed design study, mobile-enhanced groups were observed with more unexpected conversations, better fitted social conversations and a higher likelihood to use L2 in conversations, while communication performance tests showed the classroom group outperformed the mobile-enhanced group. Thus, researchers called for further investigation of the miss-match of the instruments and instructions in the Chinese learning context. Tang's study (2019) further explored teaching Chinese grammar through CMC environment. Results of the study supported Sykes' hypothesis (2005) that CMC would be beneficial for non-linguistic/affective outcomes while FTF would be superior for linguistic outcomes in the context of learning pragmatics of Chinese modal verbs due to learner's neglect of grammatical structures in casual conversational settings like CMC.

Negative effects in the current meta-analysis possibly reflect the pattern of theoretical development and maturation in the research domain. Early research conducted in the domain often featured direct and strong manipulations, and thus, might be expected to yield large effect sizes (Ioannidis, 2008). Subsequently, later published studies in a matured research domain often shifted to testing the generalizability of the intervention in various research contexts (Plonsky & Oswald, 2014). As observed, less commonly taught language and targeted L2 skills occurred in ICT integrated L2 classrooms research and demonstrated negative effects. Theoretical maturation in ICT integrated L2 classrooms might be one of the reasons for the observations of smaller effects over time. For example, L2 pragmatics were newly emerged in L2 research and results of subgroup analysis indicated negative effects of using ICT (ES = -0.2680). Less commonly investigated L1 background, such as Norwegian and Chinese, and appeared to contribute negative effects. Hence, it could be argued that this is an example of the Proteus effect (Krist, 2018) where negative effects were discovered as a sign of maturation and generalization of ICT in L2 classrooms.

On the other hand, studies with negative effects provided insights on the theoretical development of ICT integrated L2 classrooms. Early published research of ICT integrated L2 classrooms were often completed in English settings with simple experimental manipulations to determine whether an effect exists (Plonsky & Oswald, 2014; Zhao, 2003). However, studies produced negative effects in the current meta-analysis reported using innovative ICT types, as well as less commonly reported L2 contexts. Researchers also argued for the theoretical development of ICT integrated L2

classrooms, suggesting that neglects of the transferability between instructions and instruments might cast negative learning outcomes.

In summary, both positive effects and negative effects were found across included studies as theoretical development advanced researchers investigated more nuanced and less significant aspects of ICT integrated L2 classrooms. Several large effect sizes were also found in ICT integrated L2 classrooms which suggested strong manipulation and internal validity. By summarizing features of studies with negative effects, it could be concluded that ICT tools would be most effectively used in L2 classrooms when the features of applied ICT medium were compatible with designed outcome measures.

5.2.2. The magnitude of effect sizes

Additionally, the magnitudes of ICT integrated L2 classrooms were also of interest. In this section, the literal and mathematical interpretation of the pooled effect size is examined and discussed in comparison to precious meta-analysis studies. Theoretical and methodological indications of the meta-analysis results for SLA research are discussed.

Interpretation of effect size in the domain of SLA research has a different standard than other domains. Plonsky and Oswald (2014) argued that Cohen's cut points for effect size overvalued pooled effect sizes across SLA research. Thus, an adjusted rule of thumb for of effect size in SLA research was suggested for the standardized mean difference between groups. The adjusted rule of thumb for effect size equals 0.4 would be small, 0.7 as medium and 1.0 as large. Results of the current meta-analysis reported a

large positive pooled effect size (ES = 1.03) across included studies with 95% confidence interval of [0.84; 1.22]. With this difference, the mean of the treatment group is equated with approximately the 84th percentile of scores in the control group. It is also the highest pooled effect size reported among similar meta-analysis studies (e.g., Grgurović et al., 2013; Sharifi et al., 2018; Zhao, 2003).

Four reasons might contribute to the observed large effect size. First, the impact of differences in measures are overlooked and might cause unexpected results. Constructs in measurements might vary based on many factors and decisions made by the researcher, then consequently operationally and psychometrically bias effects on the reported effect sizes (Plonsky & Oswald, 2014). One of the factors that would influence the magnitude of effects is the correlation of constructs. The current meta-analysis observed 106 multivariate results reported from 30 included studies. However, not all constructs were nested in one comparison condition but often several constructs would be combined to represent the results of a comparison. Due to the lack of reporting of measurement reliability in SLA research, the current meta-analysis followed Teimouri et al.'s (2019) practice and assumed that the correlation coefficients between multivariate results were r = 0. Then the mean of multivariate results was assigned to each comparison group as the unit of analysis. However, this practice is not free from theoretical disadvantages. Theoretically, the assumption of zero correlation between items would lead to an underestimation of the variance of effect size locally (Borenstein et al., 2011), thus tend to increase the magnitude of the pooled effect size.

In addition, different types of measurement instruments used in studies might also have an impact on pooled effect size. In SLA research, larger effects were often observed with free-constructed response (Plonsky & Oswald, 2014). Results of meta-analysis support this statement as free constructed-response test type is observed to be the second most frequently researched measurement type in the current meta-analysis (k = 13). The effect size of free-constructed response group is also found to be the second largest within the measurement outcomes subgroup with significant between-groups heterogeneity. Thus, it could be argued that the popularity of adopting free-constructed response from included studies might increase the observed pooled effect size.

Second, the current meta-analysis only included primary studies using experimental design. In the general field of experimental studies, more strict and rigorous experimental manipulations, such as using strict experimental control, would be expected to produce a larger mean difference between groups (Plonsky & Oswald, 2014). Experimental design requires strict control of potential differences between treatment and control groups, thus, limiting the potential impact of unknown variables and effects other than the intervention. In practice, strict experimental designs are often characterized by strong manipulation, in which case the control-group might not receive equal amounts of instructional stimuli as the treatment groups, thus, leading to larger standardized mean difference.

