Received: 3 April 2021 | Accepted: 15 September 2021

DOI: 10.1002/rev3.3306

SYNTHESIS

Review of Education BERA

Systematic review measuring the efficacy of study abroad in undergraduate language learners on linguistic proficiency gains

Ian Moore | **Carole Torgerson** | **Nadin Beckmann**

School of Education, Durham, UK

Correspondence
Ian Moore, School of Education, Leazes Road, Durham DH1 1TA, UK.
Email: ian.moore45@btinternet.com

Funding information
ESRC, Grant/Award Number: ES/J500082/1

Abstract

Ascertaining the value of a study abroad experience on facilitating language development has long been a goal of Second Language Acquisition scholars. There is currently a plethora of studies regarding the topic, yet, despite this, the evidence remains inconclusive and contradictory. This systematic review evaluates the impact of study abroad among undergraduate language learners in both Europe and beyond on a range of linguistic outcomes compared to remaining in domestic instruction. Studies that used a randomised controlled trial design, or a quasi-experimental design which achieved baseline equivalence, were included in an in-depth review as they offer the best available evidence on impact. Nine electronic databases were searched using a series of keywords. Articles were screened using pre-specified inclusion criteria. Forty studies were identified for a mapping synthesis, with seven publications being carried forward to full data extraction and quality appraisal. The synthesis of the evidence indicated that sojourning can facilitate the development of global proficiency and oral fluency, while oral and written accuracy demonstrated less

change. The available evidence is discussed in relation to policy and practice, together with the limitations of this review.

KEYWORDS
proficiency, second language acquisition, study abroad, systematic review

Context and implications

Rationale for this study
The past two decades have led to a plethora of published material on study abroad linguistic research. However, the reliance on narrative literature reviews has led to mixed and inconclusive findings.
Why the new findings matter
The findings contribute to synthesising a broad evidence base offering insights to both practitioners and policy makers alike.
Implications for educational researchers and policy makers
This review is important for practitioners and learners alike as it indicates realistic expectations concerning linguistic proficiency pre-departure to a year studying abroad. The evidence from the review shows that, where expectations have not been met, learners can become disillusioned with learning the second language. These findings could, for example, be used in pre-departure workshops to facilitate the setting up of realistic expectations. From a policy perspective, the findings support the importance of study abroad in second language acquisition and the need for governments to ensure learners have this opportunity.

BACKGROUND AND PREVIOUS REVIEWS

Each year, thousands of language students worldwide undertake a study abroad (SA) experience which can provide an opportunity for linguistic and cultural immersion. For the purposes of this review, SA is defined as the following, taken from Kinginger (2009, p. 11):

> a temporary sojourn of pre-defined duration, undertaken for educational purposes.

While sojourning is not a new phenomenon, the past three decades have seen increasing attention given to linguistic outcomes during SA (e.g., Freed, 1995; Sanz & Morales-Front, 2018). This literature has mirrored the growing number of sojourners, which according to Tullock and Ortega (2018), can be described as reaching *maturity*, having reached a historical peak in publication numbers between 2011 and 2014. As a topic, several journals (e.g., System) have published special issues, while the journal *Study Abroad Research in Second Language Acquisition and International Education* is explicitly devoted to research concerning language learning abroad.

Despite this accruement in literature, findings have generally proven contradictory and inconclusive, much to the frustration of scholars, who, supported by second language

acquisition (SLA) theory, would purport that SA should facilitate language acquisition. This is because learners are assumed to be afforded numerous linguistic opportunities from which to actively use the learnt language (Rehner & Mougeon, 2003). The notion of just how *immersive* the experience is, has been challenged by scholars (e.g., Coleman, 2015) but, undoubtedly, the perception of the SA being an i*mmersive* experience prevails in the wider public (Hessel, 2017).

Understanding and appreciating the extent of linguistic gain on a SA serves as a fundamental component in setting realistic expectations for language learners at pre-departure. Research has shown that learners tend to begin a SA with high linguistic expectations but that these expectations are rarely met (e.g., Bädstuber & Ecke, 2009). This failure can lead to disappointment, frustration and ultimately disillusionment with the second language (L2) (Wilkinson, 1998). Consequently, systematic reviews can serve as a means of distilling and disseminating a large body of work to practitioners, which may be then used to inform language learners before departure.

Of special interest to this review is a previously published meta-analysis conducted by Yang (2016) and a scoping review conducted by Tullock and Ortega (2018). Yang (2016) conducted a meta-analysis on 11 studies that matched a set criterion (e.g., control-group design, with a dependent variable of a measure of linguistic proficiency), all of which were dated pre-2011. The weighted effect size of the 11 studies combined was $d = 0.75$, indicating a medium to large effect size (Plonsky & Oswald, 2014), with four studies reporting a large effect size of over $d = 0.8$.[1] Yang (2016) concluded that the data supported the view that SA 'could lead to greater L2 linguistic attainment compared to at-home classroom learning' (p. 78). Tullock and Ortega presented a scoping review on the topic of oral fluency. For studies to be included in this review, they must have followed a pre-post design, measuring linguistic outcomes through the means of observational, behavioural measures. In total, 401 records were screened, with 31 having a specific outcome measure of oral fluency, and as such, were carried forward to a meta-analysis. The results of the meta-analysis proved rather inconclusive. When exploring differing domains of oral fluency (e.g., speech rate, pause frequency) the results displayed inconsistencies in effect direction, with only speech rate indicating consistent positive effects. When investigating speech rate specifically, the individual effect sizes of each study fell between $d = 0.5$ and $d = 1.2$[1] representing the full magnitudes of field-specific benchmarks in the field (see Plonsky & Oswald, 2014). Tullock and Ortega (2018) concluded by stating 'we can answer, but only tentatively that students probably become more fluent after a SA experience' (p. 14). In sum, these two reviews have demonstrated the broad range of findings in the literature and the difficulties associated with determining firm conclusions regarding the extent of the gain. Both reviews included non-European evidence, which may only further exacerbate the inconclusiveness of findings. The next section details the importance in focusing on a particular context when describing findings.

THE IMPORTANCE OF THE EUROPEAN PERSPECTIVE

The prevalence of North American based literature has made it sometimes difficult for European readers to place a study's findings within a European context. Such differences were first highlighted by Coleman (1998) and reiterated by other scholars (e.g., McManus et al., 2020). North American students tend to undertake short stays (an average of six weeks), remain in their established contact groups and generally have limited prior language instruction, reflected by a low proficiency. European learners, on the other hand, tend to undertake long stays abroad (three or more months), live independently within the host community, and be of high proficiency in the L2, given the multilingual status of many Europeans. Moreover, narrative reviews

have tended to disregard the importance of participant and environmental factors in influencing the extent of linguistic change, potentially leading to varied findings in a particular outcome. This systematic review has tried to combat this and given that the wider context of this work was within a European context, the review has made a concerted effort to highlight European literature and to form conclusions that are reflective of European learners. Next, focus will be given to understanding the importance of causal inference within the evidence base.

