*Article*

LANGUAGE TEACHING RESEARCH

Language Teaching Research
1–45

DOI: 10.1177/1362168820988423
journals.sagepub.com/home/ltr
SAGE

# Revisiting the effectiveness of study abroad language programs: A multi-level meta-analysis

**Wen-Ta Tseng**
National Taiwan University of Science and Technology

**Yeu-Ting Liu**
**Yi-Ting Hsu**
**Hsi-Chin Chu**
National Taiwan Normal University

## Abstract

This study set out to re-examine the effectiveness of study abroad programs in second language (L2) acquisition through a multi-level meta-analysis. Overall, 42 primary studies published between 1995 and 2019 were identified, and in total 283 effect sizes were meta-analysed. This study implemented a three-level random effects model to account for the clustered, mutually dependent effect sizes often nested in the primary studies of L2 study abroad research. The results indicated a medium-to-large effect ($g = 0.87$) on study abroad language programs. Essentially, the featured moderators in general explained more heterogeneity variances at level 3 (i.e. the between-study level) than at level 2 (i.e. the within study level). For study abroad language learners, language acquisition is optimal when learners, in particular those of a lower proficiency level, take both formal and content-based language courses while living with host families. Learners' age and pre-program training may not moderate the effectiveness of study abroad language programs. Importantly, this study further established that the length of study abroad programs are positively associated with learners' language gains, but that an extended and prolonged domestic program does not necessarily lead to such gains. Research and pedagogical implications are further discussed based on the research findings.

## Keywords

effect size, meta-analysis, research synthesis, second language acquisition, study abroad

**Corresponding author:**
Yeu-Ting Liu, Department of English, National Taiwan Normal University, No. 162, Sec. 1, HePing East Road, Da'an District, Taipei, 106
Email: yeutingliu@ntnu.edu.tw

## I Introduction

Learning context has been recognized as a key factor influencing second language acquisition (SLA) since the 1970s (Hymes, 1974). In particular, the last few decades have seen a sharp rise in empirical research investigating the effects of study abroad language programs on L2 acquisition (Dekeyser, 1991; Freed, 1995; Freed, Segalowitz & Dewey, 2004; Hirakawa, Shibuya, & Endo, 2019; Llanes, Mora & Serrano, 2017; Schenker, 2018). It is generally believed that studying a language abroad provides L2 learners with opportunities to employ the target language in real-life situations including through local media and cultural events than studying the L2 domestically. There has been conflicting evidence found in empirical studies which failed to confirm the relationships between study abroad language learning experiences and L2 linguistic development. For instance, while some studies indicated that learners who studied the target language abroad significantly outperformed those who learned the language in their home countries (Martinsen et al., 2011; Segalowitz et al., 2004), other research results showed no substantial differences among L2 learners in foreign and domestic contexts (Freed, 1990; Serrano et al., 2011). The multidimensional nature of the L2 learning experience in both study abroad and domestic settings comprises an assortment of variables which may affect overall L2 linguistic development. Due to the large amount of research being carried out in this area, it appears to be necessary to conduct a research synthesis which explores the usefulness of study abroad programs in order to pinpoint variables which moderate success or shortcomings among programs.

Over the past decade, there have been a number of narrative syntheses of research on language learning studies abroad (Borràs & Llanes, 2019; Llanes, 2011; Tullock & Ortega, 2017; Xiao, 2015) as well as several meta-analyses undertaken to examine the effect of studying abroad on linguistic gains (Hirai, 2018; Varela, 2017; Xu, 2019; Yang, 2016). The common core value upheld by both the narrative syntheses and meta-analyses is that studying abroad stands as an important learning context in which quality and quantity of linguistic input, practice, and instruction can be uniquely proffered compared to the typical counterpart (i.e. at home context). This premise has been extensively evidenced in the methodological design of the existing study abroad research where the pedagogical potency of study abroad programs is predominantly examined vis-à-vis at home programs. As Borràs and Llanes (2019, p. 1) succinctly argue, 'The study abroad (SA) context is believed to be one of the most favorable contexts for second language (L2) learning.'

It is worth noting that among the four meta-analyses conducted to date, inconsistent results were observed when examining the overall effect of L2 development (Hirai, 2018; Varela, 2017; Xu, 2019; Yang, 2016). For instance, Yang's study achieved a medium effect size ($d = 0.75$), while Varela's study uncovered a larger effect size ($d = 0.975$). A less promising value ($g = 0.56$) was obtained in Hirai's investigation, and in Xu's analysis an even smaller value ($d = 0.37$) was observed. Inconsistent results were also noted regarding length of study. Yang's (2016) meta-analysis indicated that short-term study abroad programs are more beneficial than long-term programs, whereas both Varela's (2017) and Hirai's (2018) studies revealed that learners in long-length study abroad programs outperformed those in medium- and short-length programs.

Apart from these inconsistent findings, it should be noted that none of the aforementioned meta-analyses of study abroad account for the potential impact of the clustered, mutually-dependent effect sizes reported. Tullock and Ortega's (2017) scoping review found that up to 75 diverse and distinct methodological operationalizations were created as oral fluency measures observed in the 31 primary studies that were collected. This suggested that each primary study had on average 2.5 effect sizes with noticeable amounts of heterogeneity of variance due to methodological inconsistency. In other words, the presence of multiple, distinct effect sizes existing in every single primary study prevented the authors from synthesizing and quantifying the overall average effect of studying abroad on oral fluency. This unexpected outcome led Tullock and Ortega to remark that 'little conceptual clarity can accumulate regarding the gains that studying abroad brings about in the area of oral fluency unless the measurement is carefully reasoned within and across studies' (p. 13).

This critical remark directly points to the necessity of using a more adequate modeling framework to account for the possible methodological inconsistencies existing within and across primary studies on studying abroad. Furthermore, these remarks highlight the need for a further check of the degree to which featured moderators may account for the variability of effect sizes at both the within- and between-study levels. To bridge the research gap, the current meta-analysis aims to adopt a multi-level modeling approach (Cheung, 2015) to examine the impact of the study abroad context through a more meticulous and detailed lens by implementing rigorous exclusion and inclusion criteria, employing comprehensive meta-analytic procedures with the use of heterogeneity tests, exploring moderator variables that were excluded from previous meta-analyses, and drawing primary studies from a wider range of publication dates.

## II Literature review

### *1 Research on study abroad*

Over nearly three decades, there has been an increase in research investigating the effectiveness of study abroad programs on varied facets of language development. The foci of the prior empirical studies include the development of speaking (Freed, 1995; Llanes, 2012; Llanes et al., 2017; Solon & Long, 2018), reading (Dewey, 2004; Li, 2014), writing (Llanes, 2012; Sasaki, 2007; Serrano at al., 2016) and listening (Cubillos, Chieffo, & Fan, 2008; Taguchi, 2011) proficiency. Additional research investigated the acquisition of vocabulary knowledge (Zaytseva, Pérez-Vidal, & Miralpeix, 2018), utilizing new and different communication strategies (Lafford, 2004; Yashima & Zenuk-Nishide, 2008), and how L2 learners can gain proficiency in terms of grammatical (Isabelli-García, 2010; Marqués-Pascual, 2011; Sagarra & LaBrozzi, 2018;), pragmatic (Ren, 2015; Taguchi, 2011), and sociolinguistic abilities (Barron, 2006; Regan, 1995). Studies have also looked into how motivation (Hernández, 2010; Isabelli-García, 2010; Sasaki, 2011), attitude (Geeslin & Schmidt, 2018), confidence (Martinsen et al., 2011) and learning strategy use (Watson & Ebner, 2018) changed both before, during, or after the learners' stay in the target language community. Overall, relevant primary studies on studying abroad have shown numerous benefits for L2 learners.

## 2 Related meta-analyses on the effects of study abroad programs

Although there have been numerous meta-analyses undertaken over the past 2 decades in the context of SLA, meta-analysis that set out to examine the effects of study abroad programs are still new and relatively scarce. To date, as noted earlier, there have only been four empirical meta-analyses conducted to examine the effect of studying abroad on language development (Hirai, 2018; Varela, 2017; Xu, 2019; Yang, 2016). Although these four meta-analytic inquires shared a common statistical model (i.e. random effects model) and unanimously recognized the value-added role of studying abroad in facilitating L2 language development, they differed greatly from one another in terms of the focus of language development, the number of primary studies collected, the use of inclusion/exclusion criteria, and the number of moderators.

In the first study, Yang (2016) conducted a meta-analysis of 11 primary studies published between 2004 and 2011 to investigate the overall effect of studying abroad on general language proficiency. Yang's study compared the effect of the study abroad learning context with that of the at home learning context and found a medium effect size of $d = 0.75$, suggesting that study abroad groups outperformed at home groups in terms of L2 linguistic enhancement after their sojourn experience. Yang's study was the first empirical undertaking in the literature that reliably informed the overall trend regarding the effect of studying abroad on L2 development. In Yang's study, the duration of study abroad language programs was the only moderator in the analysis. The average effect size for short-term study abroad residence (from 11 to 13 weeks) was $d = 0.981$, and for long-term study abroad residence (more than 13 weeks) an effect size of $d = 0.458$ was found. These results imply a greater, and more positive, impact for a short-term study abroad experience.

Second, Varela's (2017) analysis also examined the effect of study abroad programs on general language proficiency, and the at home context served as the baseline upon which the comparable effect of the study abroad context could be clearly identified: a practice reminiscent of Yang (2016). However, Varela's analysis further included the studies of a within-group design (i.e. only the study-abroad group). The included studies in Varela's (2017) meta-analysis were published between 2001 and 2015. Three learning areas - cognitive acquisition, affective attitudes, and behavioral adaptation - were reported among all 33 study abroad primary studies. Varela's results showed that studying abroad yields tangible learning outcomes for cognition ($d = 0.975$), affect ($d = 0.457$), and behavior ($d = 0.649$), respectively. Varela's study first examined the moderating effect of college major and a larger effect size was obtained from language majors ($d = 1.116$) than the general mix of majors ($d = 0.975$), suggesting a larger effect for the former in terms of language acquisition. In particular, business majors ($d = 0.754$) seemed to outperform all other majors combined ($d = 0.457$) in attitudinal learning, while compatible results (business majors $d = 0.666$; mix of majors $d = 0.649$) were discovered in behavioral learning. However, the overlaps in 95% confidence intervals (CI) of effect sizes indicates no statistically significant difference between groups, suggesting that a learner's college major plays a marginal role in learning.

Varela (2017) further tested the effects of four possible moderators of learning in the study abroad experience: cultural distance, program content (i.e. in-class or out-of-class

designs), type of immersion (i.e. staying with host families or in the dormitory), and time of sojourn which yielded a mixed result. First, no correlation was found between cultural distance and learning outcomes. Second, when comparing program content, out-of-class design facilitated attitudinal learning while no significance was found between the two groups in behavioral learning due to a slight overlap of 95% CIs. Third, in terms of language acquisition, students sojourning with host families ($d = 1.305$) failed to outgain their counterparts staying in dormitories ($d = 0.751$) due to the overlap of CIs for both effect sizes (families = [0.479, 2.129]; dorms = [0.480, 1.022]). Contrary findings obtained in attitudinal learning suggest that homestay groups ($d = 0.577$; 95% CI = [0.319, 0.836]) acquired more linguistic competence than dormitory groups ($d = 0.560$; CI = [–.02, 1.14]). Additionally, due to the fact that only one study included learners in a homestay group, the result of behavioral learning remained unclear. Fourth, the modulating role of length of study abroad programs was investigated in Varela's analysis, just as it was in Yang's (2016) study. Contrary to Yang's findings, a long-term program ($>$ 1 semester; $d = 1.731$) facilitated learners' linguistic development more than mid-term (between 8 weeks and 1 semester; $d = 0.916$) and short-term ($<$ 8 weeks; $d = 0.604$) programs; however, no significance was found in attitudinal or behavioral learning with regards to length of program.

Still another, more recent meta-analysis of study abroad research was undertaken by Hirai (2018) who meta-analysed the effects of studying abroad by examining 31 primary studies that were all conducted in Japan's educational context. The studies included in Hirai's analysis were exclusively those of adopting a pretest–posttest experimental design (i.e. including both single-group and control-group designs with a pretest–posttest study feature). The results of Hirai's study showed that studying abroad could achieve a small-to-medium effect ($g = 0.56$) and revealed a trend showing a longer study abroad duration leads to better linguistic gains. Specifically, a long-term study abroad duration (6 months - 1 year) procured the largest effect size ($g = 1.77$), followed by the mid-term study abroad (1 month ~ 6 months) with $g = 0.82$, and a short-term study abroad (less than or equal to 1 month) led to the smallest effect size ($g = 0.36$). Furthermore, Hirai also found that the effects of studying abroad in the L1 (Japanese) context might be dampened by learners' predeparture L2 preparatory training. It should be pointed out that 29 of the 31 studies included in Hirai's meta-analysis were published in Japanese rather than English.

The fourth meta-analysis of study abroad research was contributed by Xu (2019). Unlike the three predecessors who focused on general language proficiency, Xu narrowed down the scope of analysis and specifically looked into the effects of studying abroad on interlanguage complexity development. In Xu's study, 28 primary studies were included and, as in Hirai's (2018) study, the primary studies included were those of both single-group and control-group experimental design with a pretest–posttest study feature. In comparison to Hirai's study results, Xu uncovered a smaller effect size ($d = 0.37$). To elaborate, the results of Xu's study revealed that studying abroad had a larger effect on oral complexity development ($d = 0.41$) than written complexity development ($d = 0.31$), and that studying abroad had a larger effect on lexical complexity development ($d = 0.29$) than syntactical complexity ($d = 0.2$).

Despite the significant findings and contributions of the four prior meta-analysis, a number of limitations still exist and many matters are yet to be examined. First, Yang's (2016) research only examined overall increase in L2 linguistic ability. Possible moderating effects of salient language features and various types of outcome measures have yet to be systematically addressed, organized, and analysed. Additionally, the only moderator examined in the study was length of sojourn, leaving many other important individual and programmatic factors potentially influencing the effects of study abroad unexplored. Furthermore, although Varela's (2017) analysis tested more moderators than Yang's study including cultural distance, program content, type of immersion, and time of sojourn, many subgroup classifications of the moderators needed to be more precisely operationalized. For instance, the classification of time spent abroad is mostly operationalized as categorical variable such as short-, mid-, and long-term, and this makes it difficult to compare across primary studies due to the fact that one's perception about the relative length of the program may vary from one to another. To resolve this classification inconsistency, it is necessary to operationalize the length of the study abroad as a continuous variable rather than a discrete variable. For instance, using the basic operational time frame/units (e.g. number of weeks/months) commonly adopted in the study abroad programs may be a possibility to operationalize the length of study abroad programs as a continuous variable. Additionally, more individual differences in language learning such as learners' language proficiency, age, and pre-program language training should also be considered and examined in order to understand their potential influences on the effects of study abroad programs.

Notably, the two more recent studies (Hirai, 2018; Xu, 2019) are not only restricted in the number of moderators identified, but also constrained by the inclusion criteria commonly seen in recent meta-analysis (i.e. Varela, 2017). That is, Hirai and Xu's investigations included primary studies which utilized within-group design features as well as studies that used between-group design features. For example, some studies examined single-group (e.g. Leonard & Shea, 2017; Mora & Valls-Ferrer, 2012), while others focused on between-group (e.g. Llanes & Muñoz, 2013; Serrano, Llanes & Tragant, 2011) pretest–posttest experimental designs.