In comparison to previous meta-analysis studies, Grgurović et al. (2013) argued that early published SLA research lacked methodological rigour in many studies in terms of random assignment which was also supported by Zhao (2003), thus resulting in

smaller pooled effect size. Sharifi et al. (2018) claimed to include experimental studies in their meta-analyses and found a larger pooled effect size than previous. However, the team did not take the random assignment into consideration, which reduced the rigor (quasi-experimental). Thus, by strictly controlling experimental conditions of included studies, the current meta-analysis observed higher pooled effect size of ICT integrated L2 classrooms. Given the positive relationship between strict experimental condition and overall effects, it could be argued that strict experimental control in ICT integrated L2 classrooms would lead to observations of larger effect size. This finding is also consistent with the previous meta-analysis and theoretical studies.

Third, methodological adjustments and improvements in a research domain would cause larger effect sizes over time (Fern & Monroe, 1996). In SLA research particularly, longer treatment durations might yield larger mean differences between-groups (Plonsky & Oswald, 2014). The findings of this study reflect the fact that included studies generally employed strong experimental manipulation in terms of larger sample sizes and longer durations for treatment groups. In addition, subgroup analysis indicated that strong experimental manipulation also contributed large effect size with significant between-groups heterogeneity. Thus, it could be observed that there are improvements of methodologies that occurred in the domain that have led to larger pooled effect size.

Although the pooled effect size would appear to be large, results of subgroup analysis indicated that included studies published between 2016 and 2017 produced larger effect sizes while later published studies may have smaller effect sizes with significant heterogeneity found between subgroups ($Q = 30.53$, $p < 0.05$). Considering

with findings of emerging innovative ICT and L2 skills across included studies, it could be argued that the surge of positive effects in studies published between 2016 and 2017 might also represent the emergence of a newer research strand. In general, there should be an inverse relationship between year of publication and the effect size due in part to the maturation of methodological practice as researchers gain more insights in a newly emerged field (Cobb, 2010). Thus, it could be concluded that the increase of pooled effect size in the current meta-analysis would be due to methodological maturation in the field. However, the maturation should be viewed with respect to long-term research agenda and fluctuations might be expected when the number of studies under consideration is low.

Fourth, publication bias in SLA research might contribute to the inflation of the pooled effect size. The phenomenon is called researchers' drawer effect (Norris & Ortega, 2006), where statistically significant results tend to be published more frequently, thus, biasing the meta-analysis results. Publication bias would be present in a meta-analysis study if included studies were an unrepresentative collection of studies. However, both statistical and visual analyses indicated that there is no significant publication bias in the current meta-analysis. First, the funnel plot indicated a visually symmetric pattern of included studies with effect size on the X-axis and the variance on the Y-axis. A trim-and-fill technique showed that the adjusted pooled effect size with smaller effect studies filled into the collection would still be considered as medium to large. Lastly, Egger's test indicated that there is no statistically significant asymmetry in the funnel plot. Hence, it could be argued that there was no significant publication bias in the current meta-analysis inflating the overall effect.

5.2.3. Heterogeneity

Research in a mature research domain would demonstrate more varied research questions and research contexts, and thus, might produce more heterogeneity in overall effects (Plonsky & Oswald, 2014). The current meta-analysis found consistent results in heterogeneity tests with significant heterogeneity ($Q = 2300.93$, $p < 0.05$; T-square = 0.3815; I-square = 98.2%). The results could be interpreted that there is not only statistically significant true variance between comparison groups retrieved from included studies, but the true variance takes 98.2% (I-square = 98.2%) of the entire variance of effect sizes. Considering the guideline provided by Higgins' team (Higgins et al., 2003), in which they suggested that values on the order of 25%, 50%, and 75% were considered as low, intermediate and high dispersion respectively, it could be argued that there is high between-groups heterogeneity in the current meta-analysis.

Many factors might contribute to high heterogeneity. One of them might be the wide scope of the current meta-analysis. A field-wide scope meta-analysis may be "blind" to the substantive differences across multiple minor strands but sensitive to the statistical variance across included studies (Plonsky & Oswald, 2014). That is to say that a field-wide meta-analysis might aggregate effects of different research topics to a single pooled effect size, and thus, results in high heterogeneity across included studies. Results of subgroup analysis are consistent with this hypothesis, where nine of ten subgroups in the current meta-analysis showed significant between-groups heterogeneity and only educational level subgroup showing insignificant heterogeneity.

Blake (2013) suggested that the interdisciplinary nature of ICT integrated L2 classrooms should results in significant heterogeneity across studies. Given that more empirical evidence on different topics in ICT integrated L2 classrooms was observed in this study, it could be argued that significant results of heterogeneity test in the current meta-analysis were consistent with similar meta-analyses studies in the domain (Grgurović et al., 2013; Sharifi et al., 2018 ; Zhao, 2003). It would also resonate with the theoretical maturation of ICT integrated L2 classrooms by reporting various minor topics across included studies.

In summary, results of the analysis provided abundant statistical inferences regarding the direction, magnitude and variance of the effectiveness of ICT integrated L2 classrooms. The study sample included both positive and negative individual effect sizes. Analysis of individual studies indicated that learning effects in ICT integrated L2 classrooms might be positively moderated by the internal validity and transferability between instructional design and outcome measures. In addition, effects retrieved from individual studies reflected methodological development and maturation of research in ICT integrated L2 classrooms. Methodological maturation in the domain, for example, stronger and more rigorous experimental manipulation, would also explain the strong magnitude of the overall instructional effect of ICT integrated L2 classrooms. The magnitude of the pooled effect size would be overall large, despite the existence of some studies with negative effects. Results of the current meta-analysis support previous arguments with a cohort of updated studies: ICT integrated L2 classrooms has positive effects on improving L2 learning outcomes. Lastly, analysis of heterogeneity showed significant variance both between comparison groups and subgroups, indicating potential

moderating effects of subgroups on the effectiveness of ICT integrated L2 classrooms. High heterogeneity between subgroups also resonated with theoretical maturation observed in the study features section.