EVALUATING LANGUAGE GAIN MADE DURING STUDY ABROAD

An outstanding question for scholars is attributing change in linguistic proficiency to that of the SA itself. Typically, scholars have employed a matched-comparison design (i.e., quasi-experimental design) whereby sojourners (SA) are matched to learners in a classroom setting (AH). Any differences obtained between these learners over time can be attributed to the learning context given the differences in the amount, type and frequency of L2 comprehension and production opportunities (Grey, 2018). While such a design can introduce a number of confounding variables, given the non-random assignment into groups (AH vs. SA), the impact on findings can be reduced (although not minimised or eliminated) by implementing a pre-test. By collecting proficiency scores at pre-test, any differences between the two groups can be controlled for in the statistical analysis. Nevertheless, one should be mindful that the larger the difference in pre-test scores between the two groups, the stronger the presence of selection bias (Shadish et al., 2002), and the less reliable overall conclusions will be. Ideally, the two groups should be non-significantly different to rule out the presence of strong selection bias.

RATIONALE

This review serves as the first systematic synthesis, to the best of our knowledge, to explore the efficacy of SA on facilitating language gain compared to remaining AH. While many reviews are narrative in design, this review differs because it employs open, transparent and systematic design and methods. To date, SLA literature has explored linguistic gains in sojourners from a number of perspectives and angles but in doing so, the available evidence appears fragmented and highly compartmentalised by outcome. There is a subsequent need to systematically synthesise the available literature so that stakeholders (e.g., practitioners, students, policy makers) can be informed of the available evidence and make evidence-based judgements regarding the value of the sojourning experience on language development. By critically evaluating the available evidence, we can determine the strength of causal inference in the conclusions made, which as aforementioned are required to fully ascertain whether SA achieves what it sets out to achieve. The review comes at a time where the need to synthesise the literature is great and it is believed to hold relevance to practitioners and learners alike. Next, an overview of the design and methods will be given.

DESIGN AND METHODS

This systematic review is guided by the following research questions:

1. How effective is the ERASMUS programme in facilitating additional language acquisition?
2. How effective are study abroad programmes in facilitating second language acquisition in language learners compared to learners who remain in the domestic classroom?

Both the structure and findings of this systematic review have been reported following the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) and the *Synthesis without Meta-analysis (SWiM) in Systematic Reviews* (Campbell et al., 2020) reporting guidelines. SWiM guidelines (see Table 1) should be implemented when a review, such as this current review, does not require a meta-analysis. It is, to the best of our knowledge, the first use of SWiM guidelines in the field of education.

INCLUSION/EXCLUSION CRITERIA

Table 2 presents the inclusion/exclusion criteria used to guide the decision-making process while screening. These criteria were developed before running the main searches. Of importance in this review was the need to capture both European orientated and non-European orientated studies. As aforementioned, the imbalance to date, regarding sample composition (i.e., the prevalence of North American undergraduate students) meant a stricter study design criterion was implemented for non-European exchange programmes. In doing so, it was hoped the review would hold generalisability for both European and non-European samples. The remaining criteria focused on the characteristics of the intervention, participants and relevant linguistic outcomes. There were no restrictions placed on publication type or date.

ELECTRONIC SEARCHING

Prior to electronic searching, a protocol was developed in accordance with PRISMA-P (Shamseer et al., 2015). Searching began in April 2019, with a broad range of databases selected, reflecting the range of disciplines interested in SA research. The following databases were searched: Web of Science; Article First; Eco; British Education Index; ERIC; PsycInfo; PsycArticles, Proquest and Scopus within one week, with records imported into Endnote (version x8, 2016) where de-duplication occurred. Once completed, records were transferred into EPPI (Thomas et al., 2010), where they were screened.

First stage of screening

At this stage, all records were screened on title and abstract, using the inclusion/exclusion criteria as guidance in the decision process. All records were double screened, and where disagreement existed, the record was sent to the third reviewer for arbitration.

Second stage of screening

At this stage, the primary researcher searched for and collected the publications of records included. All records were again double screened and in instances of disagreement, a third member of the review team arbitrated.

Third stage of screening

Studies included at this stage went forward to be included in a mapping table. Data extraction was undertaken on each study by the primary reviewer on study design, intervention,

TABLE 1 Synthesis without meta-analysis (SWiM) items

SWiM reporting item	Item description	Page in manuscript where item is reported	Other
Methods			
1. Grouping studies for synthesis	(1a) Provide a description of, and rationale for, the groups used in the synthesis (e.g., groupings of populations, interventions, outcomes, study design)	16	
	(1b) Detail and provide rationale for any changes made subsequent to the protocol in the groups used in the synthesis	NA	
2. Describe the standardised metric and transformation methods used	Describe the standardised metric for each outcome. Explain why the metric(s) was chosen and describe any methods used to transform the intervention effects, as reported in the study, to the standardised metric, citing any methodological guidance consulted	NA	
3. Describe the synthesis methods	Describe and justify the methods used to synthesise the effects for each outcome when it was not possible to undertake a meta-analysis of effect estimates	NA	
4. Criteria used to prioritise results for summary and synthesis	Where applicable, provide the criteria used, with supporting justification, to select the particular studies, or a particular study, for the main synthesis or to draw conclusions from the synthesis (e.g., based on study design, risk of bias assessments, directness in relation to the review question)	9	
5. Investigation of heterogeneity in reported effects	State the method(s) used to examine heterogeneity in reported effects when it was not possible to undertake a meta-analysis of effect estimates and its extensions to investigate heterogeneity	NA	
6. Certainty of evidence	Describe the methods used to assess the certainty of the synthesis findings	15	
7. Data presentation methods	Describe the graphical and tabular methods used to present the effects (e.g., tables, forest plots, harvest plots)	11–14	
	Specify key study characteristics (e.g., study design, risk of bias) used to order the studies, in the text and any tables or graphs, clearly referencing the studies included		
Results			
8. Reporting results	For each comparison and outcome, provide a description of the synthesised findings and the certainty of the findings. Describe the result in language that is consistent with the question the synthesis addresses, and indicate which studies contribute to the synthesis	16–21	
Discussion			
9. Limitations of the synthesis	Report the limitations of the synthesis methods used and/or the groupings used in the synthesis and how these affect the conclusions that can be drawn in relation to the original review question	23	

outcomes, instruments, participants and key findings. At this stage also, those studies which allowed for stronger causal inference were carried forward to the in-depth review. Studies that met such criteria were QEDs, with pre-test proficiency equivalence. Pre-test equivalence was observed in studies that explicitly provided mean scores at pre-test, and for which the mean difference in scores between each group was considered non-significant (i.e., $p = > 0.05$). For all of the included in-depth studies, data extraction was undertaken by two reviewers.