According to Marsden and Torgerson (2012), although this inclusion criterion might increase the number of primary studies, the use of a single-group, pretest–posttest experimental design may render the results of primary studies difficult to interpret. To be specific, Marsden and Torgerson pointed out that the outcome derived from the single group, pretest–posttest research design may likely regress toward the mean (RTM) because participants with lower baseline scores tend to make greater improvement over those with higher baseline scores and vice versa. This might be evidenced by the smaller overall effect sizes observed in Hirai's (2018) and Xu's (2019) studies in comparison to Yang's (2016). Without the presence of a comparable baseline group (e.g. at home context), forming a causal link between study abroad programs and linguistic gains based on the single group, pretest–posttest research design may be problematic.

Finally, a considerable number of the related primary studies were omitted from the four meta-analysis (for an overview, see Table 3 below; for a cross-study comparison, see Appendix 1). The current meta-analysis in total includes 42 primary studies, all of which involve the effect comparisons between study abroad context and at home context.

However, the number of primary studies adopting the same type of experimental design in the three prior meta-analyses is far fewer: 9 in Yang's (2016), 15 in Varela's (2017), and 4 in Xu's (2019) study, respectively. Appendix 1 lists the detailed table to show the gap of number of primary studies between the prior works and the current study. Clearly, the aim of the study is by no means just reanalysing the earlier studies, but conducting a new meta-analysis with a more comprehensive coverage of primary studies. More importantly, several mutually dependent effect sizes are typically observed in numerous primary studies of L2 study abroad research (Llanes & Muñoz, 2013; Ren, 2015; Schenker, 2018; Segalowitz & Freed, 2004; Serrano, Llanes, & Tragant, 2016; Taguchi, 2011). The conventional meta-analytic approach cannot account for such a clustered type of data format (Assink & Wibbelink, 2016). Recent meta-analytic research has referred to 'three-level meta-analysis' as an efficient method to handle the clustered effect sizes nested in primary studies (Cheung, 2015); therefore, the statistical validity of the two related meta-analysis still remains further corroborated.

In sum, the current meta-analysis seeks to fill these gaps in the study abroad literature by (1) including more recent studies which were not considered in previous meta-analyses; (2) identifying salient programmatic and individual variables since the effects of these factors were not sufficiently examined and explained; (3) providing a more precise definition and classification for each moderator; and (4) implementing a 'three-level meta-analysis' to account for the clustered effect sizes typically seen in the primary studies.

## III Variables and research questions

The moderator variables included in the current meta-analysis refer to type of outcome measure, type of residence, length of treatment, learners' language proficiency, measure of proficiency, learners' age, pre-program training, program content, and type of domestic language program. Likewise, the single dependent variable of effect size was derived from the primary studies collected. Under the implementation of a three-level meta-analytic framework, the current meta-analysis seeks to answer the following research questions:

- Research question 1: What are the overall effects of study abroad programs on linguistic gains?
- Research question 2: To what extent do the effects of studying abroad vary across different moderating factors identified in the study?

## IV Method

### *1 Identifying primary studies*

Pertinent primary studies were compiled from a range of reputable sources. First, studies were collected from well-known applied linguistics and educational databases such as the Educational Resource Information Center (ERIC), ProQuest, Psyclnfo, Google Scholar, and dissertation abstracts. Second, search terms included combinations of the

following terms: study abroad, second language learning, second language acquisition, international experience, mobility program and so forth. Third, electronic searches were carried out across past and current issues of journals which are highly regarded in the fields of applied linguistics and SLA including *Applied Linguistics*, *Foreign Language Annals*, *Language Learning and Technology*, *Language Learning*, *International Review of Applied Linguistics*, *Language Teaching Research*, *Studies in Second Language Acquisition*, *System*, *TESOL Quarterly*, and *The Modern Language Journal*, and, in particular, *Frontiers: The Interdisciplinary Journal of Study Abroad*, which mainly publishes articles related to education abroad. Fourth, course books, edited books, and chapters from books associated with study abroad language learning programs (DuFon & Churchill, 2006; Freed, 1995; Kinginger, 2009; Regan, Howard & Lemée, 2009; Ren, 2015; Wilkinson, 2006) were scanned, including their reference sections, in order to find promising primary research studies. Lastly, reference sections from previously published meta-analyses on study abroad language programs were carefully examined (Varela, 2017; Yang, 2016). In order to maintain search consistency and reliability, the following sets of key terms were combined and applied across different databases: 'study abroad' or 'language learning' or 'foreign language learning' or 'control group'.

## 2 Inclusion/exclusion criteria

Studies fulfilling the following criteria were included:

1. Studies with experimental-control-group designs (i.e. study abroad versus at-home contexts) were included in order to observe the comparable learning effect after treatment between the treatment and control groups.
2. The independent variables in the primary studies involved collective, shared experiences of overseas residence in a target-language setting.
3. The dependent variables in the primary studies involved measurements of participants' linguistic performance.
4. Two studies, reprinted across several sources or based on the same sample, were used only once in this meta-analysis (i.e. a journal article was published based on a dissertation).
5. The quantitative results of the study could be transformed into computation of *g* effect sizes.

Studies were excluded if they had one of the following features:

1. Linguistic outcome was not measured in the primary studies. That is, some studies investigated the construction of participants' national identity after their sojourn abroad (Diao, 2017; France & Rogers, 2012; Huang, 2015). Although study abroad experiences might possibly influence learners' national identity, this was not the focus of the present meta-analysis.
2. The research design lacked a control group despite the adoption of pretest–posttest designs. In order to make a meaningful comparison between study abroad

language programs and domestic language programs, the inclusion of a control group was necessary.
3. The impact of the study abroad experience was only investigated with regards to gaining cultural familiarity instead of linguistic understanding or practical language comprehension.
4. The reported statistical information could not be transformed into effect sizes.

## 3 Moderators

Selected primary studies were thoroughly investigated in order to compile a coding system to identify and record elements such as dependent and independent variables, methodological features, learner characteristics, and programmatic factors. These studies were then labeled and classified meticulously, as shown in Table 1. Primary studies withstood numerous cycles of coding in order to ensure inter-rater reliability. As shown in Appendix 2, individual kappa indices of all the 10 moderators ranged between 0.82 and 0.98, achieving an overall average kappa index of 0.92. Table 2 further reported the detailed coding structure for each primary study. Evidently, the need to carefully examine the overall average of effect size is clear, but the justification for examining the moderator variables needs to be further implemented. Thus, the rationale behind this inquiry is discussed in detail below.

## 4 Type of outcome measure

Treatment effect measures were coded mainly following Borràs and Llanes's (2019) category structure to understand how study abroad language programs may influence different facets of language gains. In their category structure, Borràs & Llanes listed seven linguistic domains where studying abroad may influence language learning: general L2 proficiency, vocabulary, grammar, oral skills, listening, reading, and writing. However, three noteworthy yet largely overlooked types of outcome measures were not reviewed by Borràs and Llanes: lexical-grammatical knowledge (Segalowitz et al., 2004), pragmatic knowledge (Ren, 2015; Taguchi, 2011) and working memory (Sagarra & LaBrozzi, 2018; Sunderman & Kroll, 2009). In total, up to 10 types of outcome measures were critically examined in the current study. According to Norris and Ortega (2000), outcome assessment can moderate the effects of L2 instruction. Unfortunately, none of prior meta-analyses of studying abroad considered the effects of the operationalizations of different linguistic measures such as how different operationalizations of linguistic attainments may moderate the effects of study abroad programs. Therefore, meta-analysing the moderating effects of different types of outcome measure is one of the objectives of this study.

## 5 Learners' language proficiency

Learners' pre-departure language proficiency level was coded according to the information stated in the primary studies (i.e. beginner, intermediate, advanced, and mixed). It should be noted that the measures of learners' language proficiency varied among the

**Table 1.** The coding scheme of this meta-analysis.

| Features | Descriptors |
|---|---|
| Type of outcome measure | a. Vocabulary knowledge$_{receptive}$<br>b. Grammar<br>c. Lexical-grammatical knowledge<br>d. Pragmatic knowledge<br>e. Listening<br>f. Reading<br>g. Speaking<br>h. Writing<br>i. Working memory<br>j. Mixed_01(Listening, Grammar, Reading)<br>k. Mixed_02(Listening, Grammar, Reading, Translation, Writing) |
| Learners' language proficiency | a. Beginner<br>b. Intermediate<br>c. Advance<br>d. Mixed |
| Test mechanisms | a. In-house assessment<br>b. Standardized test<br>c. ACTFL-OPI<br>d. ILR-OPI<br>e. Home spun OPI |
| Type of residence | a. Host family<br>b. School-based dormitory<br>c. Non-school-based dormitory<br>d. Mixed_01(Host family, School-based dormitory)<br>e. Mixed_02(Host family, Non-school-based dormitory)<br>f. Mixed_03(School-based dormitory, Non-school-based dormitory)<br>g. Mixed_04(Host family, School-based dormitory, Non-school-based dormitory) |
| Program content | a. Content-based course<br>b. Formal language course<br>c. Mixed |
| Learners' age | Average age of participants |
| Length of treatment | Average weeks of the study abroad program |
| Length of AH (hour) | At Home program (hours per week) |
| Target language | a. English<br>b. Chinese<br>c. Spanish<br>d. French<br>e. German<br>f. Japanese<br>g. Russian;<br>h. Mixed_01(Arabic, Chinese, Japanese, Russian)<br>i. Mixed_02(French, German, and Spanish) |
| Pre-program training | a. Yes<br>b. No |

*Notes*. ACTFL-OPI = Oral Proficiency Interview following the scale of *American Council on the Teaching of Foreign Languages*. ILR-OPI = Oral Proficiency Interview following Interagency Language Roundtable scale.

**Table 2.** The primary studies meta-analyzed in the study.

| Study | Year | Total N | Target language | Length of SA | Type of residence | Program content | Outcome measure | Learner's age | Language proficiency | Measure of proficiency | Domestic Program | Pre-program training |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hirakawa, Shibuya, & Endo | 2019 | 69 | English | 5 weeks | H | F | G | 19.35 | Intermediate | IA | 3 weeks | No |
| (2 samples) | | | English | 4 weeks | H | F | G | 19.31 | Intermediate | IA | 3 weeks | No |
| Sagarra & LaBrozzi | 2018 | 47 | Spanish | 16 weeks | H | M | WM | 25 | Basic | IA | n/a | No |
| Schenker | 2018 | 92 | German | 4 weeks | H | F | G, L, R, V, P | 20 | Advanced | ST | 16 weeks | Yes |
| Winke & Gass | 2018 | 296 | Chinese | 8 weeks | Mixed | Mixed | S | 20 | Intermediate | ACTFL-OPI | n/a | No |
| (3 samples) | | | French | n/a | Mixed | Mixed | S | 20 | Intermediate | ACTFL-OPI | n/a | No |
| | | | Spanish | n/a | Mixed | Mixed | S | 20 | Intermediate | ACTFL-OPI | n/a | No |
| Serrano, Tragant & Llanes | 2017 | 112 | English | 3 weeks | onD | F | W | 14.37 | Intermediate | IA | 4 weeks | No |
| Wu & Zhang | 2017 | 94 | English | 72 weeks | n/a | C | W | 24 | Advanced | IA | n/a | No |
| (2 samples) | | | English | 72 weeks | n/a | C | W | 24 | Advanced | IA | n/a | No |
| Köylü | 2016 | 46 | English | 16 weeks | Mixed | Mixed | W, S | 23 | Intermediate | IA | n/a | No |
| (2 samples) | | | English | 16 weeks | Mixed | Mixed | W, S G | 23 | Intermediate | IA | n/a | No |
| Serrano et al. | 2016 | 112 | English | 3 weeks | onD | Mixed | S | 14 | Intermediate | IA | 4 weeks | No |
| Llanes et al. | 2017 | 36 | English | 3 weeks | Mixed | C | S | 15 | Intermediate | IA | 4 weeks | No |
| Félix-Brasdefer & Hasler-Barker | 2015 | 37 | Spanish | 8 weeks | H | Mixed | S | 21 | Intermediate | IA | 8 weeks | No |
| Ren | 2015 | 40 | English | 40 weeks | n/a | C | P | 24.5 | Advanced | IA | n/a | No |
| Muñoz & Llanes | 2014 | 55 | English | 12 weeks | H | Mixed | S | 10.4 | Basic | IA | 12 weeks | No |
| (2 samples) | | | English | 12 weeks | offD | C | S | 22.36 | Advanced | IA | 12 weeks | No |
| Li | 2014 | 73 | Chinese | 8 weeks | n/a | Mixed | L, R, G, T, W | 21 | Basic | IA, ST | 16 weeks | No |
| (3 samples) | | | Chinese | 8 weeks | n/a | Mixed | L, R, G, T, W | 21 | Intermediate | IA, ST | 16 weeks | No |
| | | | Chinese | 8 weeks | n/a | Mixed | L, R, G, T, W | 21 | Advanced | IA, ST | 16 weeks | No |
| Llanes & Serrano | 2017 | 197 | English | 10 weeks | H | Mixed | W, S | 10.5 | Basic | IA | n/a | No |
| (3 samples) | | | English | 8 weeks | H | F | W, S | 13.5 | Intermediate | IA | n/a | No |
| | | | English | 12 weeks | Mixed | F | W, S | 20.5 | Advanced | IA | n/a | No |