5.3. Limitations and future directions

The current meta-analysis provided updated evidence supporting the effectiveness of ICT integrated L2 classrooms, with a large positive pooled effect size observed across included studies. Results of meta-analysis reflected theoretical and methodological development and maturation in the domain with potential variance and cautions. In this section, cautions and limitations when interpreting results of the current meta-analysis are discussed with suggestions for future research.

5.3.1. Sampling limitations

The current meta-analysis conducted a systematic literature search in electronic databases with an inclusive approach to capture a cohort of primary studies that is representative of the subject matter. Results of searching included 29 unpublished dissertations to which there was limited access. Although statistical and visual analysis suggested that there would be no significant publication bias present, it would be of stakeholders' interests to provide a fuller picture of substantive results in ICT integrated L2 classrooms. Thus, the current researcher suggests future researchers adopt a more inclusive approach when searching studies, including (a) documenting searching details and decisions in every step; (b) incorporating a journal-specific searching of the domain;

and (c) being creative on searching studies published in relevant domains, for more comprehensive syntheses in the future.

5.3.2. A narrower-scope meta-analysis

The practice of using ICT in L2 classrooms have accumulated over the past decades (Otto, 2017). Abundant evidence of primary studies in the domain indicated theoretical and methodological maturation with high variabilities across studies including research designs, innovative ICT types and emerging L2 skills. Given the goal of the current meta-analysis was to provide an up-to-date evaluation of ICT integrated L2 classrooms, effects of moderator variables from individual studies might be blurred. Results of heterogeneity test and subgroups analysis also indicated significant variations between 9 subgroups that have yet been fully investigated from a field-wide scope meta-analysis. Thus, the current researcher suggests future studies might focus on a narrow scope in (a) innovative ICT types; (b) subgroups with small or negative effects; and (c) internal validity and transferability of ICT-enhanced instructions, to further illuminate the effectiveness of ICT integrated L2 classrooms.

5.3.3. Validity and reliability of instruments

Lastly, different measures of L2 learning outcomes were observed across included studies. However, the impact of the validity of theoretical constructs within and between studies was overlooked thus might bias the overall precision of effects in the current meta-analysis. In general, SLA research would arbitrarily adopt either a part of standardized language test or previously designed tests for other purposes. Language test scores reported measuring the same group were often validation for constructed learning

outcomes. Overlooking measurement factors that traditional meta-analysis methods are not sensitive to, such as Chronbach's alpha and the correlation between test items, would negatively affect the precision of a meta-analysis study. The current researcher suggests future researchers might (a) report evidence of test validity; (b) consider a psychometrically adjusted meta-analysis approach; (c) consider range restrictions of observed effects when interpreting overall effects.

5.4. Conclusion

The current meta-analysis was conducted to examine the effectiveness of ICT integrated L2 classrooms. In response to the first research question, there were 10 subgroups coded that summarized substantive, methodological and reporting features of primary studies published between 2016 and 2019. Coded study features suggested that studies exploring the use of innovative ICT types and emerging L2 skills are trending in the research domain. More studies adopted stricter experimental manipulation, such as larger sample size and longer treatment length. Features and trends summarized from included studies further indicated that both theoretical and methodological development were present when compared with previous meta-analysis studies (Zhao, 2003; Grgurović et al., 2013; Sharifi et al., 2018). The second research question contributes to systematic evidence on the direction and magnitude of the effectiveness of ICT integrated L2 classrooms. Considering Cohen's and Plonsky's guidelines (Cohen, 2013; Plonsky & Oswald, 2014) for interpreting effect size in SLA research, the current meta-analysis reported a large effect size in favor of using ICT for L2 learning despite some negative results found in individual studies. Comparing with similar meta-analysis, results of the

current meta-analysis were consistent with the positive effects of using ICT with stronger magnitude and higher variability in research contexts.

The current meta-analysis also explored the influence of potential moderator variables on the effectiveness of ICT integrated L2 classrooms. Considering theoretical and empirical constructions in the domain (see e.g., Grgurović et al., 2013; Norris & Ortega, 2000; Plonsky & Ziegler, 2016; Sharifi et al., 2018), the current researcher coded examined 10 moderator variables across included studies including ICT type, targeted L2 skills, L1 backgrounds, L2 contexts, educational level, L2 proficiency level, outcome measures, treatment duration, sample size, and publication year. Innovative ICT types, such as CMC, ICALL and MALL appear to have greater effects in ICT integrated L2 classrooms. L2 grammar showed to be the most researched and found higher effects when taught in ICT integrated L2 classrooms. Other variables, except educational level, suggested moderating effects of ICT integrated L2 classrooms with significant between-groups heterogeneity.

Finally, the current meta-analysis addressed limitations and future directions in the research domain. Quantitative studies with strict experimental designs are critical in providing comparable empirical evidence of instructional effects. However, such studies are still rare in ICT integrated L2 classrooms. With theoretical and methodological maturation observed in the research domain, researchers might expect more variability to occur in the treatments offered in the ICT integrated L2 classrooms. Future meta-analysis might also adopt a narrow-scope research topic exploring effects of newly emerged ICT types, treatment approaches or L2 skills. Future researchers might also pay attention to

the impact of instruments reliability when analyzing across studies to improve the precision of meta-analysis results.