Data extraction and quality appraisal

For each study, data on intervention type (e.g., length of stay), study design, description of sample characteristics, study measures, outcomes, results and conclusions were extracted. Each study was also quality appraised using a specifically designed tool. Studies were quality appraised on accounting for the counterfactual, attrition and pre-specified outcomes. The results are presented below.

RESULTS

In total, 2548 records were found through the database and grey literature searches. After de-duplication in Endnote x8 (2016), 1533 records were imported into EPPI, on which the first stage of screening was conducted. At the first stage of screening, 1347 records (87.8%) were excluded on title and abstract, meaning 186 publications were then located for the second stage of screening: screen on full text. At the second stage of screening, 140 studies were excluded. Twenty-seven of these were excluded as the relevant publication could not be located, three due to duplication, and the remaining 110 due to not meeting the inclusion/exclusion criteria. Subsequently, 46 studies were screened at stage three. Of these, 40 were included in an evidence mapping table as four were excluded on intervention and a further two were excluded on study design. The online supplementary material highlights these 40 studies in greater detail. Of these 40 studies, six studies (seven publications) were carried forward to the in-depth review and were used to inform the findings and conclusions of this systematic review. These were chosen because they utilised a quasi-experimental design and reported pre-test equivalence (i.e., the mean proficiency scores between SA and AH at pre-test were not significantly different). Figure 1 serves as a PRISMA flow diagram.

MAPPING SYNTHESIS

The methodological landscape

A small subset of studies identified utilised a quasi-experimental design where between-group analysis was conducted to compare change between an intervention and comparison group. These studies are explored in more detail within the narrative synthesis and listed in Table 3.

Given the number of long-term projects exploring language change within European learners, such as the Study Abroad and Language Acquisition (SALA) project in Spain (e.g., Pérez-Vidal, 2014) and the Languages and Social Networks Abroad (LANGSNAP) project in the UK (e.g., Mitchell et al., 2017), the field benefits from several studies of longitudinal designs,

TABLE 2 Inclusion/exclusion criteria

Included	Excluded
Topic: Study abroad, including affiliated organisations such as the British Council	Topic: Non-study abroad related interventions
Date: No time restriction	Date: -
Publication status: All published and unpublished material which is in the public domain	Publication status: -
Study design	Study design
Non-ERASMUS: Any study design where there is a control or comparison group—RCT (individual and cluster); quasi-experiment (interrupted/control time-series designs, control group post-test only, control group pre/post-test)	Non-ERASMUS: Case-study designs; designs with only post-test and no control group; basic time-series designs. Review articles and non-empirical literature
ERASMUS: All aforementioned designs AND pre-experimental designs (e.g., pre/post-test with no control group)	ERASMUS: Case-study designs; designs with only post-test and no control group. Review articles and non-empirical literature
Participants: Undergraduate students undertaking a study abroad as part of their academic degree studies. Control students must be a comparable group (e.g., matched comparisons at pre-test) and hold characteristic similarities to those who go abroad at pre-test	Participants: Non-academic learners or are under the age of 18
Intervention: Studies which include a study abroad which is longer than five weeks in length	Intervention: Does not have a study abroad component. A length of stay five weeks or less
Outcomes: Studies where learners are measured at post-test on any linguistic skill and their relevant skill outcome, e.g., speaking, writing, reading, listening, pragmatics. This can be measured through a multitude of instruments—for example, length of utterances, length of prose, speech/written accuracy/fluency, reading score, listening score, grammatical score. The outcome must be objective (i.e., not self-report)	Outcomes: Measures not looking at linguistic gain, e.g., intercultural competency Outcomes which are self-rated/perceived change, e.g., on a scale of 1–10, how much do you believe you have improved?

examining the same learner within both an at-home (control) and study abroad (intervention) contexts. These within-subject designs have the advantage of minimising variability between subjects and require fewer participants for analysis, something which is important given the generally high attrition rates in longitudinal studies caused by participant fatigue (Field, 2013).

Other authors (e.g., Milton & Meara, 1995; Regan, 1995) have opted to implement pre/post-test designs which allow for inferential statistical analysis to be undertaken. The lack of a counterfactual group means any change identified cannot be attributed to the intervention itself but can indicate how participants develop during the SA period (although this could be due to confounders). A pre/post-test design was generally utilised in earlier dated studies.

Alternatively, a cross-sectional design has been implemented by some authors (e.g., Howard, 2005, 2008). Such a design measures attainment at one point in time, serving as an efficient, less intensive data collection design. Given that ability is examined only once, change cannot be captured, nor can a cross-sectional design derive causality.

An overview of the European landscape

Improvements in oral fluency have been consistent regardless of the length of stay or the learner's respective L1 or L2. Here, learners typically demonstrate an enhanced ability to

process language and its physical manifestations, with oral fluency typically being measured using role-play activities (e.g., Juan-Garau, 2014), picture-based tasks (Llanes et al., 2012) or interviews (Valls-Ferrer & Mora, 2014). Similarly, oral accuracy has tended to demonstrate significant improvement over time, regardless of the length of stay (Juan-Garau, 2014), suggesting that learners return home making fewer errors in their speech compared to remaining AH. Regarding pronunciation, evidence generally indicates that a period abroad has little benefit to improving foreign accent (e.g., Avello, 2014) although pronunciation accuracy after a period abroad was found to be significantly greater (e.g., Avello et al., 2012). As a skill, pronunciation has been explored to a lesser extent than fluency and accuracy.

Regarding writing, sojourning appears to have a positive effect on both fluency and accuracy. Learners typically return home with an increased ability to produce more content and make fewer errors (e.g., Mitchell et al., 2017). When compared to oral skills, writing ability tends to improve at a slower rate (Serrano et al., 2012), and is perhaps reflective of the more limited opportunities learners receive to write during the period abroad compared to speaking. Findings were also found to be inconclusive regarding the extent to which three months is a sufficient length of time to demonstrate change, although such a finding may be borne from the use of different measures to ascertain writing ability.

The receptive skills of reading and listening were found to be understudied compared to the above productive skills. Indeed, no European study was identified as examining reading skill change. Nevertheless, regarding listening, the results appeared promising with Beattie et al. (2014) finding listening skills to significantly improve while abroad compared to being at home.