*(Continued)*

**Table 2.** (Continued)

| Study | Year | Total N | Target language | Length of SA | Type of residence | Program content | Outcome measure | Learner's age | Language proficiency | Measure of proficiency | Domestic Program | Pre-program training |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Jochum | 2014 | 27 | Spanish | 16 weeks | H | F | S | 21 | Intermediate | ACTFL-OPI | 40 weeks | No |
| Llanes & Muñoz | 2013 | 139 | English | 10 weeks | H | Mixed | W, S | 11 | Basic | IA | n/a | No |
| (2 samples) | | | English | 10 weeks | Mixed | F | W, S | 21 | Advanced | IA | n/a | No |
| Llanes | 2012 | 16 | English | 8 weeks | H | Mixed | W, S | 11 | Basic | IA | n/a | No |
| Taguchi | 2011 | 62 | English | 52 weeks | n/a | C | P | 22 | Advanced | IA | 40 weeks | No |
| (2 samples) | | | English | 52 weeks | n/a | C | P | 22 | Advanced | IA | 40 weeks | No |
| Serrano et al. | 2011 | 131 | English | 52 weeks | n/a | Mixed | S, W | 20.5 | Intermediate | IA | 10 weeks | Yes |
| (2 samples) | | | English | 52 weeks | n/a | Mixed | S, W | 20.5 | Intermediate | IA | 4.5 weeks | Yes |
| Sasaki | 2011 | 37 | English | 7 weeks | n/a | F | W | 19 | Intermediate | IA | 52 weeks | Yes |
| (3 samples) | | | English | 16 weeks | n/a | Mixed | W | 19 | Advanced | IA | 52 weeks | Yes |
| | | | English | 38 weeks | n/a | Mixed | W | 20 | Advanced | IA | 130 weeks | Yes |
| Martinsen et al. | 2011 | 78 | French | 24 weeks | offD | F | S | 21 | Mixed | ACTFL-OPI | n/a | No |
| (4 samples) | | | German | 24 weeks | offD | F | S | 21.2 | Mixed | ACTFL-OPI | n/a | No |
| | | | Japanese | 24 weeks | offD | F | S | 21.83 | Mixed | ACTFL-OPI | n/a | No |
| | | | Russian | 24 weeks | offD | F | S | 20.965 | Advanced | ACTFL-OPI | n/a | No |
| Marqués-Pascual | 2011 | 40 | Spanish | 16 weeks | n/a | F | S | 20 | Intermediate | IA | 16 weeks | No |
| (2 samples) | | | Spanish | 16 weeks | n/a | F | S | 20 | Advanced | IA | 16 weeks | No |
| Jiménez-Jiménez | 2010 | 81 | Spanish | 20 weeks | n/a | F | V | 20 | Advanced | IA | n/a | n/a |
| (2samples) | | | Spanish | 60 weeks | n/a | F | V | 20 | Advanced | IA | n/a | n/a |
| Isabelli-García | 2010 | 24 | Spanish | 16 weeks | n/a | C | G | 19.5 | Intermediate | IA | 16 weeks | No |
| Foster | 2009 | 100 | English | 48 weeks | n/a | Mixed | S | 32.5 | Intermediate | IA | n/a | No |
| Sasaki | 2009 | 22 | English | 8 weeks | n/a | Mixed | W | 19 | Intermediate | IA | 182 weeks | No |
| (3 samples) | | | English | 16 weeks | n/a | Mixed | W | 19 | Intermediate | IA | 182 weeks | No |
| | | | English | 22 weeks | n/a | Mixed | W | 19 | Intermediate | IA | 182 weeks | No |
| Sunderman & Kroll | 2009 | 48 | Spanish | 15.2 weeks | n/a | n/a | WM | 20.65 | Intermediate | IA | n/a | n/a |
| Vande Berg et al. | 2009 | 1297 | Arabic, Chinese, French, German, Japanese, Spanish | 18.75 weeks | Mixed | Mixed | S | 20 | Intermediate | HOPI | n/a | No |

*(Continued)*

**Table 2.** (Continued)

| Study | Year | Total N | Target language | Length of SA | Type of residence | Program content | Outcome measure | Learner's age | Language proficiency | Measure of proficiency | Domestic Program | Pre-program training |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dewey | 2008 | 56 | Japanese | 11 weeks | H | F | V | 20 | Intermediate | IA | n/a | No |
| (2 samples) | | | Japanese | 11 weeks | H | F | V | 20 | Intermediate | IA | n/a | No |
| Cubillos et al. | 2008 | 140 | Spanish | 5 weeks | n/a | F | L | 20 | Intermediate | IA | 5 weeks | No |
| O'Brien et al. | 2007 | 43 | Spanish | 13 weeks | n/a | F | S | 21.84 | Mixed | ACTFL-OPI | 13 weeks | No |
| Sasaki | 2007 | 13 | English | 28.57 weeks | onD | Mixed | W, L | 20 | Intermediate | IA, ST | 16 weeks | No |
| Sasaki | 2004 | 11 | English | 22.67 weeks | n/a | Mixed | W | 18 | Basic | IA | 182 weeks | No |
| Segalowitz et al. | 2004 | 46 | Spanish | 13 weeks | n/a | Mixed | V, LG, S | 22 | Intermediate | IA, ST, ACTFL-OPI | 13 weeks | No |
| Segalowitz & Freed | 2004 | 40 | French | 13 weeks | n/a | Mixed | S | 22 | Intermediate | ACTFL-OPI | n/a | No |
| Freed et al. | 2004 | 28 | French | 12 weeks | Mixed | F | S | 19.88 | Mixed | HOPI | 3 weeks | No |
| (2 samples) | | | French | | Mixed | F | S | 19.88 | Intermediate | HOPI | 12 weeks | No |
| Dewey | 2004 | 30 | Japanese | 11 weeks | Mixed | F | R, V | 21.35 | Intermediate | IA | 9 weeks | No |
| Collentine | 2004 | 46 | Spanish | 16 weeks | H | F | S | 20 | Intermediate | ACTFL-OPI | n/a | n/a |
| Yashima & Zenuk-Nishide | 2008 | 177 | English | 40 weeks | n/a | Mixed | LC, G, R | 15 | n/a | ST | n/a | No |
| (2 samples) | | | English | 40 weeks | n/a | Mixed | LC, G, R | 15 | n/a | ST | n/a | No |
| Steven | 2001 | 22 | Spanish | 7 weeks | n/a | F | S | 22 | Intermediate | IA | 16weeks | No |
| (2 samples) | | | Spanish | 16 weeks | n/a | Mixed | S | 10.4 | Advanced | IA | 16 weeks | No |
| Freed | 1995 | 29 | French | 16 weeks | Mixed | Mixed | S | 20 | Intermediate | ILR-OPI | n/a | No |
| Huebner | 1995 | 22 | Japanese | 9 weeks | H | F | V, W, L, R | 22.8 | Basic | ST, IA | 9 weeks | No |

*Notes*. H = Host family. onD = school-based dormitory. offD = non-school-based dormitory. F = Formal language instruction. C = Content-based instruction. V = Receptive vocabulary knowledge. G = Grammar. LG = Lexical grammatical knowledge. P = Pragmatic knowledge. L = Listening. R = Reading. S = Speaking. W = Writing. WM = Working memory. T = Translation. IA = In-house assessment. ST = Standardized test. ACTFL-OPI = Oral Proficiency Interview following the scale of *American Council on the Teaching of Foreign Languages*. ILR-OPI = Oral Proficiency Interview following Interagency Language Roundtable scale. HOPI = Home spun Oral Proficiency Interview. n/a = Not available.

**Table 3.** Overall average effect size and heterogeneity test results of the three level random effects model.

| Model | Weighted effect size | | | 95% CI | | Heterogeneity | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *K* | *g* | SE | Lower | Upper | *Q*-value | *df* | *p*-value | $Tau^2_{within\text{-}study}$ | *p*-value | $Tau^2_{between\text{-}study}$ | *p*-value |
| Three-level random-effects | 283 | 0.87 | 0.17 | 0.53 | 1.20 | 1991.71 | 282 | < .001 | 0.57 | < .001 | 1.00 | < .001 |

*Notes.* $k$ = the number of effect sizes. $g$ = Hedges' standardized differences in means. SE = standard error.

primary studies. The details of these differences are discussed further in Section VI.7. In study abroad programs, learners vary in terms of the level of their target language proficiency which is believed to moderate the effect of studying abroad. Dekeyser (2010) advocated the need for learners' fundamental knowledge before their sojourn in order to reach noticeable L2 linguistic progress during their time abroad. Hence, the current meta-analysis sought to find the ideal starting proficiency level for learners to join study abroad programs.

## *6 Test mechanisms*

According to Keck et al.'s (2006) study, the three types of test mechanisms were coded as in-house assessment, an oral proficiency interview (OPI), and standardized test, respectively. In-house assessment refers to tests designed by the researcher or a placement test. As for standardized tests, the learners' proficiency level was assessed using established tests like the TOEFL or SAT Subject Test. An OPI could be further categorized into three variants: ACTFL-OPI (O'Brien et al., 2007; Segalowitz & Freed, 2004), ILR-OPI (Freed, 1995), and home-spun OPI (Freed et al., 2004; Vande Berg et al., 2009).

## *7 Type of residence*

Type of residence refers to the living conditions of learners in the target speech country. During a sojourn abroad, learners typically stay with a host family or live with other learners of the target language in university dormitories, but very few of them live alone in an apartment. Living with a host family offers learners more opportunities to practice the target language with native speakers. Unfortunately, using an unfamiliar language to communicate with foreigners may increase learners' language anxiety. In the dormitory, learners have chances to use both their native and the target language with peers. Hence, due to the different living surroundings while abroad, it can be argued that a study abroad experience may not have the same impact in these two different scenarios.

Various types of living conditions of learners in the target countries were coded. The type of residence was coded as 'host family' if learners lived with host family members, as 'dormitory' if learners lived in the school dormitory, and as 'mixed' if learners were

assigned to live with both a host family and in a dormitory in the primary studies. As listed in Table 1, dormitory could be further classified into an on/off-campus dormitory, and several combinations of mixed residences were listed in the primary studies on studying abroad.

### *8 Program content*

Two types of program content - formal language courses and content-based instruction - were identified. Formal language courses focus on teaching certain linguistic structures, practicing real-world language usage, or using the target language to complete meaningful tasks. On the other hand, content-based instruction is designed to teach participants using the target language without focusing on linguistic development (Ellis, 2003). Hence, study abroad language learning programs were coded as a 'formal language course' if their main focus was the acquisition of linguistic features and meaningful practice of language skills. Programs that involved content-based language learning were coded as a 'content-based course'. Such instruction included the learning of history, culture, and other subjects that were not directly related to the acquisition of certain linguistic features or skills. Since classroom instruction has been proven to have an effect out-of-class performance (McMeekin, 2004), it is of interest to determine which type of instruction can facilitate learners' L2 linguistic development more effectively during the time of sojourn.

### *9 Learners' age*

Learners' age was modeled as a continuous variable to avoid loss of information by categorization. Participants' estimated age was recorded if the primary studies reported learners' enrollment at school. For instance, 12 years old for sixth graders and 18 for freshmen in university. The average age was recorded if primary studies reported a range of ages such as '18 to 22 years old'. For this continuous variable, a meta-regression analysis would be conducted.

### *10 Length of treatment*

In the study abroad research, length of treatment refers to length of time spent in a target speech community, and ranges from 3 weeks (Serrano, Llanes, & Tragant, 2016) to a year and a half (Jiménez-Jiménez, 2010). Yang's (2016) study suggested that learners can best facilitate their L2 linguistic development in a short-term study abroad (i.e. less than 13 weeks), but Varela (2017) found that learners in a long-term study abroad period (i.e. more than 1 semester) outperformed students in both short- and mid-term programs. Apparently, existing studies adopted different categorical cut-off time points to operationalize the definitions of short-, mid-, and long-term durations. Such categorical operationalization makes it difficult for scholars and practitioners to interpret the effects of length of treatment across studies. Specifically, the average number of weeks was employed as the scale unit to operationalize the length of study abroad programs in the current meta-analysis. Weeks used as the scale unit across all studies helped depict a full

scope of treatment length distribution without loss of informative differences which may likely arise in the case of segmenting the length of treatment into several unfounded categories. Importantly, this operationalization allows researchers and practitioners to translate the effects of length treatment into practice.

### *11 Length of domestic programs*

Length of domestic programs was also operationalized as a continuous variable in order to objectively depict the moderating role of domestic programs. Average hours per week were used as the metric unit of length for domestic programs. Notably, this enabled a direct comparison of effect size and duration length between both study abroad and domestic programs.

### *12 Pre-program training*

Pre-program training was conceptualized as the provision of formal instruction *ad hoc* for the study abroad programs (Serrano et al., 2014; Llanes & Serrano, 2017). Despite the consensus among researchers that pre-program preparation offers benefits to study abroad learners, this facet has not been the focus of empirical investigations. Therefore, the current study aims to determine the potential influence that pre-program training may have on the effectiveness of study abroad programs. For this moderator analysis, yes/no classification was used for counting frequency of pre-program preparation in the primary studies.

### *13 Target language*

Despite the phenomena of English as the lingua franca, study abroad programs were also established for other modern languages including French (Freed et al., 2004; Segalowitz & Freed, 2004), Spanish (Félix-Brasdefer & Hasler-Barker, 2015; Isabelli-García, 2010), Japanese (Dewey, 2008; Huebner, 1995), Russian (Martinsen et al., 2011), Chinese (Liu, 2014), and German (Schenker, 2018). Notably, different combinations of more than one target language could also be featured in study abroad programs (Vande Berg et al., 2009; Winke & Gass, 2018). Since English has become the lingua franca, it is of interest to examine whether study abroad language programs are more beneficial for learning English or other modern languages.

## V Three-level meta-analysis

The current study took a multi-level modeling approach to conducting the meta-analysis (Cheung, 2015; Hox, Moerbeek, & van de Schoot, 2018). Typically, multi-level modeling can be used to explore and account for the modulating effects of context in which individuals are nested, and its application has been trending in language learning research (Khajavy, MacIntyre & Barabadi, 2018; Pfenninger & Singleton, 2016; Sasaki, Kozaki, & Ross, 2017). While meta-analysing existing study abroad primary studies, a three-level random effects model was implemented to account for the effect of the multiple

clustered, mutually-dependent effect sizes reported in the same primary study. It was noted that 283 effect sizes were available from the 42 primary studies collected for the present investigation. On average, each study contributed around 7 (283/42 = 6.7) effect sizes, suggesting the existence of mutual dependence among the effect sizes underlying most of the primary studies (Tullock & Ortega, 2017).

The phenomena of effect size dependence could be configured in at least four forms in the study abroad research: (1) the same linguistic construct operationalized by *multiple measures* (Köylü, 2016), (2) the performance of the same control group against *multiple control groups* (Dewey, 2008), (3) the performance of *multiple study abroad groups* against *multiple control groups* (Llanes & Serrano, 2017), and (4) the performance of *multiple study abroad groups* over *multiple time points* (Sasaki, 2004, 2009, 2011). For instance, Köylü (2016) employed *multiple measures* (i.e. accuracy, lexical complexity and syntactic complexity) to operationalize the same constructs: oral production and written production, respectively. Dewey (2008) compared the effects of studying abroad against intensive domestic immersion and academic-year formal classroom learning. Sasaki (2004, 2009, 2011) not only tracked Japanese learners' writing performance in study abroad programs through *multiple time points*, but also compared *multiple treatment groups* (i.e. different lengths of study abroad programs) against the same control group. Llanes and Serrano (2017) applied multiple measures to assessing both oral and written performance across three different age groups of studies abroad (children, adolescent, and adults) and compared those study abroad effects against their corresponding at home groups.

It is clear that the primary studies mentioned above are featured by different forms of effect size dependence. Hence, if all the 283 effect sizes were treated as mutually-independent, and no dependency was assumed among the clustered effect sizes, the result of analysis would become inflated and over-powered (Assink & Wibbelink, 2016) which would lead to a biased estimate of the overall average effect size. However, the clustered effect sizes in Hirai's (2018) study, Varela's study (2017), and Xu's (2019) study were all operationalized as mutually-independent. Although early approaches such as one single effect size per study and averaging the dependent effect sizes could be taken, they are much less realistic than the three-level meta-analytic approach since 'informative differences between effect sizes are lost and can no longer be identified in the analyses' (Assink & Wibbelink, 2016, p. 155). Unfortunately, clustered effect sizes in Yang's study (2017) were averaged out for each primary study. Hence, to preserve all the effect sizes across the primary studies as well as to model the potential heterogeneity of clustered effect sizes within the primary studies, the current research adopted a three-level meta-analytic framework to model the dependency effects existing in the primary studies. To elaborate, level 1 modeled the sampling variance of all the 283 extracted effect sizes; level 2 further modeled the variance in relation to the clustered effect sizes belonging to the same study; finally, level 3 modeled the variance of effect sizes across the 42 primary studies. By taking a three-level meta-analysis, the analyses of both the within-study level (level 2) and between-study level (level 3) of heterogeneity can be synchronized in the same meta-analytic framework such that the effects of meaningful and influential moderators can be likewise modeled at the within-study and between-study levels. This suggests that the modulating effects of salient moderators, as well as the variance explained by the moderators, can be modeled at both the within- and between-study levels.