In summary, the current meta-analysis illuminated the effectiveness of using ICT in L2 classrooms in the period 2016-2019. Theoretical and methodological maturation was observed in the domain, and large effect size was obtained in favour of ICT integrated L2 classrooms with significant between-groups heterogeneity. Subgroup analysis suggested ICT types, targeted L2 skills, L1 backgrounds, L2 context, L2 proficiency level, outcome measures, treatment duration, sample size, and publication year would moderate learning effects of ICT integrated L2 learning. Future meta-analysis might further explore the effects of ICT integrated L2 classrooms with a narrower scope, as well as the impact of testing instruments on the precision of effect sizes observed.

References

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, *21*(3), 199–226. https://doi.org/10.1080/09588220802090246

Anderson, S. J. (1976). *TICCIT Project.* https://eric.ed.gov/?id=ED134226

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bax, S. (2003). CALL—past, present and future. *System, 31*(1), 13–28.

Blake, R. (2016). Technology and the four skills. *Language Learning & Technology*, *20*(2), 129.

Blake, R. J. (2013). *Brave new digital classroom: Technology and foreign language learning*. Georgetown University Press.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348–393. https://doi.org/10.1111/lang.12224

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL*, *27*(01), 4–20. https://doi.org/10.1017/S0958344014000159

Borenstein, M., Cooper, H., Hedges, L., & Valentine, J. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis*, *2*, 221–235.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Cerezo, L., Baralt, M., Suh, B.-R., & Leow, R. P. (2014). Does the medium really matter in L2 development? The validity of CALL research designs. *Computer Assisted Language Learning, 27*(4), 294–310. https://doi.org/10.1080/09588221.2013.839569

Chang, M.-M., & Lin, M.-C. (2013). Strategy-oriented web-based English instruction—A meta-analysis. *Australasian Journal of Educational Technology*, *29*(2). https://doi-org.ezproxy.library.uvic.ca/10.14742/ajet.67

Chapelle, C. A. (2006). Interactionist SLA theory in CALL research. In J. L. Egbert & G. M. Petrie (Eds.), *CALL research perspectives* (pp. 65–76). Routledge.

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: The effectiveness of DGBL. *British Journal of Educational Technology*, *49*(1), 69–77. https://doi.org/10.1111/bjet.12526

Chiu, Y., Kao, C., & Reynolds, B. L. (2012). The relative effectiveness of digital game-based learning types in English as a foreign language setting: A meta-analysis: Colloquium. *British Journal of Educational Technology*, *43*(4), E104–E107. https://doi.org/10.1111/j.1467-8535.2012.01295.x

Chiu, Y.-H. (2013). Computer-assisted second language vocabulary instruction: A meta-analysis. *British Journal of Educational Technology*, *44*(2), E52–E56. https://doi.org/10.1111/j.1467-8535.2012.01342.x

Cho, K., Lee, S., Joo, M.-H., & Becker, B. J. (2018). The effects of using mobile devices on student achievement in language learning: A meta-Analysis. *Education Sciences*, *8*(3), 105. https://doi.org/10.3390/educsci8030105

Chun, D. M. (2016). The role of technology in SLA research. *Language Learning & Technology*, *20*(2), 98–115.

Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences. Routledge. https://doi.org/10.4324/9780203771587

Cook, T. D, & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D.Dunnette (Eds.), *Handbook of industrial and organizational psychology* (pp. 223-326). Chicago: Rand McNally.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

Cope, B., & Kalantzis, M. (2017). *e-Learning Ecologies: Principles for New Learning and Assessment*. Taylor & Francis.

Dickersin, K. (2006). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In W.Viechtbauer (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (pp. 11–34). John Wiley & Sons.

Duval, S., & Tweedie, R. (2000). 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

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

Fern, E. F., & Monroe, K. B. (1996). Effect-Size Estimates: Issues and Problems in Interpretation. *Journal of Consumer Research, 23*(2), 89. https://doi.org/10.1086/209469

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Sage Publications.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*(2), 165–198. https://doi.org/10.1017/S0958344013000013

Harrer, M., Cuijpers, P., Furukawa, T. & Ebert, D. D. (2019). dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R' (R package version 0.0.9000). http://dmetar.protectlab.org.

Hart, R. S. (1995). The Illinois PLATO foreign languages project. *CALICO Journal*, *12*(4), 15–37.

Haythornthwaite, C., & Andrews, R. (2011). *E-Learning Theory and Practice*. Sage Publications. https://doi.org/10.4135/9781446288566

Hedges, L. V. (1994). Fixed effects models. In R. J. Light (Eds.), *The Handbook of Research Synthesis*, 285, 299.

Hedges, L. V., & Olkin, I. (2014). *Statistical Methods for Meta-analysis.* Academic press.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hudson, T., & Llosa, L. (2015). Design issues and inference in experimental L2 research *Language Learning, 65*(S1), 76–96. https://doi.org/10.1111/lang.12113

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology, 19*(5), 640–648. JSTOR.

Iyengar, S., & Greenhouse, J. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis*, 417–433.

Jia, J. (2019). Theories and Literatures for Technology-Enhanced Language Instruction. In C Fuchs (Eds.), *Computer-Assisted Language Learning: Concepts, Methodologies, Tools, and Applications* (pp. 50–82). IGI Global.

Kaufman, D. (2004). 14. Constructivist issues in language learning and teaching. *Annual Review of Applied Linguistics*, *24*, 303–319.

Krist, A. H. (2018). "Needs More Research"—Implications of the Proteus Effect for Researchers and Evidence Adopters. *Mayo Clinic Proceedings, 93*(3), 273–275. https://doi.org/10.1016/j.mayocp.2018.01.013

Kukulska-Hulme, A., Lee, H., & Norris, L. (2017). Mobile Learning Revolution. In C. A. Chapelle & S. Sauro (Eds.), *The Handbook of Technology and Second Language Teaching and Learning* (pp. 217–233). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118914069.ch15

Lan, Y.-J., & Lin, Y.-T. (2016). Mobile Seamless Technology Enhanced CSL Oral Communication. *Educational Technology & Society, 19*(3), 335.