Similarly, studies have indicated both active and receptive vocabularies to significantly grow during a period abroad, suggesting that learners return home with a significantly increased ability to both comprehend a word and actively use it in its correct context (Ife et al., 2000; Milton & Meara, 1995). Grammatical skills have, on the other hand, demonstrated more mixed evidence and improvements are more nuanced to the grammatical aspect under study. For example, Edmonds and Gudmestad (2018) found target-like rates of gender markings to significantly increase across the duration of the year abroad, whereas Howard (2008) found learners demonstrated little improvement over a stay-at-home counterfactual group when using the subjunctive in spoken French. Regarding grammar, it can perhaps be argued that the skills acquired later by native L1 speakers are those which prove most troublesome for L2 learners and require an extended stay of longer than a year to drive forward change.

Only one study was identified as focusing on pragmatics (Barron, 2019). Here, it was found that a sojourn had both positive and negative consequences on pragmatic development. Exploring the specific feature of apologies, Barron (2019) demonstrated the use of some apologies to remain stable throughout the year, some to increase in their usage and some to decrease. Barron (2019, p. 13) described the findings as being 'complex' and 'non-homogeneous', although more research is required to form stronger conclusions.

Lastly, European studies have generally found a period abroad to significantly improve a learner's general proficiency (Hessel, 2016; Rees & Klapper, 2007), as typically measured through means of a C-test. As a caveat, however, these gains have been subject to substantial individual differences, suggesting gains to be non-homogeneous and dependent on a number of factors, often outside the remit of the study itself. This research has also tended to indicate that the most substantial gains are seen in the first three months of the study abroad period and that the rate of the gain slows, the longer an individual is abroad. This has subsequently led to questions regarding the extent to which long-term study abroad stays are beneficial to linguistic development.

In sum, European studies have tended to demonstrate that a period abroad can potentially facilitate and foster linguistic change, particularly in the domains of oral ability and

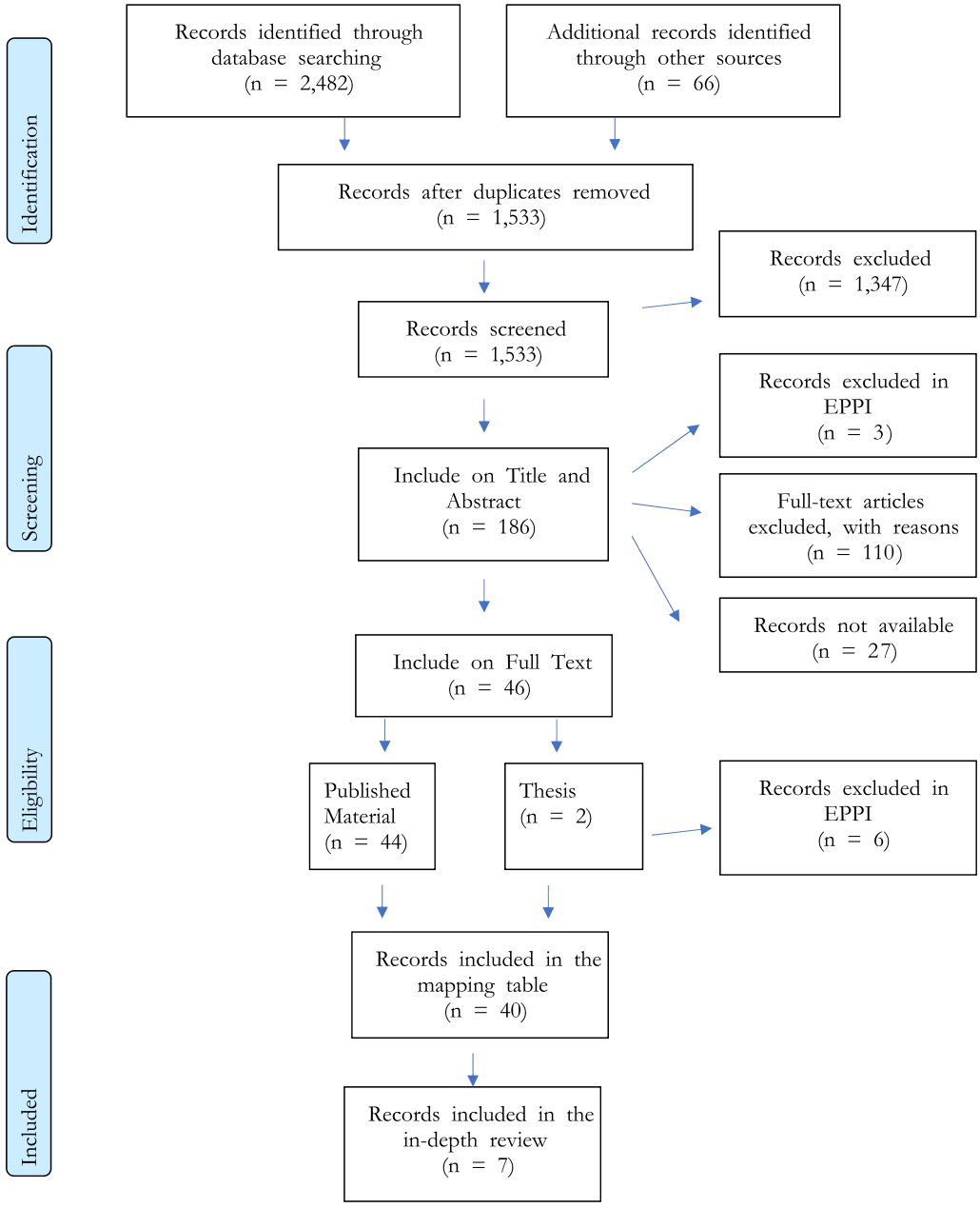

FIGURE 1 PRISMA flow-diagram from: Moher et al. (2009)

general proficiency. This is generally in line with findings outside of Europe, most pertinently the United States (US), indicating that despite often substantial differences in the learners' characteristics, the findings of US-orientated studies are somewhat generalisable to the European context. However, regardless of the extent of change, the literature described above cannot, due to limitations in study design, causally attribute the change identified to the sojourn experience itself. This has typically been a weakness of studies included in narrative reviews, and as such this new systematic review provides an in-depth synthesis of literature that has the potential to infer causality in respect of the intervention and outcome.