The conceptual formulas of three-level meta-analysis were essentially based on Cheung's (2015) study and presented as follows:

Level 1: $y_{ij} = \lambda_{ij} + e_{ij}$

Level 2: $\lambda_{ij} = f_j + u_{(2)ij}$

Level 3: $f_j = \beta_0 + u_{(3)ij}$

At level 1, $y_{ij}$ refers to the observed effect size ($i_{th}$) nested within the primary study ($j_{th}$); $\lambda_{ij}$ denotes the true effect size for the observed corresponding effect size(s) in the same study; $e_{ij}$ represents the known sample variance associated with the observed effect size ($i_{th}$) in the clustered study ($j_{th}$). At level 2, $\lambda_{ij}$ becomes the dependent variable, upon which the clustered study effect size($j_{th}$) predicts(($f_j$)), whereas $u_{(2)ij}$ refers to the random effects observed at level 2. At level 3, $f_j$ points to the outcome variable, upon which average population effect ($\beta_0$) predicts, while $u_{(3)ij}$ refers to the random effect observed at level 3. Since the random effects at levels 2 and 3 are modeled in the current three-level meta-analysis, variance of $u_{(2)ij}$ and variance of $u_{(3)ij}$, symbolized as $\tau^2_{(2)}$ and $\tau^2_{(3)}$ respectively, also need to be estimated and modeled in the analysis.

Importantly, given the estimation of $\tau^2_{(2)}$ and $\tau^2_{(3)}$, it is possible to further obtain the variance explained by level 2 and level 3. The formulae are shown below:

$$R^2_{(2)} = 1 - \frac{\hat{\tau}^2_{(2)1}}{\hat{\tau}^2_{(2)0}} \quad (1)$$

$$R^2_{(3)} = 1 - \frac{\hat{\tau}^2_{(3)1}}{\hat{\tau}^2_{(3)0}} \quad (2)$$

$R^2_{(2)}$ and $R^2_{(3)}$ refer to variance explained by level 2 and level 3, respectively. Notably, $\hat{\tau}^2_{(2)0}$ and $\hat{\tau}^2_{(2)1}$ are the estimated variance with and without moderators at level 2. By the same token, $\hat{\tau}^2_{(3)0}$ and $\hat{\tau}^2_{(3)1}$ are $ri?^2_{(3)1}$ the estimated variance with and without moderators at level 3. Therefore, $R^2_{(2)}$ and $R^2_{(3)}$ can be interpreted as the percentage of the variance of heterogeneity that can be accounted for by the moderators at level 2 and level 3, respectively. This modeling feature has the advantage of identifying modulating weights of moderators at these two distinct levels.

The current three-level random effects meta-analysis was enacted via the *R* package, *metaSEM* (Cheung, 2015). Hedges' *g* was used as the effect size metric in the investigation.

*Q*-test was performed to determine whether any heterogeneity exists among the effect sizes estimated by the primary studies. Specifically, the heterogeneity variances featured in the multiple clustered, mutually dependent effects within primary studies ($tau^2_{within\text{-}study}$) as well as the heterogeneity variances across independent primary studies ($tau^2_{between\text{-}study}$) would be reported. Meta-regression analyses were further performed to determine whether

continuous moderator variables could predict the magnitude of effect size. To establish coding quality and reliability, two raters specializing in second language acquisition coded the 42 studies independently. Cohen's kappa coefficient was calculated and found to be .95 for the entire coding process, demonstrating a fairly high degree of inter-rater agreement.

## VI Results

### *1 The research synthesis*

The current work analysed a total of 42 primary studies (i.e. 24 published articles, 3 PhD theses, and 15 empirical studies from book chapters) that were published between 1995–2019, and up to 4,068 L2 learners were included. Furthermore, 283 effect sizes were extracted from the primary studies. Table 3 summarizes the coded constructs of methodological features, learner characteristics, and programmatic factors which were operationalized as different moderator variables. Participants were on average 19.65 years old. The native and target language of the participants of the primary studies were varied.

### *2 Publication bias*

Publication bias occurs in situations where the sample in the primary studies does not demonstrate the complete population of interest. This can lead to a greater likelihood of publishing a favorable research outcome. If publication bias exists, an inflated overall effect size may be present (Plonsky & Oswald, 2014), and the research outcome may become questionable. Aside from two unpublished PhD theses, all the primary studies collected are published articles and book chapters. In order to assess the likelihood of publication bias in the current work, the fail-safe N index (Rothstein, Sutton, & Borenstein, 2005) and trim-and-fill analysis (Duval & Tweedie, 2000) were employed to check the likelihood of publication bias.

The fail-safe N index refers to the number of unpublished studies required in order for the overall average effect computed in a meta-analysis to become zero. For the current meta-analysis, the safe number would be $N_{fs} = 5k + 10 = 5 \times 283 + 10 = 1425$. The results showed that 11,708 more effect sizes would have been required to invalidate the significant result of the study, and this number exceeded the safe number by more than eight times. Furthermore, the researchers also conducted Duval and Tweedie's (2000) trim-and-fill analysis which trimmed the asymmetric and extreme studies to adjust the estimate of the overall effect size. As shown in Figure 1, only seven missing studies marked in white circles needed to be imputed.

To check whether the overall effect size of study abroad language programs would significantly change with the inclusion of the seven imputed studies, we tested the null hypothesis of 'no missing studies on the left side' under the framework of the random effects model. We compared the original overall effect size ($g = 0.87$, see below for full information) against the 'trimmed' overall effect size ($g = 0.81$), and the results showed that the difference ($g_{\Delta} = 0.06$) between the two effect sizes did not reach statistical significance ($p > .05$). Given the small amount of change in effect size, the results of the current meta-analysis could be taken as trustworthy and reliable.

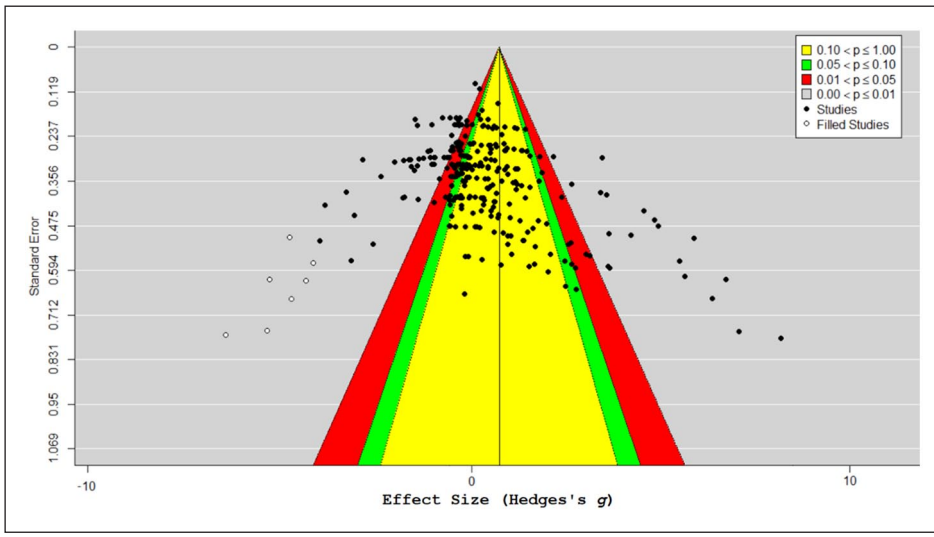


**Figure 1.** The funnel plot of trim-and-fill analysis.

## 3 The effects of study abroad language programs

Table 3 presents information on the overall effectiveness of study abroad language programs including standard error, the number of mean effect sizes, 95% CIs, *Q*-test values, and tau$^2$ values both at the within- and between-study levels. Since a high heterogeneity ($Q = 1991.71$, $df = 282$, $p = .000$) was presented, the random-effects model could be assumed. As seen in Table 3, the three-level random-effects model presented a robust and large overall average effect size ($g = 0.87$). The 95% CIs were far above zero (0.53, 1.20), indicating a reliable effect of study abroad language programs on L2 development. The results of *Q*-test values indicated a substantial variability between the primary studies and suggesting the need for moderator analysis. Importantly, the values of both $tau^2_{within\text{-}study}$ and $tau^2_{between\text{-}study}$ were 0.57 and 1.01, respectively, and statistically significant which suggests significant heterogeneity variances existing at level 2 (within-study level) and level 3 (between-study level).

## 4 Moderator analyses

The analysed moderator variables include outcome measure of linguistic gains, learners' language proficiency, measure of proficiency, type of residence, length of treatment, program content, target language, length of domestic programs, learners' age, and pre-program training. The effect sizes of different levels of the moderators as well as the $R^2$ values of the two distinct levels were organized in Table 4 and Table 5, respectively.

## 5 Type of outcome measure

Meaningful differences were uncovered among the 11 types of outcome measures. Significant effects were observed on *pragmatic knowledge*, *listening*, *speaking*, *writing*,

**Table 4.** Effect sizes and $R^2$ results of the moderators in the primary studies.

| | *k* | Effect size | | | 95% CI | | Explained variances ($R^2$) | |
|---|---|---|---|---|---|---|---|---|
| | | *g* | SE | *p* | Lower | Upper | Level $2_{within\text{-}study}$ | Level $3_{between\text{-}study}$ |
| All studies | 283 | | | | | | | |
| *Type of outcome measure*: | | | | | | | 0.02 | 0.07 |
| Grammar | 13 | 0.58 | 0.44 | | −0.27 | 1.43 | | |
| Lexical-grammatical knowledge | 1 | 0.90 | 0.89 | | −0.84 | 2.63 | | |
| Listening+grammar+reading | 2 | 1.43 | 1.12 | | −0.77 | 3.63 | | |
| Listening+grammar+reading+translation+writing | 3 | 0.04 | 0.78 | | −1.49 | 1.56 | | |
| Pragmatic knowledge | 12 | 0.62 | 0.50 | < .05 | 0.35 | 1.60 | | |
| Listening | 6 | 0.74 | 0.37 | < .05 | 0.07 | 1.54 | | |
| Reading | 6 | 0.13 | 0.53 | | −1.17 | 0.90 | | |
| Speaking | 146 | 1.02 | 0.20 | < .001 | 0.64 | 1.40 | | |
| Writing | 65 | 0.87 | 0.22 | < .001 | 0.44 | 1.30 | | |
| Vocabulary knowledge$_{receptive}$ | 23 | 0.98 | 0.40 | < .05 | 0.20 | 1.76 | | |
| Working memory | 6 | 0.40 | 0.77 | | −1.10 | 1.91 | | |
| *Language proficiency*: | | | | | | | 0.01 | 0.30 |
| Beginner | 42 | 0.95 | 0.23 | < .001 | 0.50 | 1.41 | | |
| Intermediate | 172 | 0.66 | 0.17 | < .001 | 0.33 | 0.99 | | |
| Advanced | 59 | 0.90 | 0.21 | < .001 | 0.48 | 1.31 | | |
| Mixed | 10 | 3.04 | 0.63 | < .001 | 1.81 | 4.27 | | |
| *Measure of proficiency*: | | | | | | | 0.01 | 0.21 |
| IA | 198 | 0.67 | 0.18 | < .001 | 0.33 | 1.02 | | |
| ST | 17 | 0.69 | 0.30 | < .05 | 0.07 | 1.32 | | |
| ACTFL-OPI | 43 | 1.84 | 0.34 | < .001 | 1.17 | 2.51 | | |
| ILR-OPI | 1 | 0.63 | 1.23 | | −1.78 | 3.03 | | |
| HOPI | 24 | 0.89 | 0.66 | | −0.41 | 2.18 | | |
| *Type of residence*: | | | | | | | 0.14 | 0.02 |
| Host family | 84 | 0.75 | 0.17 | < .001 | 0.41 | 1.09 | | |
| On-campus dormitory | 17 | 0.50 | 0.34 | | −0.17 | 1.17 | | |
| Off-campus dormitory | 5 | 0.60 | 0.47 | | −0.31 | 1.52 | | |
| On-campus dormitory+host family | 6 | 0.55 | 0.39 | | −0.21 | 1.31 | | |
| Off-campus dormitory+host family | 6 | 0.42 | 0.57 | | −0.69 | 1.53 | | |
| On-campus dormitory+off-campus dormitory | 11 | 0.51 | 0.30 | | −0.08 | 1.11 | | |
| On-campus dormitory+off-campus dormitory+host family | 52 | 0.63 | 0.21 | < .001 | 0.23 | 1.03 | | |
| *Program content*: | | | | | | | 0.01 | 0.00 |
| Content-based course | 24 | 0.62 | 0.45 | | −0.27 | 1.51 | | |
| Formal language course | 126 | 0.81 | 0.22 | < .001 | 0.39 | 1.24 | | |
| Mixed | 131 | 0.99 | 0.21 | < .001 | 0.59 | 1.40 | | |
| *Target language*: | | | | | | | 0.00 | 0.16 |
| English | 144 | 0.59 | 0.23 | < .05 | 0.14 | 1.04 | | |
| Spanish | 76 | 1.08 | 0.28 | < .001 | 0.53 | 1.63 | | |

*(Continued)*

**Table 4.** (Continued)

| | *k* | Effect size | | | 95% CI | | Explained variances ($R^2$) | |
|---|---|---|---|---|---|---|---|---|
| | | *g* | SE | *p* | Lower | Upper | Level $2_{within\text{-}study}$ | Level $3_{between\text{-}study}$ |
| Japanese | 15 | 1.44 | 0.59 | < .05 | 0.29 | 2.59 | | |
| French | 28 | 1.58 | 0.46 | < .001 | 0.68 | 2.49 | | |
| Chinese | 6 | 0.09 | 0.98 | | −1.84 | 2.02 | | |
| German | 6 | 0.75 | 0.71 | | −0.64 | 2.15 | | |
| Russian | 1 | 0.62 | 1.13 | | −1.60 | 2.85 | | |
| Mixed | 7 | 0.91 | 0.55 | | −0.16 | 1.98 | | |
| *Pre-program training*: | | | | | | | | |
| Yes | 24 | 0.53 | 0.61 | | −0.67 | 1.72 | 0.001 | 0.008 |
| No | 245 | 0.86 | 0.18 | < .001 | 0.51 | 1.22 | | |

*Notes.* $k$ = the number of effect sizes. $g$ = Hedges' standardized differences in means. SE = standard error.