Lantolf, J. P., Thorne, S. L., & Poehner, M. E. (2015). Sociocultural theory and second language development. In VanPatten, B., Keating, G. D., & Wulff, S. (Eds.), *Theories in Second Language Acquisition: An Introduction*, 207–226.

Leow, R. P. (2019). *The Routledge Handbook of Second Language Research in Classroom Learning*. Taylor & Francis Group.

Levy, M., Hubbard, P., Stockwell, G., & Colpaert, J. (2015). Research challenges in CALL. *Computer Assisted Language Learning*, *28*(1), 1–6. https://doi.org/10.1080/09588221.2014.987035

Liberati, A., Altman, D. G., Tetzlaff, J., & Mulrow, C. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine, 6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

Light, R. J., & Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research.* Harvard University.

Lin, H. (2015a). A meta-synthesis of empirical research on the effectiveness of computer-mediated communication (CMC) in SLA. *Language Learning & Technology*, *19*(2), 85–117.

Lin, H. (2015b). Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis. *ReCALL*, *27*(3), 261–287. https://doi.org/10.1017/S095834401400041X

Lin, J.-J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *32*(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

Lin, W. C., Huang, H. T., & Liou, H. C. (2013). The effects of text-based SCMC on SLA: A meta analysis. *Language Learning & Technology, 17*(2), 123-142.

Liu, M., Moore, Z., Graham, L., & Lee, S. (2002). A look at the research on computer-based technology use in second language learning: A review of the literature from 1990–2000. *Journal of Research on Technology in Education*, *34*(3), 250–273.

Lin, H., Chen, T., & Liou, H.-C. (2018). Transparency of reporting in CALL meta-analyses between 2003 and 2015. *ReCALL, 30*(3), 253–277. https://doi.org/10.1017/S0958344017000271

Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In RJ Light (Eds.), *The Handbook of Research Synthesis*, (pp. 111–123).

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209. https://doi.org/10.1037/0003-066X.48.12.1181

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-analysis* (pp. ix, 247). Sage Publications, Inc.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic Reviews and Meta-analysis.* Oxford University Press.

Loewen, S. (2020). *Introduction to Instructed Second Language Acquisition*. Routledge.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In V Ritchie, W. Bhatia (Eds.), *The New Handbook of Second Language Acquisition*, (pp. 413–468). Brill.

Long, M. H. (2007). Recasts in SLA: The story so far. *Problems in SLA*, 75–116. Lawrence Erlbaum Associates.

Lüdecke. (2019). esc: Effect Size Computation for Meta Analysis (R package version 0.5.1). CRAN. https://github.com/strengejacke/esc/issues

Mahdi, H. S. (2018). Effectiveness of mobile devices on vocabulary learning: A meta-analysis. *Journal of Educational Computing Research*, *56*(1), 134–154. https://doi.org/10.1177/0735633117698826

Management Association, I. R. (Eds.). (2019). *Computer-Assisted Language Learning: Concepts, Methodologies, Tools, and Applications*. IGI Global. https://doi.org/10.4018/978-1-5225-7663-1

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The Effectiveness of Online and Blended Learning: A Meta-Analysis of the Empirical Literature. *Teachers College Record, 115*(3).

Mørch, A. I., Engeness, I., Cheng, V. C., Cheung, W. K., & Wong, K. C. (2017). EssayCritic: Writing to Learn with a Knowledge-Based Design Critiquing System. *Journal of Educational Technology & Society; Palmerston North, 20*(2), 213–223.

Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720–739. https://doi.org/10.1016/j.system.2013.07.013

National Standards in Foreign Language Education Project, N. (1996). *Standards for Foreign Language Learning: Preparing for the 21st Century*. ERIC Clearinghouse.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-analysis. *Language Learning, 50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing Research on Language Learning and Teaching*, (pp. 3-50). John Benjamins Publishing.

Otto, S. E. K. (2017). From Past to Present. In *The Handbook of Technology and Second Language Teaching and Learning* (pp. 10–25). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118914069.ch2

Picciano, A. G. (2019). Blending with purpose: The multimodal model. *Online Learning*, *13*(1). https://doi.org/10.24059/olj.v13i1.1673

Pigott, T. D. (2012). Power for the Test of Homogeneity in Fixed and Random Effects Models. In T. D. Pigott (Eds.), *Advances in Meta-Analysis* (pp. 55–66). Springer US. https://doi.org/10.1007/978-1-4614-2278-5_5

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

Plonsky, L. & Oswald, F. L. (2015). Meta-analyzing second language Research. In L. Plonsky (Eds.) *Advancing Quantitative Methods in Second Language Research*, Routledge. https://doi.org/10.4324/9781315870908-6

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA Interface: Insights from a Second-order Synthesis. *Language Learning & Technology*, *20*(2), 17.

Rosenthal, R. (1991). Meta-analysis: A review. *Psychosomatic Medicine, 53*(3), 247–271

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and Correlations in Effect-Size Estimation. *Psychological Science, 11*(6), 446–453. https://doi.org/10.1111/1467-9280.00287

Salaberry, M. R. (2001). The use of technology for second language learning and teaching: A retrospective. *The Modern Language Journal*, *85*(1), 39–56.