TABLE 3 Mapping table (in-depth review)

Study and author	Study design	Intervention	Outcomes and instrument	Participants	Key findings
Hessel (2016). The impact of participation in ERASMUS study abroad in the UK on students' overall English language proficiency, self-efficacy, English use anxiety and self-motivation to continue learning English: a mixed methods investigation AND Hessel and Vanderplank (2018). What difference does it make? Examining English proficiency gain as an outcome of participation in ERASMUS study abroad programmes in the UK	Pre-test/post-test, quasi-experimental design. Non-random group assignment. Participation voluntary	ERASMUS exchange programme. L1 German learners of English studying at an L1 English speaking university. Intervention group consists of those who stay abroad for 3 months and 9 months. Control group consists of domestic-based learners who failed in their application onto ERASMUS	General L2 proficiency measured via a c-test	143 L2 learners of English split across three groups. Short stay ($n = 45$); Long-stay ($n = 54$); Control ($n = 44$) Mean previous experience of English: 8.69 Average starting proficiency of all groups was B2 (upper intermediate)	Both YA groups experienced significant improvement in overall L2 proficiency across 3 months. AH group made no significant change during this time Long-stay group maintained proficiency gains—significant difference between T2 and T3. AH group also made significant gains between T2 and T3 and between-group differences at T3 were not significant

(Continues)

TABLE 3 (Continued)

Study and author	Study design	Intervention	Outcomes and instrument	Participants	Key findings
Llanes and Muñoz (2013). Age Effects in a Study Abroad Context: Children and Adults Studying Abroad and at-home	Pre-test/post-test, quasi-experimental design. Group assignment non-random. Participation voluntary	ERASMUS exchange programme. L1 Spanish learners of English spending 2 or 3 months living in an English-speaking country	Oral and written complexity, accuracy and fluency Writing: Produce two descriptive essays with learners given 15 min to complete task Oral: Picture-elicited narrative task	66 L2 learners of English split across two contexts SA ($n = 46$); AH intensive ($n = 20$) Mean age of all adults = 20.9; average age of onset = 8.42 and all had received over 1620 h of formal instruction	SA context more beneficial than the AH context for the improvement in oral skills but no real difference in writing skills
Serrano et al. (2011). Analysing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs study abroad in Europe	Pre-test/post-test, quasi-experimental design. Group assignment non-random. Participation voluntary	ERASMUS exchange programme. L1 Spanish learners of English spending time either abroad or at home Intervention group consists of those who stay abroad for 3 months. Control group consists of two types of domestic learner: (a) semi-intensive AH (b) intensive AH—groups differ on contact hours with L2	Oral and written complexity, accuracy and fluency Writing: Produce two descriptive essays (one pre-test, one post-test) with learners given 15 min to complete task Oral: Picture-elicited narrative task	131 L2 learners of English split across three groups. SA ($n = 25$); AH intensive ($n = 69$); AH semi-intensive ($n = 37$)	No differences between intensive AH and SA groups on writing and oral measures Compared to AH semi-intensive, SA group significantly developed more in written and oral productions in terms of fluency and lexical complexity

TABLE 3 (Continued)

Study and author	Study design	Intervention	Outcomes and instrument	Participants	Key findings
Li (2014). Language Proficiency, Reading Development, and Learning Context	Pre/post-test, quasi-experimental design. Group assignment non-random. Participation voluntary	Non-ERASMUS exchange programme L1 English learners of Chinese Intervention group stays abroad for eight weeks. Control group remains in formal domestic instruction (1 semester)	Chinese language global proficiency test capturing components of listening, grammar, reading, translation and writing Reading comprehension test: 10 multiple choice questions	73 L2 learners of Chinese split across three proficiency groups. SA group ($n = 35$); mean age 20.7. AH group ($n = 38$); mean age 21.4	Proficiency: Beginners showed no change in either context. Sojourners in Intermediate and Advanced groups significantly improved more than AH group Reading: Significant difference in Intermediate group only
Jochum (2014). Measuring the Effects of a Semester Abroad on Students' Oral Proficiency Gains: A Comparison of At-Home and Study Abroad	Pre/post-test, quasi-experimental design. Group assignment non-random. Purposive sampling based on a set criterion with students volunteering to partake	Non-ERASMUS exchange programme L1 English learners of Spanish studying at an L1 Spanish speaking university. Intervention group stays abroad for 3 months. Control group remains in formal domestic instruction	Oral proficiency as measured by the Oral Proficiency Interview (OPI)	18 L2 learners of Spanish. SA group ($n = 9$); AH group ($n = 9$) All had completed 1–8 semesters of Spanish study with mean proficiency at Int-low for both groups	SA group improved their oral proficiency more than those AH and at post-test significant between-group differences were found

(Continues)

TABLE 3 (Continued)

Study and author	Study design	Intervention	Outcomes and instrument	Participants	Key findings
Segalowitz and Freed (2004). Context, contact, and cognition in oral fluency acquisition	Pre/post-test, quasi-experimental design. Group assignment non-random. Purposive sampling based on a set criterion with students volunteering to partake	Non-ERASMUS exchange programme L1 English learners of Spanish studying at an L1 Spanish speaking university. Intervention group stays abroad for 3 months. Control group remains in formal domestic instruction	Oral proficiency: Oral Proficiency Interview (OPI) Cognitive tasks include attention control and lexical access	40 L2 learners of Spanish. SA group (n = 22); mean age 20.68 AH group (n = 18); mean age 23.39	SA students made significant gains in five of the eight oral measures—OPI, Turn, Rate, Filler-free, and Fluent-run; students in the AH did not show significant gain in any of these measures. Interaction effect for Turn; Rate; and Filler-free suggesting that the SA changed significantly more than AH group

Attention is now given to the seven publications in this systematic review which employed a counterfactual comparison (i.e., stay-at-home) group, and in doing so allowed for stronger causal inferences to be made regarding the extent to which study abroad can account for proficiency changes witnessed.

Study characteristics

The key characteristics of the six studies (seven publications) included in the in-depth review are shown in Table 3. Of these, four (Hessel, 2016; Hessel & Vanderplank, 2018; Llanes & Muñoz, 2013; Serrano et al., 2011) investigated the research question using a European sample while three (Jochum, 2014; Li, 2014; Segalowitz & Freed, 2004) investigated the question using North American undergraduates. Four studies (Hessel, 2016; Hessel & Vanderplank, 2018; Llanes & Muñoz, 2013; Serrano et al., 2011) explored changes in L2 English; two studies investigated change (Jochum, 2014; Segalowitz & Freed, 2004) in learners of Spanish and one study (Li, 2014) focused on L2 Chinese learners. In line with the review's research question, the seven publications covered a broad range of linguistic constructs, as shown in Table 4.

Study quality

All of the included studies met a minimum standard of rigour which warranted causal inference. Most pertinently, all included studies explored linguistic ability in both a treatment group (i.e., SA group) and a comparison group (i.e., AH group), in which pre-test proficiency equivalence was met. Table 5 presents the methodological quality of each study from which the validity of conclusions was drawn. Each included study is limited by the potential of self-selection bias introduced due to the lack of randomisation to form intervention and control groups. As aforementioned, this limitation is borne out of the perception of moral and ethical issues posed by randomly allocating learners to either the SA or AH grouping. In the four European-based studies (Hessel, 2016; Hessel & Vanderplank, 2018; Llanes & Muñoz, 2013; Serrano et al., 2011), it is known that the SA group was formed by individuals who had gained a scholarship or were successful in applying for the programme. In the North American articles, however, it is unclear how participants may have differed in each group. Only one study (Hessel, 2016) explored attrition, although it is unclear how those who left were different to those who remained (e.g., in terms of pre-attainment). No study had pre-specified outcomes.