**Table 5.** Three-level multiple meta-regression analysis on length of treatment, length of domestic program, and age with language proficiency as covariate.

| | | Effect size | | 95% CI | | | Explained variances ($R^2$) | |
|---|---|---|---|---|---|---|---|---|
| | *k* | β | SE | Lower | Upper | *p*-value | Level $2_{within\text{-}study}$ | Level $3_{between\text{-}study}$ |
| Length of treatment (week) | 194 | 0.20 | 0.27 | −0.32 | 0.72 | $p > .05$ | 0.10 | 0.41 |
| Length of domestic program (hours/week) | | −0.28 | 0.09 | −0.46 | −0.11 | $p < .001$ | | |
| Age | | −0.01 | 0.03 | −0.06 | 0.05 | $p = .75$ | | |
| Beginner | | 0.92 | 0.26 | 0.40 | 1.43 | $p < .001$ | | |
| Intermediate | | 0.67 | 0.17 | 0.34 | 1.00 | $p < .001$ | | |
| Advanced | | 0.93 | 0.24 | 0.47 | 1.40 | $p < .001$ | | |
| Mixed | | 3.06 | 0.63 | 1.82 | 4.30 | $p < .001$ | | |

*Notes.* $k$ = the number of effect sizes. β = unstandardized beta coefficient. $g$ = Hedges' standardized differences in means. SE = standard error. * Language proficiency was modeled as a covariate rather than a moderator.

and *receptive vocabulary knowledge measures*. The largest effect was found on *speaking* ($g = 1.02$ [0.64, 1.40], $k$ [number of effect sizes] = 146), with the CI excluding zero, $p < .001$; followed by *receptive vocabulary knowledge* ($g = 0.98$ [0.20, 1.76], $k = 23$), with the CI locating above zero, $p < .05$. By way of comparison, the $g$ values of pragmatic knowledge ($g = 0.62$ [0.05, 1.60], $k = 12$), *listening* ($g = 0.74$ [0.07, 1.54], $k = 6$) and *writing* ($g = 0.87$ [0.44, 1.30], $k = 65$) were comparatively smaller, and their CIs covered more areas than those of *speaking* and *receptive vocabulary knowledge*.

Notably, the insignificant types of outcome measures included *grammar*, *lexical-grammatical knowledge*, *two general proficiency measures* (i.e. *listening + grammar + reading; listening + grammar + reading + translation + writing*), *reading*, and *working memory*. Overall, the explained variances ($R^2$) contributed by 'type of outcome measure' at Level $2_{within\text{-}study}$ were small ( = 0.02), whereas at Level $3_{between\text{-}study}$ 7% of variances were observed, indicative of a certain level of practical significance.

## 6 Learners' language proficiency

The results of this moderator analysis showed, first, that the beginner level could benefit significantly from study abroad programs ($g = 0.95$ [0.50, 1.41], $k = 42$), with the CI reliably locating above zero, $p < .001$. Next, the advanced level ($g = 0.90$ [0.48, 1.31], $k = 59$), with zero also firmly falling outside CI, $p < .001$. Finally, the intermediate level ($g = 0.66$ [0.33, 0.99], $k = 172$), with CI excluding zero, $p < .001$. The grouping of mixed language proficiency achieved the largest effect size in this moderator analysis ($g = 3.04$, [1.81, 4.27], $k = 10$), $p < .001$. However, there was no way of knowing the exact combinations of language proficiency levels regarding the largest effect size. By way of comparison, the ranges of CIs across the beginner, intermediate, and advanced levels intersected to a great extent and overlapped substantially which indicates similar, high facilitative effects of study abroad programs on the three different proficiency levels of learners. Overall, the explained variances ($R^2$) contributed by *language proficiency* at Level $2_{within\text{-}study}$ were very small ( = 0.01), whereas at Level $3_{between\text{-}study}$ 30% of variances were detected, indicative of a noticeable level of practical significance.

## 7 Measure of proficiency

A close inspection of CIs of the five levels revealed that the largest effect size was found for *ACTFL OPI* ($g = 1.84$, [1.17, 2.51], $k = 43$), with zero falling far outside the CI, $p < .001$, followed by *standardized tests* ($g = 0.69$ [0.07, 1.32], $k = 17$), with the CI also excluding zero, $p < .05$, and then by *in-house assessments* ($g = 0.67$ [0.33, 1.02], $k = 198$), with the CI locating above zero, $p < .001$. Although the other two OPI-related proficiency measures (i.e. *ILR-OPI* and *home-spun OPI*) did not reach statistical significance, it was noted that most of the areas of the CI of home-spun OPI ($g = 0.89$ [–0.41, 1.32], $k = 24$) were above zero, suggesting some consistent effect of the measure.

Essentially, the *g* value of oral proficiency interview was not located in the CIs of both standardized tests and in-house assessments, and vice versa, suggesting a superior effect of the oral proficiency interview. Notably, the *g* values of both standardized tests and in-house assessments were mutually inclusive in their CIs, and the ranges of their CIs were also largely overlapped, suggesting little noticeable, reliable difference in effects between the two proficiency measures. The explained variances ($R^2$) endorsed by *measure of proficiency* Level $2_{within\text{-}study}$ and Level $3_{between\text{-}study}$ were 0.01 and 0.21, respectively, indicating that *measure of proficiency* in actuality explained the heterogeneity variances that were purely present at Level $3_{between\text{-}study}$ rather than at Level $2_{within\text{-}study}$. Hence, conceivably and quite possibly, the effectiveness of study abroad programs may have been moderated by different proficiency measures that researchers adopted or designed.

## 8 Type of residence

Substantial differences were found among the seven types of effect sizes regarding type of residence. Notably, significant effects were detected only on *host family* and *Mixed_04* (*on-campus* + *off-campus* + *host-family dormitory*). By contrast, it was important to note that the effect size of both *on-campus dormitory* ($g = 0.50$ [–0.17, 1.17], $k = 17$, $p > .05$) and *off-campus dormitory* ($g = 0.60$ [–0.31, 1.52], $k = 5$, $p > .05$) failed to reach statistical significance. Hence, there seemed no reliable difference in effect between *on-campus dormitory* and *off-campus dormitory.* This finding clearly suggests that type of residence did moderate the effectiveness of study abroad programs. Importantly, type of residence could explain more heterogeneity variances arising from Level $2_{within\text{-}study}$ ( = 0.14) than from Level $3_{between\text{-}study}$ ( = 0.02), indicative of noticeable variability existing *within* rather than *between* primary studies.

## 9 Program content

It was found that the *g* values of both formal language courses ($g = 0.81$, [0.39, 1.24], $k = 126$) and mixed courses ($g = 0.99$ [0.59, 1.40], $k = 131$) were large and statistically significant, with zero reliably falling outside their CIs, suggesting convincing evidence for trust-worthy instructional effects of the two program types. By contrast, content-based instruction obtained the smallest effect size of this moderator analysis ($g = 0.62$ [–0.27, 1.51], $k = 24$), with zero being included in the CI, suggesting a meaningful and conclusive difference regarding its instructional effects as compared to the two counterparts. However, the explained variances ($R^2$) contributed by *program content* at Level $2_{within\text{-}study}$ ( = 0.01) and Level $3_{between\text{-}study}$ ( = 0.00) were both extremely small, indicative of little practical significance.

## 10 Target language

Critical differences were uncovered among the eight types of effect sizes regarding target language. Salient effects were observed on *English*, *Spanish*, *Japanese*, and *French*. Among these four significant target languages, the largest effect was found on *French* ($g = 1.58$ [0.68, 2.49], $k = 28$), with the CI locating above zero, $p < .001$; followed by *Japanese* ($g = 1.44$ [0.29, 2.59], $k = 15$), with the CI excluding zero, $p < .05$; followed by *Spanish* ($g = 1.08$ [0.53, 1.63], $k = 76$). Notably, the effect of study abroad programs on English learning could reach statistical significance ($g = 0.59$ [0.14, 1.04], $k = 144$), but the effect was comparatively much smaller than those of French, Japanese, and Spanish.

On the contrary, the insignificant target languages included Chinese, German, and Russian. Overall, the explained variances ($R^2$) contributed by *target language* at Level $2_{within\text{-}study}$ were zero ( = 0.00), whereas at Level $3_{between\text{-}study}$ 16% of variances were observed, indicative of a noticeable level of practical significance.

## 11 Pre-program training

The results showed that the effect size of the study abroad programs *with* pre-program training was not statistically significant ($g = 0.53$ [–0.67, 1.72], $k = 24$), whereas the

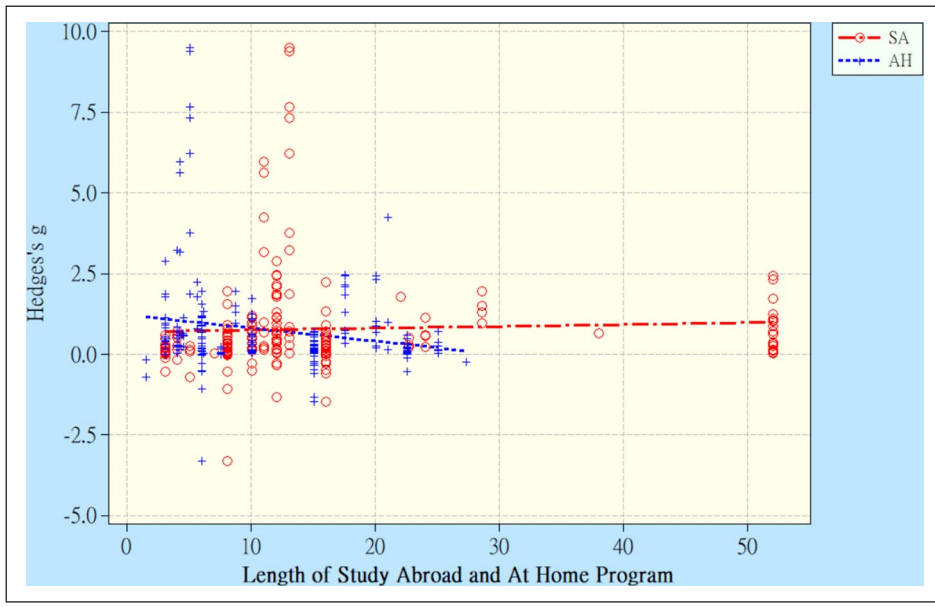


**Figure 2.** Length of study abroad and at home program.

effect size of the study abroad programs *without* pre-program training reached statistical significance ($g = 0.86$ [0.51, 1.22], $k = 245$). This finding suggests that pre-program training might not be able to add extra benefits to the study abroad programs.

Researchers were concerned about potential multicollinearity among the moderators of age, length of study abroad programs, length of domestic programs, and language proficiency. Therefore, a comprehensive three-level meta-regression was run on these four focal moderators simultaneously, and this still enabled a large set of 194 effect sizes to be analysed in the same run. As reported in Table 5, it was found that the length of study abroad programs (unstandardized $\beta = 0.20$, [–0.32, 0.72], $k = 194$, $p = 0.24$) might not have a reliable effect, whereas the length of domestic programs (unstandardized $\beta = –0.28$, [–0.46, –0.11], $k = 194$, $p < 0.01$) could produce a significant yet negative impact. However, the statistical insignificance notwithstanding, the range of the 95% CI of length of study abroad programs covered more areas above zero. As depicted in Figure 2, the directions of their regression effects were opposite. This suggests that the longer the duration of study abroad programs, the more pronounced the effect. Conversely, the longer the duration of domestic programs, the less promising the effect. The current study is the first of its kind to unveil this salient finding. Importantly, it is interesting to note that the two regression lines intersected at 12 on the x-axis, suggesting that the effects of study abroad programs less than 12 weeks are comparatively smaller than those of at home programs in the same primary studies. At least 13 weeks were required for the effects of studying abroad to become superior to those of at home programs. Furthermore, by treating proficiency level as covariate, the effect size of age (unstandardized $\beta = –0.06$, [–0.18, 0.14], $k = 194$, $p = 0.71$) did not reach statistical significance.

These results suggest that the included primary studies were heterogeneous and the effect of study abroad language programs was distinctively moderated by the subgroups operationalized in the study. It should also be noted that some of the learner characteristics, methodological features, and programmatic factors were not equally presented because this information was lacking in the primary studies. For instance, some primary studies did not mention the type of residence, while others did not present the length of pre-program training. Hence, the results were carefully examined with this in mind.

## VII Discussion

The current three-level meta-analysis examined the effectiveness of study abroad language programs on L2 acquisition as well as other salient moderators underlying the effectiveness of such programs. The present study not only included a greater number of primary studies with a wider age range of participants, but also adopted a more precise meta-analytic model to pinpoint the overall linguistic gains of study abroad programs. On the whole, the current meta-analysis shows that learners who study abroad have greater linguistic gains than those who stay in their home country for domestic study. The present study further analysed the moderating effect of 10 variables, and in general the featured moderators explained various degrees of heterogeneity variances at level $3_{between\text{-}study}$ than at level 2 Level $2_{within\text{-}study}$.

### *1 Research questions*

Research question 1: What are the overall effects of study abroad programs on linguistic gains with a more comprehensive coverage and inclusion of primary studies? The overall average effect size ($g = 0.87$) fell between the effect sizes evidenced in Varela's (2017) study ($d = 0.98$) and Yang's (2016) study ($d = 0.75$), and was far greater than the ones reported in Hirai's (2018) study ($g = 0.56$) and Xu's (2019) study ($d = 0.37$). The total number of L2 learners involved in the current meta-analysis amounted to 3,938, far exceeding the number of participants available in the four prior meta-analyses. It is therefore important to discuss the practical significance of the overall $g$ value of study abroad programs. Essentially, the current study follows Plonsky and Oswald's (2014) guidelines regarding the interpretation of magnitude of effect size in the context of L2 acquisition (i.e. 0.40 as small effect, 0.70 as medium effect, and 1.0 as large effect). Thus, the effect size level of the overall average $g$ value of 0.87 meta-analysed in the current study can be interpreted as medium-to-large and meaningful regarding its practical significance. This suggests that, by comparison, 82% of the participants in a control group (i.e. at home learners) would obtain lower scores in a linguistic measure than any average participant in an experimental group (i.e. study abroad learners) (Coe, 2002). This result clearly argues for the implementation of study abroad programs to facilitate and advance L2 learning. Importantly, these outcomes may be taken as a more reliable quantitative synthesis because the current study adopted a more realistic statistical model (Tullock & Ortega, 2017). The three-level random effects model adopted by the current work achieved a more unbiased effect estimate and precluded the confounding effect which might have been brought by the inclusion of a pretest–posttest experimental design without a control group.

Essentially, the current meta-analysis models the heterogeneity that is likely to arise at both the within- and between-study levels, reliably mirroring the genuine effect size structure of variability existing within and across the primary studies. The approach adopted by Yang (2016) took averages of the clustered effect sizes within primary studies, and this may have failed to identify the potential informative differences existing among the effect sizes of any primary study, thus underestimating the effectiveness of study abroad programs. On the other hand, the approach taken by Hirai's (2018), Varela's (2017) and Xu's (2019) studies intentionally treated the clustered effect sizes as mutually-independent, which could have inflated the effectiveness of study abroad programs. The above observations may explain why Varela obtained a large effect size in his study.

Unlike Hirai's (2018) and Xu's (2019) studies which included primary studies with both an experimental-vs.-control group design and a within-group design (i.e. study-abroad group only), the present meta-analysis only included studies with an experimental-vs.-control group design in order to make a reliable and meaningful comparison between study abroad language programs and domestic language programs. As far as the effects of study abroad are concerned, the focal issue should be on the comparison of program effectiveness between study abroad and domestic language programs. The within-group design may have its value in understanding the effects of study abroad programs; however, the lack of control group design (i.e. normally at home language programs) makes it unlikely to create a common baseline upon which the effects of study abroad programs may be compared against. Notably, a single group with pretest–posttest experimental design may suffer considerably from a systematic bias such as regression toward the mean (RTM) (Marsden & Torgerson, 2012). In other words, participants with higher pretest scores tend to make less improvement in test performance, whereas those with lower pretest scores consistently make more progress in test scores. Hence, because Hirai's (2018), Varela's (2017) and Xu's (2019) studies may have confounded the study abroad program effects with the participant effects, their results should be interpreted with cautions.