Sauro, S., & Chapelle, C. A. (2017). Introduction to the *The handbook of technology and second language teaching and learning*. In Chapelle C. & Sauro S. (Eds), The Handbook of Technology and Second Language Teaching and Learning (pp.1-9). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118914069.ch2

Schmidgall, J. E., & Powers, D. E. (2017). Technology and High-stakes Language Testing. In Chapelle C. & Sauro S. (Eds), *The Handbook of Technology and Second Language Teaching and Learning* (pp.317-331). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118914069.ch2

Schwarzer. (2020). meta: General Package for Meta-Analysis (R package version 4.14-0). CRAN. https://github.com/guido-s/meta

Shadish, W. R., & Haddock, C. R. (1994). *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Sharifi, M., Rostami AbuSaeedi, A., Jafarigohar, M., & Zandi, B. (2018). Retrospect and prospect of computer assisted English language learning: A meta-analysis of the empirical literature. *Computer Assisted Language Learning*, *31*(4), 413–436. https://doi.org/10.1080/09588221.2017.1412325

Smith, B. (2017). Technology-Enhanced SLA Research. In Chapelle C. & Sauro S. (Eds), *The Handbook of Technology and Second Language Teaching and Learning* (pp. 444–458). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118914069.ch29

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

Suppes, P. (1968). Computer-assisted instruction: An overview of operations and problems. *IFIP Congress* (2), 1968, 1103–1113.

Sykes, J. M. (2005). Synchronous CMC and Pragmatic Development: Effects of Oral and Written Chat. *CALICO Journal, 22*(3), 399–431. JSTOR.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science, 12*(2), 257-285. https://doi.org/10.1207/S15516709COG1202_4

Taj, I. H., Sulan, N. B., Sipra, M. A., & Ahmad, W. (2016). Impact of mobile assisted language learning (MALL) on EFL: A meta-Analysis. *Advances in Language and Literary Studies*, *7*(2), 76–83.

Tang, X. (2019). The effects of task modality on L2 Chinese learners' pragmatic development: Computer-mediated written chat vs. face-to-face oral chat. *System, 80*, 48–59. https://doi.org/10.1016/j.system.2018.10.011

Taylor, A. (2013a). The effects of CALL versus traditional L1 glosses on L2 reading comprehension. *CALICO Journal*, *23*(2), 309–318. https://doi.org/10.1558/cj.v23i2.309-318

Taylor, A. M. (2009). CALL-based versus paper-based glosses: Is there a difference in reading comprehension? *CALICO Journal*, *27*(1), 147–160. https://doi.org/10.11139/cj.27.1.147-160

Taylor, A. M. (2013b). CALL versus Paper: In Which Context Are L1 Glosses More Effective? *CALICO Journal*, *30*(1), 63–81. https://doi.org/10.11139/cj.30.1.63-81

Taveggia, T. C. (1974). Resolving research controversy through empirical cumulation: Toward reliable sociological knowledge. *Sociological Methods & Research, 2*(4), 395–407.

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second Language Anxiety and Achievement: A meta-analysis. *Studies in Second Language Acquisition, 41*(2), 363–387. https://doi.org/10.1017/S0272263118000311

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

Thornton, P., & Houser, C. (2005). Using mobile phones in English education in Japan. *Journal of Computer Assisted Learning, 21*(3), 217–228. https://doi.org/10.1111/j.1365-2729.2005.00129.x

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching, 31*(2), 57–71. https://doi.org/10.1017/S0261444800012970

Viechtbauer, W. (2020). metafor: Meta-Analysis Package for R (R package version 2.4-0). CRAN. https://github.com/wviechtb/metafor/issues

Xu, Z., Banerjee, M., Ramirez, G., Zhu, G., & Wijekumar, K.. (2019). The effectiveness of educational technology applications on adult English language learners' writing quality: A meta-analysis. *Computer Assisted Language Learning*, *32*(1–2), 132–162. https://doi.org/10.1080/09588221.2018.1501069

Yang. (2017). Effects of Attention Cueing on Learning Speech Organ Operation through Mobile Phones. *Journal of Educational Technology & Society, 20*(4), 112–125.

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*(1), 39–58. https://doi.org/10.1080/09588221.2010.523285

Zhao, Y. (2003). Recent Developments in Technology and Language Learning: A Literature Review and Meta-analysis. *CALICO Journal*, *21*(1), 7–27. JSTOR.

Ziegler, N. (2015). Interaction in synchronous computer-mediated communication. In A Mackey, E Marsden (Eds.), *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages* (pp.197–223). Routledge.

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553–586. https://doi.org/10.1017/S027226311500025X

Ziegler, N., Meurers, D., Rebuschat, P., Ruiz, S., Moreno-Vega, J. L., Chinkina, M., Li, W., & Grey, S. (2017). Interdisciplinary Research at the Intersection of CALL, NLP, and SLA: Methodological Implications From an Input Enhancement Project: Research at the Intersection of CALL, NLP, and SLA. *Language Learning*, *67*(S1), 209–231. https://doi.org/10.1111/lang.12227

Appendices

Appendix A

Inclusion Criterion for Searched Studies

Inclusion criteria	Searched study were included in meta-analysis if all conditions were met including:
	• The study used a strict experimental design implying random selection and assignment of participants. • The study included adequate reports of independent variables including (a) at least one type of ICT tool in L2 intervention; (b) targeted at specific language skills; (c) included treatment and control groups. • The study included adequate reports of dependent variables, measured participants' performance on objective language testing instruments using the same quantitative tests for treatment and control groups. • The study included primary studies published between 2016 and 2020. Studies published after May 2020 were not available in published form at the time of searching.

Exclusion criteria	Searched study were excluded for meta-analysis if any condition were met including:
	• The study did not provide sufficient (missing any of the statistics including sample size, mean of the measured outcome, and corresponding standard deviations) statistical information to support effect size calculation. • The study was primarily descriptive, focusing on the implementation of ICT in L2 classrooms. • The study did not incorporate a placebo/control group (in experimental design studies). • Dependent variables in the study did not measure L2 skills as part of the learning outcomes of instructional treatments. • The study was a conference proceeding publication or technical report that lack of rigorous methodology, formal theoretical design or peer review.