TABLE 4 In-depth studies organised by study outcome

Outcome	No.	Author(s)
Oral	4	Llanes and Muñoz (2013); Jochum (2014); Segalowitz and Freed (2004); Serrano et al. (2011)
Writing	2	Llanes and Muñoz (2013); Serrano et al. (2011)
Reading	1	Li (2014)
Grammar	2	Llanes and Muñoz (2013); Serrano et al. (2011)
Vocabulary	2	Llanes and Muñoz (2013); Serrano et al. (2011)
General proficiency	3	Hessel (2016); Hessel and Vanderplank (2018); Li (2014)

NARRATIVE SYNTHESIS

The narrative synthesis presents the findings and quality appraisal of the studies selected for the in-depth review. Due to the homogeneity of the intervention focus, studies are presented in relation to the linguistic outcome they studied.

Findings by linguistic outcome

Oral

General oral proficiency

Two studies (Jochum, 2014; Segalowitz & Freed, 2004) evaluated the impact of sojourning on general oral proficiency. Both studies operationalised oral proficiency through means of the Oral Proficiency Interview (OPI), while Segalowitz and Freed (2004) also measured the total number of words spoken (Total), duration of speech (Duration) and longest turn of spoken language (Turn) during an extract of speech. The results of both studies indicated general oral proficiency to develop as a direct effect of the sojourning experience. Jochum (2014) found the nine sojourners scored higher at post-test than the nine non-sojourners ($d = 0.56$), while there was an improvement in those who studied abroad ($d = 0.98$), but not in those who remained in domestic instruction ($d = 0.55$). The findings of this paper are limited in their reliability and validity, given the small sample. Similarly, Segalowitz and Freed (2004) found a pre/post-test difference in OPI scores for the SA group, but not for those AH. There were, however, substantial individual differences in the extent of this change. For example, in the SA group, 12 students improved, but 10 did not. The total number of words and duration of speech appeared unaffected by a period abroad. From a methodological perspective, this study is of higher quality than Jochum's (2014) study and shows promising evidence that sojourning can have a positive impact on particular oral components.

Oral fluency

Three studies evaluated whether sojourning facilitated oral fluency development (Llanes & Muñoz, 2013; Segalowitz & Freed, 2004; Serrano et al., 2011). Oral fluency is typically concerned with capturing the communicative competencies of a learner and can be measured through a multitude of constructs including speech rate, pause frequency and speech repair frequency (see Tullock & Ortega, 2018). Llanes and Muñoz (2013) and Serrano et al. (2011) did so through means of pruned syllables per minute, exploring change over one semester in L1 speaking Catalan learners of English. Segalowitz and Freed (2004) captured a number of domains of oral fluency (rate, hesitation-free, filler-free and fluent run) during one semester in 40 participants who were L2 Spanish speakers. The results of these three studies were all somewhat similar. Llanes and Muñoz (2013)[2] found SA adults became more fluent across time ($d = 0.77$), whereas no change was seen in the AH learners ($d = 0.10$). Segalowitz and Freed (2004) concluded that sojourning facilitated oral fluency development more than remaining at home. These studies were of moderate methodological quality and as such indicate promising evidence that sojourning can enhance oral fluency.

Oral accuracy

Two studies investigated the influence of SA on oral accuracy development (Llanes & Muñoz, 2013; Serrano et al., 2011), measuring accuracy by the number of errors per t-unit. Llanes and Muñoz (2013) found no learner demonstrated change over the semester (SA adults: $d = 0.15$; AH adults: $d = 0.24$).

TABLE 5 Methodological quality of studies included in the in-depth review

Studies	Grouping strategy	Sampling bias	Attrition/impact on results	Pre-specified outcomes
Hessel (2016)	Allocation dependent on learning context—students self-selected into learning context	Comparison group formed of individuals who wanted to be in treatment group but were unsuccessful in application. Prior pre-test differences considered minimising sampling bias on results	Low attrition (5.5–11.4%). Possible predictors of attrition not discussed although attrition highest in AH group. Impact to results not discussed—unclear if link between proficiency and attrition—states drop out due to 'personal reasons'. No other reasoning given	No
Hessel and Vanderplank (2018)	Allocation dependent on learning context—students self-selected into learning context	Comparison group formed of individuals who wanted to be in treatment group but were unsuccessful in application. Prior pre-test differences considered minimising sampling bias on results	Attrition not reported	No
Llanes and Muñoz (2013)	Allocation dependent on learning context—students self-selected into learning context	Authors used all available ERASMUS students to take part. These students had been successful in gaining a scholarship. Unclear whether the AH group chose to stay at home or were unsuccessful in gaining a scholarship	Attrition not reported	No
Serrano et al. (2011)	Allocation dependent on learning context—students self-selected into learning context	Authors used all available ERASMUS students to take part. These students had been successful in gaining a scholarship. Unclear whether the AH group chose to stay at home or were unsuccessful in gaining a scholarship	Attrition not reported	No
Li (2014)	Allocation dependent on learning context—students self-selected into learning context	Sample divided according to proficiency bands (e.g., intermediate SA vs. intermediate AH) no other pre-existing differences controlled for	Attrition not reported	No

(Continues)

TABLE 5 (Continued)

Studies	Grouping strategy	Sampling bias	Attrition/impact on results	Pre-specified outcomes
Jochum (2014)	Allocation dependent on learning context—students self-selected into learning context	Presence of sampling bias noted by author, but no measure or analysis undertaken to account for this presence	Attrition not reported	No
Segalowitz and Freed (2004)	Allocation dependent on learning context—students self-selected into learning context	Steps were undertaken to minimise sample bias, with participant recruitment purposive, based on a criterion. Stronger inference on the role of learning context in results	Attrition not reported	No

In evaluation of the studies previously discussed, all are limited by small sample sizes and, as such, there is a high possibility of a Type II error (i.e., accepting an effect exists, when in reality an effect does not exist) in each study. Jochum (2014) employed the smallest sample size ($n = 18$), while Llanes and Muñoz had the largest ($n = 66$). Jochum (2014) acknowledged the small size, stating the reason stemmed from a lack of uptake in the sojourning programme. The study is further limited by the delayed testing procedure. Individuals were measured up to six weeks after departure abroad, and up to six weeks before arrival home. It is as such questionable whether the scores achieved are truly representative of the ability prior to departure abroad and return home. Given such limitations, this study was evaluated to be of limited quality. Segalowitz and Freed (2004) analysed a larger sample ($n = 40$); however, their study is limited by the presence of a known confounding variable. Sojourners participated in three language courses per week compared to only one for non-sojourners. Consequently, it cannot be ruled out that the effects found were as a result of the extra formal instruction, and not due to a treatment effect. As such, this study can be considered of moderate quality. Similarly, the study by Llanes and Muñoz (2013) was considered to be of moderate quality. Serrano et al.'s (2011) study was limited by a possible practice effect because two control groups were used. Those in the treatment group (SA) undertook the post-test twice with roughly a six-week gap. As such, the results may be skewed by a task repetition effect favouring performance in sojourners. As a result, this study may be considered of limited to moderate quality.