To reiterate, the current study hallmarks as the first empirical undertaking that enables the modeling of heterogeneity both within and between primary studies in a meta-analytic framework. As indicated in Table 3, the heterogeneity variances ($Tau^2$) of both the within-study level and between-study level reached statistical significance. This finding lends the direct support necessary for modeling the variability of clustered effects within different primary studies, and suggests that the featured moderators of the study may have distinct roles in explaining heterogeneity variances at the two study levels.

Research question 2: To what extent do the effects of studying abroad vary across different moderating factors identified in the study? In total, 10 moderators were featured for analysis in the present study. This number is far greater than those of previous works (Hirai, 2018; Varela, 2017; Xu, 2019; Yang, 2016). Clearly, an investigation of the modulating effects of the 10 moderators greatly helped identify and understand the degree to which the effects of study abroad language programs may vary as a function of salient factors. Importantly, the selection of the 10 moderators investigated in the current study was essentially motivated by theoretical relevance rather than chosen in a speculative manner. Hence, the variance in effect sizes accounted for by the 10 moderators can be deemed a more practical significance than a trivial nuance, either at the within-study level or at the between-study level.

### 2 Length of study abroad programs & length of at home programs

In the beginning, with regard to the length of treatment, the current study revealed a positive value ($\beta$ = 0.20, [–0.32, 0.72]) for its moderating effects, despite an inclusion of zero in the 95% CI. Because most of the range of the 95% CI of the moderator covered the areas above zero (i.e. [0.72/(0.32+0.72) = 69%]) the findings of the current study are generally consistent with those of Hirai's (2018) and Varela's (2017) meta-analyses, suggesting that a longer length of study abroad language program results in greater benefits. On the other hand, in Yang's (2016) research, a shorter length (i.e. less than 13 weeks) study abroad experience was shown to better facilitate target language learning. These inconsistent findings may be explained by the different sets of primary studies included in the three meta-analyses as previously discussed. Yang (2016) only included 11 primary studies, whereas there are 42 in the present study. It is argued that a more comprehensive inclusion of primary studies in the current meta-analysis offered more convincing findings. Importantly, according to the results of Figure 2, it appears that the effects of study abroad programs less than 12 weeks are comparatively smaller than those of at home programs in the same primary studies. In other words, the effects of at home programs less than 12 hours per week are comparatively greater than those of study abroad programs in the same primary studies. Notably, the outcome of Figure 2 further suggests that 13 weeks were the minimum required in order for the effects of a studying abroad to become superior to those of at home programs.

Since there is a descending growth trajectory in the interplay between the duration of at home programs and the magnitude of effect size. This directly supports the research findings of some primary studies (Dewey, 2008; Freed et al., 2004; Jiménez-Jiménez, 2010; Serrano et al., 2011) which indicated a larger effect size for intensive programs when compared with semi-intensive and regular domestic programs. The suggests that both researchers and practitioners need to consider the optimal duration for a domestic language program since an extended and prolonged program may exhaust learners' motivation and interest in language learning. Such exhaustion could result in undesirable learning outcomes like fatigue or boredom. Therefore, if learners cannot afford to study abroad and choose to study at home, then it seems that an intensive program may be an ideal option compared with longer, less intensive programs. Overall, because the outcome of Figure 2 was plotted on the basis of 194 effect sizes that juxtaposed length of the study abroad and at home in the same primary studies, the theoretical underpinnings reasoned from the above discussions were empirically founded and reliable.

### 3 Type of residence

In terms of type of residence, contrary to Varela's (2017) dichotomous category, the current study identified 7 types of residences from the literature and tested their effects. The results show that only two types of residence may have facilitative effects: *host family* and one mixed residence (*on-campus* + *off-campus* + *host-family dormitory*). Neither on-campus dormitory nor off-campus dormitory had a significant modulating effect on linguistic gains. However, it should be noted that most of the 95% CIs of the two types of dormitory cover areas above zero, suggesting some reliable effects from the

dorm-based residence. By way of comparison, the outcome of the study clearly suggests that there is a reliable difference in effect between host family and dorm-based residence. This finding clearly indicates that type of residence did moderate the effectiveness of study abroad programs.

To elaborate, in comparison to the learners living the dorm-based residence, learners staying with a host family may gain a greater amount of higher-quality input and output of the target language from native speakers and, in turn, achieve a higher level of proficiency. In the primary studies with learners living in the dormitory (O'Brien et al., 2007; Sasaki, 2004, 2007; Serrano, Tragant, & Llanes, 2014; Serrano et al., 2016), most of the learners lived with non-native learners, and only a few stayed with teachers of the target language and/or with local students. That is, students living in the dormitory either used the target language or their native language when speaking with other non-native speakers. On the other hand, learners who stayed with a host family used the target language almost exclusively with native speakers. Therefore, living in the dormitory may not be as beneficial to L2 development as staying with a host family since learners have fewer opportunities to practice the target language and tend to speak a common native language with other learners in the school dormitory.

In addition, according to the Contact Hypothesis, staying with a host family provides the social and cultural interactions conducive to language acquisition and acculturation process (Allen, Dristas, & Mills, 2006; Knight & Schmidt-Rinehart, 2002; Rivers, 1998). For instance, in Knight and Schmidt-Rinehart's (2002) study, Spanish and Mexican host mothers took on the role of conversation partners and teachers who assisted American students linguistically and culturally. Also, in Law's (2003) study, learners reported that they received ample input, practice, and feedback by spending more time conversing with their host mothers.

Since time spent with a host family is beneficial to language acquisition, program duration could also have a positive impact on forming a relationship with a host family as described in previous studies (Schmidt-Rinehart & Knight, 2004; Schmidt-Rinehart & Knight, 2002). These researchers illustrated that learners who are part of a longer program have fewer problems with their host family, which could be attributed to an increased amount of time available to work through cultural differences and language barriers. Indeed, as shown in Figure 2, there is an ascending growth trajectory in the interplay between the duration of study abroad programs and the effect of study overseas. This research finding lends support to the superior effect of living with a host family to that of living in a dormitory, thereby confirming the potential effects that homestay placements have on L2 linguistic development.

### *4 Type of outcome measure*

In terms of the modulating effects of different types of outcome measures on linguistic gains, the current study in principle adopts Borràs and Llanes's (2019) classification structure to examine how the effects of studying abroad may be moderated by different components of language gains. Overall, this study identified 7 types of outcome measures from Borràs and Llanes's category structure and 3 additional types of outcome

measures which were overlooked in Borràs and Llanes's classification scheme yet actually used in some other primary studies.

The outcome of the current study showed that study abroad language programs could have significant effects on developing L2 learners' speaking, writing, listening, receptive vocabulary knowledge, and pragmatic knowledge; whereas, study abroad programs were not found to have a significant effect on acquisition or development of grammar, lexical-grammatical knowledge, general language proficiency, or working memory. Among the significant linguistic measures speaking, writing, and receptive vocabulary knowledge can benefit greatly from study abroad programs, as evidenced by the large effects they received. Listening and pragmatic knowledge received somewhat smaller yet still appreciable effect sizes. Notably, despite the insignificant effect sizes of grammar and lexical-grammatical knowledge, most of the 95% CIs associated with these factors cover above zero, suggesting some reliable effects of study abroad programs.

An examination of the length of study abroad programs in the primary studies including grammar and lexical-grammatical knowledge as linguistic measures (Hirakawa, Shibuya, & Endo, 2019; Isabelli-García, 2010; Schenker, 2018; Segalowitz et al., 2004) revealed that positive and large effect sizes were obtained from the primary studies where length of the study abroad programs was typically around one-semester (i.e. four months) (Isabelli-García, 2010; Segalowitz et al., 2004). On the other hand, negative or very small effect sizes were observed from the primary studies where the length of the study abroad programs were just 3–5 weeks (Hirakawa et al., 2019; Schenker, 2018). This micro analysis not only points to the salient effects of study abroad programs on developing L2 grammar, but also explicitly suggests the conditioned time frame necessary for study abroad programs to really take effect on promoting L2 grammar.

In a similar vein, reading ability was also less likely to benefit from study programs and its 95% CI covers more areas below zero. A further check of the primary studies focusing on reading (Dewey, 2004; Huebner, 1995; Li, 2004; Schenker, 2018) indicated that the length of these programs varied between 4 and 11 weeks, and the target languages were mainly Chinese, German, and Japanese. In the case of reading, the length of study abroad programs and types of target languages may jointly determine the effects of study abroad programs on reading. In sum, the results of the modulating effects of different types of outcome measures on linguistic gains not only systematically establish the empirical grounding to timely inform the narrative review provided by Borràs and Llanes (2019), but also critically update and advance the understanding of the way in which the effects of studies abroad may be modulated by different types of outcome measures.

## *5 Test mechanisms*

With regard to test mechanisms, the results revealed that in-house assessment occupied most of the effect sizes of this moderator analysis ($k = 198$), followed by standardized tests ($k = 17$), ACTFL-OPI ($k = 52$), home-spun OPI ($k = 24$), and ILR-OPI ($k = 1$). Additionally, the results found that ACTFL-OPI obtained the largest effect size and was substantially larger than that of either in-house assessments or standardized tests. This finding was further evidenced by the significant amount of heterogeneity variances explained at level 3 (i.e. the between-study level). Because of this, the effectiveness of

study abroad programs on L2 acquisition may vary as a function of different test mechanisms that primary studies employed. OPI essentially measures how well a person speaks a language by assessing his/her performance on a range of language tasks against a list of specified criteria. For instance, the performance on ACTFL-OPI is typically assessed by an interviewer/rater based on an ordered set of ability descriptors which range from (1) Novice Low, (2) Novice Mid, (3) Novice High, (4) Intermediate Low, (5) Intermediate Mid, (6) Intermediate High, (7) Advanced Low, (8) Advanced High, to (9) Superior. On the contrary, the scoring of standardized tests such as TOEFL is typically based on learners' listening, speaking, reading and writing. Understandably, the test content of TOEFL is much more comprehensive, and the scoring metric adopts an interval scale. Conversely, the ACTFL-OPI employs a categorical scale which scores on the principle of categorical ability descriptors with only speaking performance. Furthermore, based on the scoring descriptors of the ACTFL-OPI, it is comparatively much easier to develop from Intermediate Low to Intermediate Mid than from Advanced Low to Advanced High (Dekeyser, 2007). Because most learners' proficiency levels are identified as 'intermediate' in the primary studies adopting OPI-based measures, learners assessed by OPIs may be more advantageous than those assessed by standardized tests.

It is perceivable that primary studies of L2 study abroad research may have different linguistic targets to examine such as pragmatic skills (Taguchi, 2011), vocabulary (Dewey, 2004), or grammatical accuracy (Collentine, 2004). Nonetheless, to accurately pinpoint learners' linguistic gains in different aspects, it is suggested that psychometric evidence such as reliability and validity need to be convincingly addressed and verified regardless of types of proficiency measures used. This way, the content of proficiency measures can be subject to expert scrutiny, and the test validity can be empirically obtained, thus ensuring the trustworthiness of a well-planned study abroad program.

## *6 Program content*

In terms of program content, the results revealed that the effect size of content-based instruction was smaller than that of formal language instruction and mixed content programs. In this case, the current work does not claim that study abroad language programs with content-based courses may be the least helpful to language acquisition. Rather, it is suggested that accumulated language instruction time is a significant variable in predicting learners' ultimate language proficiency. It can be inferred that, given the limited number of effect sizes ($k = 24$) available in the content-based category, the effect of delivering content-based courses in study abroad language programs has yet to be fully realized. More primary studies are warranted in order to address this line of research enquiry.

## *7 Age*

The results indicated that age is not a significant variable in predicting learners' linguistic gains in study abroad language programs. This finding did not support the results of three primary studies (Llanes & Muñoz, 2013; Llanes & Serrano, 2017; Muñoz & Llanes, 2014), which indicated that child learners did improve their oral proficiency more than adult learners. It should be noted that the participants in primary studies of the present

meta-analysis were between 10 and 33 years old. Therefore, the results suggest that learners in this age range do not need to worry too much about the age effect on language acquisition in study abroad language programs.

## 8 Pre-program training

This meta-analysis found that the inclusion of pre-program training failed to predict learners' linguistic gains. It therefore appears that the additional time of domestic language learning before a sojourn does not guarantee that study abroad programs will be more effective. As previously discussed, extended and prolonged domestic language studies may demotivate learners in terms of language acquisition. In light of these findings, further pre-program training is not recommended.

## 9 Language proficiency

In terms of the modulating effects of learners' language proficiency, both beginner ($g = 0.95$) and advanced ($g = 0.90$) level learners showed a comparatively larger effect size than that of the intermediate level learners ($g = 0.66$). Notably, the effect sizes of beginner and advanced learners were quite similar. Although the literature has shown that lower proficiency learners gain more compared with higher proficiency learners in study abroad programs due to the fact that lower proficiency learners have more room to improve their L2 skills (Brecht & Robinson, 1995; Dyson, 1988; Freed, 1995; Lapkin, Hart, & Swain, 1995; Li, 2014; Milton & Meara,1995), the outcome of the current meta-analysis is not consistent with those of the primary studies. Advanced learners may comparatively engage in more communication activities with native speakers (Brecht & Robinson, 1995) and proactively learn how to uncover different resources to acquire more input of target language (Rivers, 1998; Segalowitz & Freed, 2004). Dekeyser (2007, pp. 211–212) also remarks that 'It may very well be that the more advanced students are indeed the ones that are learning more in the long run and that the weaker students make the quickest progress at the beginning.' Collectively, therefore, based on the empirical analysis with a large number of effect sizes ($k = 273$) of the studies, as well as the theoretical underpinnings put forward by Dekeyser, the current study would like to formally propose that there may exist a U-shaped relationship between levels of language proficiency and the effects of study abroad programs.

## 10 Target language

Finally, with regard to the modulating effects of target language, the current study found that the effects of study abroad programs varied across different types of target languages. Notably, the effect sizes of French, Japanese, and Spanish consistently went beyond 1, indicative of large effect, whereas the effect size of English was just around half of that. To seek explanations for the unexpected discrepancy of effect sizes, the researchers checked the test mechanisms implemented in the primary studies and found that all the study abroad programs of English language adopted in-house assessments, and none of the study abroad programs of English used any type of OPI mechanisms. On

the other hand, all three different types of OPI test mechanisms were consistently adopted by most of the study abroad programs in the context of learning languages other than English as seen in Table 2. Essentially, ACTFL-OPI obtained the largest effect size in the moderator analysis with the test mechanism, whereas the in-house assessment mechanism received the smallest effect size. Dekeyser (2007) suggests that the linguistic gains of weaker students can be favorably projected by the OPI. It is therefore understandable that all three OPI-related test mechanisms in principle assess learners' speaking performance, and speaking also receives the largest effect size in the moderator analysis of type of outcome measure. The above analysis may explain why the effects of the study abroad programs in French, Japanese and Spanish are so salient and pronounced. In connection to the discussion of the results of the current moderator analysis, the researchers very much agree with Nassaji's (2020, p. 742) remark that 'Statistically significant results do not necessarily mean that results are practically or theoretically useful.' Along these lines, judicious deliberation behind the '*p*-value' simply needs to be well-exercised.