Appendix B

Included Studies for Meta-analysis

Study ID	Included study
1	Ziegler, N., Meurers, D., Rebuschat, P., Ruiz, S., Moreno-Vega, J. L., Chinkina, M., ... & Grey, S. (2017). Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological implications from an input enhancement project. *Language Learning*, *67*(S1), 209-231.
12	Schenker, T. (2017). Synchronous telecollaboration for novice language learners: Effects on speaking skills and language learning interests. Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication, 20(2).
24	Yu, W., & Du, X. (2019). Implementation of a blended learning model in content-based EFL curriculum. International Journal of Emerging Technologies in Learning (iJET), 14(05), 188-199.
35	Choi, I. C. (2016). Efficacy of an ICALL tutoring system and process-oriented corrective feedback. Computer Assisted Language Learning, 29(2), 334-364.

50	de Vries, B. P., Cucchiarini, C., Bodnar, S., Strik, H., & van Hout, R. (2016). Effect of corrective feedback for learning verb second. International Review of Applied Linguistics in Language Teaching, 54(4), 347-386.
52	Zibin, A., & Altakhaineh, A. R. M. (2019). The effect of blended learning on the development of clause combining as an aspect of the acquisition of written discourse by Jordanian learners of English as a foreign language. Journal of Computer Assisted Learning, 35(2), 256-267.
59	Tang, X. (2019). The effects of task modality on L2 Chinese learners' pragmatic development: Computer-mediated written chat vs. face-to-face oral chat. System, 80, 48-59.
69	Ashiyan, Z., & Salehi, H. (2016). Impact of WhatsApp on learning and retention of collocation knowledge among Iranian EFL learners. Advances in Language and Literary Studies, 7(5), 112-127.
71	Granena, G. (2016). Individual versus interactive task-based performance through voice-based computer-mediated communication. Language Learning & Technology, 20(3), 40-59.
73	Chang, C. C., Warden, C. A., Liang, C., & Chou, P. N. (2018). Performance, cognitive load, and behaviour of technology-assisted

	English listening learning: From CALL to MALL. Journal of Computer Assisted Learning, 34(2), 105-114.
77	Carlotto, T., & Jaques, P. A. (2016). The effects of animated pedagogical agents in an English-as-a-foreign-language learning environment. International Journal of Human-Computer Studies, 95, 15-26.
79	Lee, H., Lee, H., & Lee, J. H. (2016). Evaluation of electronic and paper textual glosses on second language vocabulary learning and reading comprehension. The Asia-Pacific Education Researcher, 25(4), 499-507.
81	Lan, Y. J., & Lin, Y. T. (2016). Mobile seamless technology enhanced CSL oral communication. Journal of Educational Technology & Society, 19(3), 335-350.
93	Andújar-Vaca, A., & Cruz-Martínez, M. S. (2017). Mobile instant messaging: WhatsApp and its potential to develop oral skills. Comunicar. Media Education Research Journal, 25(1).
111	Alibakhshi, G., & Mohammadi, M. J. (2016). Synchronous and Asynchronous Multimedia and Iranian EFL Learners' Learning of Collocations. Applied Research on English Language, 5(2), 237-254.

113	Huang, X. (2019). WeChat-based teaching for an immersion cultural exchange program–a case study in CFL. Smart Learning Environments, 6(1), 7.
131	Lin, P. H., Liu, T. C., & Paas, F. (2017). Effects of spell checkers on English as a second language students' incidental spelling learning: a cognitive load perspective. Reading and Writing, 30(7), 1501-1525.
137	Rezaee, A. A., Alavi, S. M., & Razzaghifard, P. (2019). The impact of mobile-based dynamic assessment on improving EFL oral accuracy. Education and Information Technologies, 24(5), 3091-3105.
146	Calvo-Ferrer, J. R. (2017). Educational games as stand-alone learning tools and their motivational effect on L 2 vocabulary acquisition and perceived learning gains. British Journal of Educational Technology, 48(2), 264-278.
147	Awada, G. (2016). Effect of WhatsApp on critique writing proficiency and perceptions toward learning. Cogent Education, 3(1), 1264173.
149	Alhujaylan, H. (2019). An Assessment of the Effectiveness of CALL in Teaching English Language Writing Skills in Saudi Arabia. Arab World English Journal (AWEJ) Special Issue on CALL, (5).

160	Wang, Y. H. (2016). Promoting contextual vocabulary learning through an adaptive computer-assisted EFL reading system. Journal of Computer Assisted Learning, 32(4), 291-303.
161	Wu, T. T., & Huang, Y. M. (2017). A mobile game-based English vocabulary practice system based on portfolio analysis. Journal of Educational Technology & Society, 20(2), 265-277.
163	Mørch, A. I., Engeness, I., Cheng, V. C., Cheung, W. K., & Wong, K. C. (2017). EssayCritic: Writing to learn with a knowledge-based design critiquing system. Journal of Educational Technology & Society, 20(2), 213-223.
167	Shang, H. F., & Chen, Y. Y. (2018). The impact of online autonomous learning on EFL students' reading skills. International Journal on E-Learning, 17(2), 227-249.
168	Hwang, W. Y., Shih, T. K., Ma, Z. H., Shadiev, R., & Chen, S. Y. (2016). Evaluating listening and speaking skills in a mobile game-based learning environment with situational contexts. Computer Assisted Language Learning, 29(4), 639-657.
169	Parra, G. (2019). Automated Writing Evaluation Tools in the Improvement of the Writing Skill. International Journal of Instruction, 12(2), 209-226.