In sum, there were relatively uniform improvements in oral fluency indicating a direct effect of sojourning on facilitating the improvement in oral fluency. Similarly, general oral proficiency as measured by the OPI indicated that those who completed a SA became more proficient than those who remained AH. For oral accuracy, the results were more mixed and inconclusive. Indeed, the included studies indicated that sojourning had little effect on development and a period of one semester may be an insufficient length of time to witness a large change.

Writing

Writing fluency

Two studies (Llanes & Muñoz, 2013; Serrano et al., 2011) evaluated the impact of sojourning on writing fluency over one semester as a result of learning context. In each, written fluency was operationalised by words per t-unit. Serrano et al. (2011) found sojourners to score higher at post-test than non-sojourners. Conversely, Llanes and Muñoz (2013) found AH adults made higher gains than SA adults. Similarly, the average adult learner in each of the learning contexts did not improve over time (SA adults: $d = 0.26$; AH adults: $d = 0.24$).

Writing accuracy

Two studies (Llanes & Muñoz, 2013; Serrano et al., 2011) also investigated the role of learning context in facilitating writing accuracy. Both studies found no evidence to suggest that a sojourn of three months had a direct effect on developing written accuracy. In each study, sojourners were not more proficient in this domain at post-test than those who stayed at home.

In sum, there is evidence to indicate that sojourning can have a facilitative effect on writing fluency, but not writing accuracy. The methodological quality of both studies was considered as moderate but further evidence is required to affirmatively conclude the extent to which sojourning benefits writing development.

Reading

One study evaluated the impact of sojourning on the development of reading ability. Li (2014) explored changes in reading ability over eight weeks in 73 L2 learners of Chinese who were split into three groups according to proficiency. While the sample sizes were fairly balanced across the three proficiency bands, they were small in scale ranging from 9 to 15 students. Reading ability was captured using a 10-item multiple-choice reading comprehension measure. Sojourners scored higher at post-test on average than non-sojourners. Further simple main effect analysis was conducted on the proficiency banding, finding only intermediate sojourners to score higher than their non-sojourning peers at post-test. However, this study can be considered of limited quality due to the instrument used. Ceiling effects were present in each learning context and proficiency band at pre-test, with pre-test scores ranging between 9.12 and 9.51 out of 10. Such scores indicate that the reading measure may not have been sensitive enough to capture actual reading ability and suggests the measure was too easy for learners. In light of this finding, we cannot have confidence in this finding and more robust evidence is required to ascertain whether the SA context facilitates reading development more than remaining AH.

Vocabulary

Two included studies (Llanes & Muñoz, 2013; Serrano et al., 2011) explored whether sojourning aids vocabulary development as operationalised by Giraud's Index of Lexical Richness. This measure is designed to capture and quantify the extent to which a user is using a varied and large vocabulary (Laufer & Nation, 1995). Both studies explored lexical complexity in both learners' oral and written compositions.

Regarding oral lexical complexity, the findings were mixed. Serrano et al. (2011) found sojourners to score higher at post-test than non-sojourners. Although Llanes and Muñoz (2013) demonstrated no adult sufficiently improved over time in either learning context (SA adults: $d = 0.35$; AH adults: $d = 0.31$), the analysis did indicate that SA adults made greater gains than AH adults. Concerning lexical complexity in writing, Serrano et al. (2011) found an effect between learning context and time. Sojourners were found to obtain higher scores in written lexical complexity than non-sojourners at post-test. Conversely, Llanes and Muñoz (2013) found AH adults to score higher at post-test ($d = 0.89$) while SA adults displayed a moderate change ($d = 0.39$).

In evaluation of both studies, the evidence may be considered of moderate strength, suggesting promising evidence that sojourning can foster positive vocabulary growth. Nevertheless, there is a reliance on one measure of vocabulary, namely that of lexical richness, and such an instrument may under- or over-represent the construct of vocabulary.

Syntactic complexity

Two studies (Llanes & Muñoz, 2013; Serrano et al., 2011) investigated changes in syntactic complexity as result of a SA experience compared to AH. Syntactic complexity refers to the range and sophistication of syntactic structure employed in speech or writing, and each study was measured by clauses per t-unit. Concerning the domain of speaking, both studies found effects indicating sojourning to have little benefit on oral syntactic complexity. Regarding writing, Serrano et al. (2011) found an effect at post-test. Similarly, Llanes and Muñoz (2013) indicated that the average SA and AH learner did not show an improvement over time (SA adults: $d = 0.01$; AH adults: $d = 0.17$).

On the whole, the available evidence points towards sojourning having little facilitative effect on the development of syntactic complexity in both oral and writing domains. Similar to that of vocabulary, criticism may be levied at having only one measurement of syntactic complexity (i.e., clauses per t-unit), and consequently may be an imprecise measurement of syntactic complexity, given the multitude of ways the outcome can be operationalised.

General proficiency

Two studies (three articles) investigated the role of sojourning on the development of general proficiency (Hessel, 2016; Hessel & Vanderplank, 2018; Li, 2014). Two of these studies operationalised general proficiency through means of a C-test while Li (2014) used a specifically designed proficiency tool comprising five sections including listening, grammar and translation. Hessel (2016) and Hessel and Vanderplank (2018) explored change in three groups. The first represented short-term sojourners who spent one semester abroad ($n = 44$); the second represented long-term sojourners, who spent the academic year abroad ($n = 52$), while the third group represented the stay-at-home control group ($n = 40$). The results of Hessel (2016) and Hessel and Vanderplank (2018) indicated that during the first semester, both short-term ($d = 0.60$) and long-term sojourners ($d = 0.55$) made gains over the first three months. For AH students an improvement was also found ($d = 0.12$). When comparing learning contexts, the regression analysis showed learning context to be a significant predictor of proficiency gain, in which short-term sojourners improved by 6.50 marks more than those AH, while long-term sojourners improved by 6.24 marks over those AH on average. When exploring the subsequent six months, Hessel (2016) and Hessel and Vanderplank (2018) found long-term sojourners to improve ($d = 0.45$), together with non-sojourners ($d = 0.17$). When comparing gain scores, the regression analysis indicated that learning context no longer predicted L2 proficiency development. Inferring causality from the study is strengthened because the authors controlled for a range of possible confounding variables in their regression analysis, including pre-test proficiency and learner characteristics. Regarding Li (2014), sojourners, on the whole, scored higher at post-test compared to non-sojourners. When simple main effects were run regarding proficiency banding, a difference was found for intermediate and advanced learners, but not for beginners. In evaluation of the aforementioned studies, Hessel (2016) and Hessel and Vanderplank (2018) studies were judged to be of high quality as they controlled for a number of confounding variables, including pre-test proficiency enabling for stronger causal inference. While the sample size was not large enough to detect small effects, a limitation noted by Hessel (2016), the sample was the largest of the included studies, with participants from a range of institutions across Germany, increasing the generalisability of findings. Li (2014), on the other hand, described the sample as originating from a *prestigious* institution which, despite the ambiguity, minimised the generalisability of findings to other, less *prestigious* institutions.