The above observations and discussions not only bolster previous results indicating the overall effects of study abroad language programs on L2 acquisition, but also explain the possible effects of salient moderators identified in the study. Further suggestions for both study abroad and domestic study learners are given as follows. For study abroad language learning, learners can best facilitate their language learning by taking formal language courses in addition to content-based courses. It is recommended that learners live with host families, regardless of learner proficiency level, for a longer study overseas. In order to fully assess the linguistic gains of learners with different levels of proficiency, practitioners and researchers should employ multiple outcome measures in order to obtain a full profile of linguistic gains (Borràs & Llanes, 2019; Dekeyser, 2007). With regard to learners who cannot study abroad, domestic language instruction should focus on the optimal, rather than maximal, length of the whole language learning program since extended and prolonged instruction might demotivate learners in terms of language acquisition. In this way, any pre-program training on top of domestic language instruction is not encouraged.

Admittedly, considering the number of moderators included in the current study, the modeling process was quite complex. Ten moderators were featured for analysis which is far more than those of prior similar works (Hirai, 2018; Varela, 2017; Xu, 2019; Yang, 2016). Clearly, an investigation of the modulating effects of the 10 moderators greatly helped identify and understand the degree to which the effects of study abroad language programs may vary as a function of salient factors.

However, there could be a trade-off in the implementation of a large-set moderator analysis as evidenced by Myers (1990, pp. 178–180) who stated 'a model that is too simple may suffer from biased coefficients and prediction, while an overly complicated model can result in large variances, both in the coefficients and in the prediction.' Instead of finding the 'best' and most 'economical' model, this study, like other *exploratory* meta-analyses, aimed to identify possible candidate explanatory variables (i.e. either the ones examined in study abroad studies or issues noted by study abroad scholars) in context. This, hopefully, will provide informative and empirically-established variables to be further examined in future study abroad research.

Despite the potential weaknesses in the outcome model construction, it is argued that the selection of moderators included in this exploratory meta-analysis is judiciously

determined based on a thorough analysis and understanding of study abroad language programs (see Table 2). That is, the selection of the 10 moderators investigated in this study was essentially motivated by their relevance in the study abroad context (i.e. either the variables investigated or issues highlighted by study abroad scholars) rather than chosen in a speculative manner without recourse to the parameters underlying study abroad language programs. In this way, moderator variables used in the current study have been carefully reviewed rather than arbitrarily labeled. Furthermore, the justifications for moderator variable selection were explicitly constructed and well-grounded in theory. Selecting moderators based on their relevance to the literature has been deemed a crucial step to be taken in constructing a regression model (Stevens, 2009; Weisberg, 2013). Hence, it is believed that the practice of this study will increase the explained variance in effect sizes and critically inform the future direction of primary research of study abroad language programs.

In sum, the current meta-analysis has a three-fold significance. First of all, as Oswald and Plonsky (2010) emphasize, a quantitative research synthesis can provide a holistic picture of studies showing how variables respond to stimuli in the most methodical and unbiased manner. By examining the moderating effect of 10 moderator variables and including widely collected primary studies published between 1995 and 2019, this meta-analysis is empowered to detect how individual, programmatic, and methodological related factors influence opportunities for contact with native speakers.

In addition, the database in this study is unique in its ability to present evidence for comparing L2 linguistic development between the domestic L2 learning context and study abroad context for the following reasons. In particular, this study utilized carefully collected before and after measures of the four language modalities (i.e. speaking, listening, reading and writing) in order to provide an in-depth assessment of linguistic gains. In addition, this research synthesis is rich in individual and programmatic variables. These strengths allow the current work to aid L2 researchers to receive a critical update regarding the overall effects of study abroad programs, and present L2 practitioners with helpful pedagogical implications.

Despite the merits unveiled by the current study, it is further noted that the current study improves significantly over the four prior similar works in a way that the impact of clustered effect sizes typically featured in nearly all of the primary studies in the current analysis is effectively incorporated into a three-level meta-analytic framework. The extent to which the 10 moderators may explain the heterogeneity variances existing at both the within-study level and between-study level is showcased and discussed in the study.

## VIII Conclusions

The present meta-analysis delved into issues which were under-explored or overlooked in earlier meta-analyses. Through careful examination, it has been shown that study abroad language programs may outweigh domestic language programs in achieving linguistic gains. The current meta-analysis recognizes that there are specific concerns which future research in this field should attend to. First, some primary studies did not assess the target structure before treatment, or did not offer a record of learners' pretest scores. Due to these missing pieces of essential information, it is impossible to gauge how effective the treatment was for learners.

Second, many studies did not include a control group, meaning that it is not possible to make meaningful comparisons of the study abroad language programs and domestic language programs. It is worth noting that during the process of screening out ineligible studies quite a few empirical studies, especially those published in 2019, were implemented without the inclusion of an experimental group. Hence, the results of these studies should be interpreted tentatively rather than definitely, particularly for those with large effect sizes.

Third, researchers should more closely examine young, beginner, and advanced learners since many of the primary studies focused only on intermediate and adult learners. Finally, there is a dearth of research investigating attitudinal and behavioral learning such as learner motivation, learner anxiety, willingness to communicate, and intercultural competence in primary studies. To conclude, given the effectiveness of L2 study abroad programs in the SLA context, it is important for scholars to further explore elements of these programs which might moderate their effectiveness.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD


Yeu-Ting Liu https://orcid.org/0000-0001-8055-0587

## References

*Note*. * references with an asterisk indicate the primary studies included for meta-analysis in the study

Allen, H., Dristas, V., & Mills, N. (2006). Cultural learning outcomes and summer SA. In Mantero, M. (Ed.), *Identity and second language learning: Culture, inquiry, and dialogic activity in educational contexts* (pp. 187–214). Charlotte, NC: Information Age Publishing.

Assink, M., & Wibbelink, C.J.M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*, 154–174.

Barron, A, (2006). Learning to say 'you' in German: The acquisition of sociolinguistic competence in a study abroad context. In DuFon, M.A., & E.E. Churchill (Eds.), *Language learners in study abroad contexts* (pp. 59–88). Bristol: Multilingual Matters

Borràs, J., & Llanes, À. (2019). Re-examining the impact of study abroad on L2 development: A critical overview. *The Language Learning Journal*. Epub ahead of print 26 July 2019. DOI: 10.1080/09571736.2019.1642941.

Brecht, R.D., & Robinson, J.L. (1995). On the value of formal instruction in study abroad. *Second language acquisition in a study abroad context* (pp. 317–334). Amsterdam: John Benjamins.

Cheung, W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. Chichester: John Wiley.

Coe, R. (2002). It's the effect size, stupid. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, UK. Available at: http://www.leeds.ac.uk/educol/documents/00002182.htm (accessed February 2021).

*Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, *26*, 227–248.

*Cubillos, J., Chieffo, L., & Fan, C. (2008). The impact of short-term study abroad programs on L2 learning comprehension skills. *Foreign Language Annuals*, *41*, 157–185.

Dekeyser, R. (1991). Foreign language development during a semester abroad. In Freed, B. (Ed.), *Foreign language acquisition research and the classroom* (pp. 104–119). Lexington, MA: D.C. Heath.

Dekeyser, R.M. (2007). Study abroad as foreign language practice. In Dekeyser, R.M. (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 208–226). Cambridge: Cambridge University Press.

Dekeyser, R.M. (2010). Monitoring processes in Spanish as a second language during a study abroad program. *Foreign language Annuals*, *43*, 80–92.

*Dewey, D.P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 303–327.

*Dewey, D.P. (2008). Japanese vocabulary acquisition by learners in three contexts. *Frontiers: The interdisciplinary Journal of Study Abroad*, *15*, 127–148.

Diao, W. (2017). Between the standard and non-standard: Accent and identity among transnational Mandarin speakers studying abroad in China. *System*, *71*, 87–101.

DuFon, M.A., & Churchill, E (Eds.). (2006). *Language learners in study abroad contexts*. Bristol: Multilingual Matters.

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*, 455–463.

Dyson, P. (1988). The year abroad. Unpublished report for the Central Bureau for Educational Visits and Exchanges, Oxford University Language Teaching Centre, Oxford, UK.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

*Félix-Brasdefer, J.C., & Hasler-Barker, M. (2015). Complimenting in Spanish in a short-term study abroad context. *System*, *48*, 75–85.

Foster, P. (2009). Lexical diversity and native-like selection: The bonus of studying abroad. In n: Richards, B., Daller, M.H., Malvern D.D., et al. (Eds.), *Vocabulary studies in first and second language acquisition* (pp. 91–106). London: Palgrave Macmillan.

France, H., & Rogers, L. (2012). Cuba study abroad: A pedagogical tool for reconstructing American national identity. *International Studies Perspectives*, *13*, 390–407.

Freed, B.F. (1990). Language learning in a study abroad context: the effects of interactive and non-effective out-of-class contact on grammatical achievement and oral proficiency. In Atlantic, J. (Ed.), *Linguistics, language teaching and language acquisition: The interdependence of theory, practice and research* (pp. 459–477). Georgetown University Round Table on Languages and Linguistics. Washington, DC: Georgetown University Press.

Freed, B.F. (Ed.). (1995). *Second language acquisition in a study abroad context*. Philadelphia, PA: John Benjamins.

*Freed, B.F., Segalowitz, N., & Dewey, D.P. (2004). Contexts of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26*, 275–301.

Geeslin, K.L., & Schmidt, L.B. (2018). Study abroad and L2 learner attitudes. In Sanz, C., & A. Morales-Front (Eds.). *The Routledge handbook of study abroad research and practice* (pp. 387–405). New York: Routledge.

Hernández, T.A. (2010). The relationship among motivation, interaction, and the development of second language oral proficiency in a study abroad context. *The Modern Language Journal*, *94*, 600–617.

Hirai, A. (2018). The effects of study abroad duration and predeparture proficiency on the L2 proficiency of Japanese university students: A meta-analysis approach. *JLTA Journal*, *21*, 102–123.

*Hirakawa, M., Shibuya, M., & Endo, M. (2019). Explicit instruction, input flood or study abroad: Which helps Japanese learners of English acquire adjective ordering? *Language Teaching Research*, *23*, 158–178.

Hox, J.J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. 3rd edition. New York: Routledge.

Huang, S.C. (2015). National identity (re)construction and negotiation and cosmopolitanism in the intercultural study-abroad context: Student sojourners from Taiwan in the UK. Unpublished doctoral dissertation, Durham University, Durham, UK.

*Huebner, T. (1995). The effects of overseas language programs: Report on a case study of an intensive Japanese course. In Freed, B.F. (Ed.), *Second language acquisition in a study abroad context* (pp. 171–193).

Hymes, D. (1974) *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia, PA: University of Pennsylvania Press

*Isabelli-García, C. (2010). Acquisition of Spanish gender agreement in two learning contexts: Study abroad and at home. *Foreign Language Annals*, *43*, 289–303.

*Jiménez-Jiménez, A.F. (2010). A comparative study on second language vocabulary development: Study abroad vs. classroom settings. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *19*, 105–123.

*Jochum, C.J. (2014). Measuring the effects of a semester abroad on student's oral proficiency gains: A comparison of at-home and study abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *24*, 93–104.

Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition. In Norris, J., & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Amsterdam: John Benjamins.

Khajavy, G.H., MacIntyre, P.D., & Barabadi, E. (2018). Role of the emotions and classroom environment in willingness to communicate: Applying doubly latent multilevel analysis in second language acquisition research. *Studies in Second Language Acquisition*, *40*, 605–624.

Kinginger, C. (2009). *Language learning and study abroad: A critical reading of research*. New York: Palgrave Macmillan.

Knight, S.M., & Schmidt-Rinehart, B.C. (2002). Enhancing the homestay: Study abroad from the host family's perspective. *Foreign Language Annals*, *35*, 190–201.

*Köylü, Z. (2016). The influence of context on L2 development: The case of Turkish undergraduates at home and abroad. Unpublished doctoral dissertation, University of South Florida, Tampa, FL, USA.

*Lafford, B.A. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition*, *26*, 201–225.

Lapkin, S., Hart, D., & Swain, M. (1995). A Canadian interprovincial exchange: Evaluating the linguistic impact of a three-month stay in Quebec. In Freed, B.F. (Ed.), *Second language acquisition in a study abroad context* (pp. 67–94). Amsterdam: John Benjamins.

Law, M.E. (2003). A case study of study abroad: University students learning Spanish in context. Unpublished PhD thesis, University of South Alabama, Mobile, AL, USA.

Leonard, K.R., & Shea, C.E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern Language Journal*, *101*, 179–193.

*Li, L. (2014). Language proficiency, reading development, and learning context. *Frontiers: The Interdisciplinary Journal of Study*, *24*, 73–92.

Llanes, À. (2011). The many faces of study abroad: An update on the research on L2 gains emerged during a study abroad experience. *International Journal of Multilingualism*, *8*, 189–215.
*Llanes, À. (2012). The short-and long-term effects of a short study abroad experience: The case of children. *System*, *40*, 179–190.
*Llanes, À., & Muñoz, C. (2013). Age effects in a study abroad context: Children and adults studying abroad and at home. *Language Learning*, *63*, 63–90.
*Llanes, À., & Serrano, R. (2017). The effectiveness of classroom instruction 'at home' versus study abroad for learners of English as a foreign language attending primary school, secondary school and university. *The Language Learning Journal 45*: 434–446.
*Llanes, À., Mora, J.C., & Serrano, R. (2017). Differential effects of SA and intensive AH courses on teenagers' L2 pronunciation. *International Journal of Applied Linguistics*, *27*, 470–490.
*Marqués-Pascual, L. (2011). Study abroad, previous language experience, and Spanish L2 development. *Foreign Language Annals*, *44*, 565–582.
Marsden, E., & Torgerson, C.J. (2012). Single group, pre- and post-test research. designs: Some methodological concerns. *Oxford Review of Education*, *38*, 583–616.
*Martinsen, R.A., Baker, W., Bown, J., & Johnson, C. (2011). The benefits of living in foreign language housing: The effect of language use and second-language type on oral proficiency gains. *The Modern Language Journal*, *95*, 274–290.
McMeekin, A.L. (2004). NS–NNS negotiation and communication strategy use in the host family versus the study abroad classroom. Unpublished doctoral dissertation, University of Hawaii, Honolulu, HI, USA.
Myers, R. (1990). *Classical and modern regression with applications*. 2nd edition. Boston, MA: Irwin.
Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL – International Journal of Applied Linguistics*, *107*, 17–34.
Mora, J.C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*, 610–641.
*Muñoz, C., & Llanes, Á. (2014). Study abroad and changes in degree of foreign accent in children and adults. *The Modern Language Journal*, *98*, 432–449.
Nassaji, H. (2020). Statistical significance tests in language teaching research. *Language Teaching Research*, *24*, 739–742.
Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.
*O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, *29*, 557–581.
Oswald, F.L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistic*, *30*, 85–110.
Pfenninger, S.E., & Singleton, D. (2016). Affect trumps age: A person-in-context relational view of age and motivation in SLA. *Second Language Research*, *32*, 311–345.
Plonsky, L.D., & Oswald, F.L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.
Regan, V. (1995). The acquisition of sociolinguistic native speech norms. In Freed, B.F. (Ed.), *Second language acquisition in a study abroad context* (pp. 245–267). Amsterdam: John Benjamins.
Regan, V., Howard, M., & Lemée, I. (2009). *The acquisition of sociolinguistic competence in a study abroad context*. Bristol: Multilingual Matters.
*Ren, W. (2015). *L2 pragmatic development in study abroad contexts*. Bern: Peter Lang AG, International Academic Publishers.