171	Chang, B., & Lu, F. C. (2018). Social media facilitated English prewriting activity design and evaluation. The Asia-Pacific Education Researcher, 27(1), 33-42.
177	Çakmak, F., & Erçetin, G. (2018). Effects of gloss type on text recall and incidental vocabulary learning in mobile-assisted L2 listening. ReCALL, 30(1), 24-47.
192	Yang, Y. F. (2016). Transforming and constructing academic knowledge through online peer feedback in summary writing. Computer Assisted Language Learning, 29(4), 683-702.

Appendix C

Boolean search term

(TitleCombined:(“informational and communication technology” OR “ICT” OR “E-learning” OR “educational technology” OR “mobile” OR “MALL” OR “computer” OR “CALL” OR “CAI” OR “CMC” OR “ICALL”, OR “blended learning” OR “telecollaboration” OR “natural language processing” OR “NLP” OR “learning management system” OR “LMS” OR “electro*” OR “informati*” OR “technolog*”) AND (“teaching” OR “learning” OR “training” OR “intervention” OR “instruction” OR “instructional treatment” OR “classroom” OR “programme” OR “implementation”) AND (“second language acquisition” OR “L2” OR “SLA” OR “foreign language” OR “FL” OR “second language” OR “SL”)))

Appendix D

Codebook of Included Studies

Table D1

Substantive features

Study ID	ICT type	Targeted L2 skill	Educational level	Language proficiency	L1	L2
1	ICALL	L2 grammar	post	intermediate	mixed	ESL
12	CAI	L2 speaking	secondary	beginner	English	German
24	CAI	mixed	post	intermediate	Chinese	EFL
35	ICALL	L2 grammar	secondary	beginner	Korean	EFL
50	ICALL	L2 speaking	post	beginner	mixed	Dutch
52	CAI	L2 writing	post	advanced	Arabic	EFL
59	CMC	L2 pragmatics	post	beginner	mixed	Chinese
69	MALL	L2 grammar	post	intermediate	Persian	EFL
71	CMC	L2 grammar	post	intermediate	Spanish	EFL
73	MALL	L2 listening	post	intermediate	Chinese	EFL
77	CAI	L2 grammar	post	intermediate	Portuguese	EFL
79	CAI	mixed	post	intermediate	Korean	EFL
81	MALL	L2 speaking	post	beginner	mixed	Chinese

93	MALL	L2 speaking	post	intermediate	Spanish	EFL
111	CMC	L2 grammar	post	intermediate	Persian	EFL
113	MALL	L2 pragmatics	post	beginner	English	Chinese
131	CAI	L2 vocabulary	post	intermediate	ESL	Chinese
137	MALL	L2 speaking	post	intermediate	Arabic	EFL
146	DGBL	L2 vocabulary	post	mixed	Spanish	EFL
147	MALL	L2 writing	post	advanced	Arabic	EFL
149	CAI	L2 writing	post	intermediate	Arabic	EFL
160	ICALL	L2 vocabulary	post	intermediate	Chinese	EFL
161	MALL	L2 vocabulary	post	miss	Chinese	EFL
163	ICALL	L2 writing	secondary	missing	Norwegian	EFL
167	ICALL	L2 reading	post	intermediate	Chinese	EFL
168	MALL	mixed	secondary	beginner	Chinese	EFL
169	ICALL	L2 writing	post	mixed	Spanish	EFL
171	MALL	L2 writing	post	intermediate	Chinese	EFL
177	MALL	L2 vocabulary	post	beginner	Turkish	EFL
192	CMC	L2 writng	post	intermediate	Chinese	EFL

Table D2

Methodological features

Study ID	Outcome measures	Treatment Duration
1	Mixed	Medium
12	Free-constructed-response	Medium
24	Mixed	Long
35	Mixed	Medium
50	Meta-linguistic judgement	Short
52	Free-constructed-response	Long
59	Mixed	Short
69	Selected-response	Missing
71	Constrained-constricted-response	Short
73	Meta-linguistic judgement	Medium
77	Free-constructed-response	Short
79	Mixed	Medium
81	Free-constructed-response	Medium
93	Free-constructed-response	Short
111	Selected-response	Medium
113	Mixed	Medium

131	Mixed	Medium
137	Free-constructed-response	Medium
146	Constrained-constricted-response	short
147	Free-constructed-response	Medium
149	Free-constructed-response	Medium
160	Constrained-constricted-response	Medium
161	Mixed	Long
163	Free-constructed-response	Short
167	Selected-response	Long
168	Mixed	Medium
169	Free-constructed-response	Medium
171	Free-constructed-response	Long
177	Mixed	Short
192	Free-constructed-response	Long

Note. Short treatment duration (less than 1.25 weeks), medium treatment duration (1.25 to 13.5 weeks) and long treatment duration (more than 13.5 weeks) were based on the cutoff points of the first and the third quartile of sample distribution.

Table D3

Reporting features

Study ID	Publication year	Sample size
1	2017	Medium
12	2017	Small
24	2019	Medium
35	2016	Large
50	2016	Small
52	2019	Medium
59	2019	Small
69	2016	Medium
71	2016	Large
73	2017	Large
77	2018	Medium
79	2016	Large
81	2016	Small
93	2017	Large
111	2016	Large
113	2019	Small

131	2017	Large
137	2019	Large
146	2017	Medium
147	2016	Medium
149	2019	Medium
160	2017	Medium
161	2017	Medium
163	2017	Medium
167	2018	Medium
168	2016	Small
169	2019	Small
171	2017	Large
177	2017	Large
192	2016	Small

Note. Small sample size (less than 43 participants), medium sample size (43 to 80 participants) and large sample size (more than 80 participants) were based on the cutoff points of the first and the third quartile of sample distribution.