Rivers, W.P. (1998). Is being there enough? The effects of homestay placements on language gain during study abroad. *Foreign Language Annals*, *31*, 492–500.

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (Eds.) (2005). Publication bias in meta-analysis: Prevention, assessment, and adjustments. Chichester: Wiley.

*Sagarra, N., & LaBrozzi, R. (2018). Benefits of study abroad and working memory on L2 morphosyntactic processing. In Sanz, C, & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 149–164). New York: Routledge.

*Sasaki, M. (2004). A multiple-data analysis of the 3.5-year development of EFL student writers. *Language Learning*, *54*, 525–582.

*Sasaki, M. (2007). Effects of study-abroad experiences on EFL writers: A multiple data analysis. *The Modern Language Journal*, *91*, 602–620.

*Sasaki, M. (2009). Changes in English as a foreign language students' writing over 3.5 years: A sociocognitive account. In Manchón, R.M. (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 49–76). Clevedon: Multilingual Matters.

*Sasaki, M. (2011). Effects of varying lengths of study-abroad experiences on Japanese EFL students' L2 writing ability and motivation: A Longitudinal Study. *TESOL Quarterly*, *45*, 81–105.

Sasaki, M., Kozaki, Y., & Ross, S.J. (2017). The impact of normative environments on learner motivation and L2 reading ability growth. *The Modern Language Journal*, *101*, 163–178.

*Schenker, T. (2018). Making short-term study abroad count-effects on German language skills. *Foreign Language Annals*, *51*, 411–429.

Schmidt-Rinehart, B.C., & Knight, S.M. (2004). The homestay component of study abroad: Three perspectives. *Foreign Language Annals*, *37*, 254–262.

*Segalowitz, N., & Freed, B.F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173–199.

*Segalowitz, N., Freed, B., Collentine, J. et al. (2004). A comparison of Spanish second language acquisition in two different learning context: Study abroad and the domestic classroom. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *10*, 1–18.

*Serrano, R., Llanes, À., & Tragant, E. (2011). Analyzing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs. study abroad in Europe. *System*, *39*, 133–143.

*Serrano, R., Llanes, À., & Tragant, E. (2016). Examining L2 development in two short-term intensive programs for teenagers: Study abroad vs. 'at home'. *System*, *57*, 43–54.

*Serrano, R., Tragant, E., & Llanes, À. (2014). Summer English courses abroad versus 'at home'. *ELT Journal*, *68*, 397–409.

Solon, M., & Long, A.Y. (2018). Acquisition of phonetics and phonology abroad: What we know and how. In Sanz, C., & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 71–85). New York: Routledge.

*Steven, J.J. (2001). The acquisition of L2 Spanish pronunciation in a study abroad context. Unpublished PhD dissertation, University of Southern California, Los Angeles, CA, USA.

Stevens, J.P. (2009). *Applied multivariate statistics for the social sciences*. 5th edition. New York: Routledge.

*Sunderman, G., & Kroll, J.F. (2009). When study abroad experience fails to deliver: The internal resources threshold effect. *Applied Psycholinguistics*, *30*, 79–99.

*Taguchi, N. (2011). The effect of L2 proficiency and study-abroad experience on pragmatic comprehension. *Language Learning*, *61*, 904–939.

Tullock, B., & Ortega, L. (2017). Fluency and multilingualism in study abroad: Lessons from a scoping review. *System*, *71*, 7–21.

*Vande Berg, M.V., Connor-Linton, J., & Paige, R.M. (2009). The Georgetown consortium project: Interventions for student learning abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *18*, 1–75

Varela, O.E. (2017). Learning outcomes of study-abroad programs: A meta-analysis. *Academy of Management Learning & Education*, *16*, 531–561.

Watson, J.R., & Ebner, G. (2018). Language-learning strategy use by learners of Arabic, Chinese, and Russian during study abroad. In Sanz, C., & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 226–244). New York: Routledge.

Weisberg, S. (2013). *Applied linear regression*. 4th edition. New York: Wiley.

Wilkinson, S. (2006). *AAUSC 2006: Insights from study abroad for language programs*. Boston, MA: Heinle & Heinle.

*Winke, P., & Gass, S. (2018). When some study abroad: How returning students realign with the curriculum and impact learning. In Sanz, C., & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 527–544). New York: Routledge.

*Wu, H., & Zhang, L.J. (2017). Effects of different language environments on Chinese graduate students' perceptions of English writing and their writing performance. *System*, *65*, 164–173.

Xiao, F. (2015). Adult second language learners' pragmatic development in the study-abroad context: A review. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *25*, 132–149.

Xu, Y. (2019). Changes in interlanguage complexity during study abroad: A meta-analysis. *System*, *80*, 199–211.

Yang, J.S. (2016). The effectiveness of study-abroad on second language learning: A meta-analysis. *Canadian Modern Language Review*, *72*, 66–94.

*Yashima, T., & Zenuk-Nishide, L. (2008). The impact of learning contexts on proficiency, attitudes, and L2 communication: Creating an imagined international community. *System*, *36*, 566–585.

Zaytseva, V., Pérez-Vidal, C., & Miralpeix, I. (2018). Vocabulary acquisition during study abroad: A comprehensive review of the research. In Sanz, C., & A. Morales-Front (Eds.). *The Routledge handbook of study abroad research and practice* (pp. 210–224). New York: Routledge.

**Appendix 1.** A comparison of primary studies collected in the current meta-analysis and the prior meta-analyses.

| | *Yang (2016) | Varela (2017) | Xu (2019) | Current study |
|---|---|---|---|---|
| Hirakawa, Shibuya & Endo (2019) | | | | ✓ |
| Sagarra & LaBrozzi (2018) | | | | ✓ |
| Schenker (2018) | | | | ✓ |
| Winke & Gass (2018) | | | | ✓ |
| Wu & Zhang (2017) | | | | ✓ |
| Köylü (2016) | | | | ✓ |
| Llanes et al. (2017) | | | | ✓ |
| Serrano et al. (2016) | | | ✓ | ✓ |
| Félix-Brasdefer & Hasler-Barker (2015) | | | | ✓ |
| Ren (2015) | | | | ✓ |
| Llanes & Serrano (2017) | | | ✓ | ✓ |
| Jochum (2014) | | ✓ | | ✓ |
| Li (2014) | | ✓ | | ✓ |
| Muñoz & Llanes (2014) | | | | ✓ |
| Serrano, Tragant & Llanes (2014) | | | | ✓ |
| Llanes & Muñoz (2013) | | | ✓ | ✓ |
| Llanes (2012) | | ✓ | | ✓ |
| Marqués-Pascual (2011) | ✓ | | | ✓ |
| Martinsen et al. (2011) | ✓ | ✓ | | ✓ |
| Sasaki (2011) | ✓ | | | ✓ |
| Serrano, Llanes & Tragent (2011) | ✓ | ✓ | ✓ | ✓ |
| Taguchi (2011) | ✓ | ✓ | | ✓ |
| Isabelli-García (2010) | ✓ | | | ✓ |
| Jiménez-Jiménez (2010) | | ✓ | | ✓ |
| Foster (2009) | | | | ✓ |
| Sunderman & Kroll (2009) | | | | ✓ |
| Vande Berg et al. (2009) | | ✓ | | ✓ |
| Cubillos et al. (2008) | | ✓ | | ✓ |
| Dewey (2008) | ✓ | ✓ | | ✓ |
| O'Brien et al. (2007) | ✓ | ✓ | | ✓ |
| Sasaki (2007) | | ✓ | | ✓ |
| Collentine (2004) | ✓ | | | ✓ |
| Dewey (2004) | | ✓ | | ✓ |
| Freed et al. (2004) | | ✓ | | ✓ |
| Sasaki (2004) | | | | ✓ |
| Segalowitz & Freed (2004) | ✓ | ✓ | | ✓ |
| Segalowitz et al. (2004) | | | | ✓ |
| Yashima & Zenuk-Nishide (2008) | | | | ✓ |
| Steven (2001) | | | | ✓ |
| Huebner (1995) | | | | ✓ |
| Freed (1995) | | | | ✓ |

*Note.* * One primary study (Lafford, 2004) included in Yang's meta-analysis measured communication strategy rather than pragmatic proficiency; Lafford's study was therefore excluded from the current meta-analysis.

**Appendix 2.** The Inter-Rater Coding Scheme

| Study names | Outcome measure | | Language proficiency | | Test mechanism | | Type of residence | | Program content | | Learners' age | | Length of SA | | Length of AH | | Target language | | Preprogram training | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rater | A | B | A | B | A | B | A | B | A | B | A | B | A | B | A | B | A | B | A | B |
| Hirakawa et al. (2019) | 3 | 1 | 3 | 2 | 3 | 4 | 3 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 3 | 2 | 3 | 2 |
| Schenker (2018) | 11 | 11 | 3 | 3 | 5 | 5 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 4 | 4 | 1 | 1 |
| Winke & Gass (2018) | 6 | 6 | 2 | 2 | 1 | 1 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 7 | 7 | 2 | 2 |
| Sagarra & LaBrozzi (2018) | 8 | 8 | 1 | 1 | 4 | 4 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 6 | 6 | 2 | 2 |
| Wu & Zhang (2017) | 9 | 9 | 3 | 3 | 4 | 4 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| Llanes et al. (2017) | 6 | 6 | 2 | 2 | 4 | 4 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Serrano et al. (2016) | 11 | 11 | 2 | 2 | 6 | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Köylü (2016) | 11 | 11 | 2 | 2 | 4 | 4 | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| Ren (2015) | 4 | 4 | 3 | 3 | 4 | 4 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Félix-Brasdefer & Hasler-Barker, 2015 | 6 | 6 | 2 | 2 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 6 | 6 | 2 | 2 |
| Liu (2014) | 11 | 8 | 4 | 4 | 6 | 3 | 1 | 3 | 1 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 3 | 2 | 3 |
| Llanes & Serrano (2017) | 6 | 6 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Jochum (2014) | 6 | 6 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 6 | 6 | 2 | 2 |
| Muñoz & Llanes (2014) | 6 | 6 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Serrano et al. (2014) | 9 | 9 | 2 | 2 | 4 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 4 | 2 | 2 | 2 |
| Llanes & Muñoz (2013) | 11 | 11 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Llanes (2012) | 11 | 11 | 1 | 1 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| Taguchi (2011) | 4 | 4 | 3 | 3 | 4 | 4 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| Serrano et al. (2011) | 6 | 6 | 2 | 2 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| Marqués-Pascual (2011) | 6 | 6 | 4 | 4 | 4 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 6 | 6 | 2 | 2 |
| Martinsen et al. (2011) | 6 | 6 | 4 | 4 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 7 | 2 | 2 |
| Sasaki (2011) | 9 | 9 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |

*(Continued)*

**Appendix 2.** (Continued)

| Study names | Outcome measure | | Language proficiency | | Test mechanism | | Type of residence | | Program content | | Learners' age | | Length of SA | | Length of AH | | Target language | | Preprogram training | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rater | A | B | A | B | A | B | A | B | A | B | A | B | A | B | A | B | A | B | A | B |
| Isabelli-García et al. (2010) | 1 | 1 | 2 | 2 | 4 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 6 | 6 | 2 | 2 |
| Jiménez-Jiménez (2010) | 7 | 7 | 3 | 3 | 4 | 4 | 2 | 3 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 6 | 6 | 3 | 3 |
| Foster (2009) | 6 | 6 | 2 | 2 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Vande Berg et al. (2009) | 6 | 6 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 7 | 7 | 2 | 2 |
| Sunderman & Kroll (2009) | 8 | 8 | 2 | 2 | 4 | 4 | 4 | 4 | 4 | 4 | 2 | 2 | 1 | 1 | 1 | 1 | 6 | 6 | 3 | 3 |
| Sasaki (2009) | 9 | 9 | 2 | 2 | 4 | 4 | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| Cubillos et al. (2008) | 3 | 3 | 2 | 2 | 4 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 6 | 6 | 2 | 2 |
| Dewey (2008) | 7 | 7 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 5 | 5 | 2 | 2 |
| Sasaki (2007) | 11 | 11 | 2 | 2 | 4 | 4 | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| O'Brien et al. (2007) | 6 | 6 | 4 | 4 | 6 | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 6 | 6 | 2 | 2 |
| Segalowitz et al. (2004) | 11 | 11 | 2 | 2 | 1 | 1 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 2 | 1 | 1 | 6 | 6 | 2 | 2 |
| Collentine (2004) | 6 | 6 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 6 | 6 | 2 | 2 |
| Dewey (2004) | 5 | 5 | 2 | 2 | 4 | 4 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 5 | 5 | 2 | 2 |
| Freed et al. (2004) | 6 | 6 | 2 | 2 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 2 | 2 |
| Segalowitz & Freed (2004) | 6 | 6 | 2 | 2 | 1 | 1 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 2 | 2 | 2 | 3 | 3 | 2 | 2 |
| Sasaki (2004) | 9 | 9 | 1 | 1 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| Yashima & Zenuk-Nishide (2008) | 10 | 10 | 1 | 1 | 5 | 5 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Steven (2001) | 6 | 6 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 6 | 6 | 2 | 2 |
| Huebner (1995) | 11 | 11 | 1 | 1 | 6 | 6 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 5 | 5 | 2 | 2 |
| Freed (1995) | 6 | 6 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 2 | 2 |
| *Kappa* | 0.92 | | 0.98 | | 0.93 | | 0.88 | | 0.88 | | 0.95 | | 0.90 | | 0.95 | | 0.94 | | 0.82 | |
| *Average of Kappa* | 0.92 | | | | | | | | | | | | | | | | | | | |

**Appendix 2.** (Continued)

| Moderators | Coding content |
|---|---|
| Outcome measure | 1 = Grammar<br>2 = Lexical-grammatical knowledge<br>3 = Listening<br>4 = Pragmatic knowledge<br>5 = Reading<br>6 = Speaking<br>7 = Vocabulary knowledge$_{receptive}$<br>8 = Working memory<br>9 = Writing<br>10 = Listening + Grammar + Reading<br>11 = Other |
| Language proficiency | 1 = Basic<br>2 = Intermediate<br>3 = Advanced<br>4 = Mixed |
| Test mechanism | 1 = ACTFL-OPI<br>2 = ILR-OPI<br>3 = Home-spun OPI<br>4 = In-house assessment<br>5 = Standardized test<br>6 = Mixed |
| Type of residence | 1 = Host family<br>2 = School-based dormitory<br>3 = Non-school-based dormitory<br>4 = School-based dormitory + Host family<br>5 = Non-school-based dormitory + Host family<br>6 = School-based dormitory + Non-school-based dormitory<br>7 = School-based dormitory + Non-school-based dormitory + Host family<br>8 = n/a<br>9 = Other |

*(Continued)*

**Appendix 2.** (Continued)

| Moderators | Coding content |
|---|---|
| Program content | 1 = Content-based course<br>2 = Form-based course<br>3 = Mixed<br>4 = n/a |
| Learners' age | 1 = Agree<br>2 = Disagree |
| Length of SA | 1 = Agree<br>2 = Disagree |
| Length of AH | 1 = Agree<br>2 = Disagree |
| Target language | 1 = Chinese<br>2 = English<br>3 = French<br>4 = German<br>5 = Japanese<br>6 = Spanish<br>7 = Mixed |
| Preprogram training | 1 = Yes<br>2 = No<br>3 = n/a |