From a methodological perspective, the domain of general proficiency has been most robustly evaluated and evidenced, given the study designs implemented. Given this, the strongest causal inference may be made pertaining to the notion that SA does enhance general proficiency.

DISCUSSION

This review has evaluated the efficacy of SA on facilitating L2 acquisition in undergraduate language learners compared to remaining at home. The review identified six studies, all of which were QEDs, and demonstrated L2 proficiency pre-test equivalence. This low number

is perhaps indicative of the 'poor state of affairs in quantitative SA research' (p. 3) as described by Sanz and Morales-Front (2018) and is in line with the limited number of studies ($n = 11$) found by Yang (2016).

The review is, to the best of our knowledge, the first to employ the SWiM guidelines in educational research, and the first systematic synthesis to be conducted in the field of SLA. Typically, scholars have previously undertaken narrative reviews which lack the rigour and limitation of bias in a systematic review design. This systematic review uses transparent and replicable methods. Author bias can impact both the selection of studies included (e.g., only those known to the authors), and preconceived knowledge can bias the interpretation of the results presented in the review (Mallett et al., 2012).

Linguistic aspect under study

This review had indicated that the most reliable evidence of causal claim concerns the SA facilitating growth in oral fluency and general proficiency given the moderate-to-high quality studies included in these domains. Regarding oral fluency, the significant gains in ability would indicate that learners, at least in the included studies, are afforded opportunities to converse in the L2 and that learners return home more confident in their speech as a result of living in the host community. The observed gain in general proficiency may be less conducive to clear explanation but does indicate that sojourning facilitates skill acquisition. In line with the Skill Acquisition Theory, closely aligned to the Adaptive Control of Thought model (Anderson, 1982), it is posited that an experience abroad facilitates the rate at which procedural knowledge becomes automatised. This process occurs through repeated practice and may once again reflect the notion that language learners should be afforded more opportunities to practise the L2 than those in a classroom setting.

Conversely, the findings indicated sojourning to have little benefit on the development of writing, reading and grammar. Reasons for this may be manifold, including the amount of time spent practising the relevant skill, non-correction of incorrect linguistic forms by native L2 speakers leading to fossilised errors and influence of dialect on the acquisition of requiring non-standard linguistic forms (DeKeyser, 2007), although no such claims are suggested by this review.

The findings discussed here corroborate much of the available literature, both primary and secondary (e.g., Llanes, 2011; Mitchell et al., 2017). The originality of this review comes in the methodological approaches undertaken; the review has challenged and corroborated previous findings using an open, transparent and systematic design.

The European perspective

This review is timely and relevant, for it reflects the burgeoning literature that has focused on a European sample in the last two decades. European studies have focused on a range of linguistic outcomes and have done so by employing a range of study designs. The findings are in line with those studies that have employed a counterfactual comparison group, in suggesting that sojourning within Europe is particularly beneficial in developing oral fluency and general proficiency. Although the prevalence of American-based literature has often been considered an issue for European readers, these findings suggest that the broader literature base may be generalisable to the European reader and should not be overlooked.

The significance of this review to practitioners and learners

Both empirical and anecdotal evidence has consistently shown learners to be disappointed and disillusioned by the *perceived* lack of development in their linguistic abilities and skills, suggesting that at pre-departure, learners' expectations are too high (Mendelson, 2004). This review can be used by practitioners, for example in pre-departure workshops, to inform learners on how they may expect their proficiency to change. It can help move practitioners away from using terms such as 'fluent' or 'native speaker' to describe progress, terms which may only inflate learners' expectations. From a learner's perspective, this review can hopefully dispel the general assumption that SA delivers linguistic growth by default (Hessel, 2016). Developing realistic expectations in learners is important because they can be a determining factor in how the experience is perceived by the learner (Wilkinson, 1998). As such, this review can facilitate these for both learner and practitioner alike.

The significance of this review to policy

Policy should be informed by evidence-based decision making. This review demonstrates the need for policy to support students in completing a SA for it can have demonstrable impact on their language proficiency development. The introduction of the Alan Turing scheme by the UK government and the continuation of ERASMUS+ is welcomed and suggests that policy makers continue to see the benefit sojourning can have on learners.

Moreover, the in-depth review has indicated the potential of short-term sojourning (one semester) on linguistic proficiency, particularly that of oral fluency (Llanes & Muñoz, 2013; Segalowitz & Freed, 2004; Serrano et al., 2011). A perceived barrier to SA participation has been the requirement to extend one's academic studies, subsequently delaying graduation, and increasing costs associated with study (Kim & Goldstein, 2005). Consequently, if the goal of the learner is to become communicatively competent, policy makers may well be advised to be flexible in their approach of structuring SA programmes.

Strengths and limitations

A strength of this review is that it has aimed to tackle a broad and diverse field in a systematic way and in doing so has included studies of both European and non-European populations. As a result, this review has attempted to tackle the issues regarding the heterogeneity of studies selected in past reviews, resulting in at times contradictory findings. Nonetheless, the review is limited by the overall small samples found in the studies included, ranging from 18 (Jochum, 2014) to 143 (Hessel, 2016). The observed small samples have also been criticised in other works (e.g., Sanz & Morales-Front, 2018). Moreover, the lack of any extant randomised controlled trial (RCT) to include in the review means that ultimately, determining strong causal effect from the review is extremely difficult, and the results should be used as a general guideline when exploring whether study abroad facilitates linguistic gain. A further potential limitation is the specific focus given to research that has focused on between-group differences. While it is was not feasible to discuss the notion of within-person differences or the factors which may contribute to accounting for individual differences in gain seen by learners abroad, it should be acknowledged that past literature (e.g., Baker-Smemoe et al., 2014) has explored this, finding that variables (e.g., cultural sensitivity and social networking) could influence the extent to which an individual improves while abroad.

CONCLUSIONS

To conclude, this systematic review has demonstrated that sojourning can have a positive impact on linguistic ability and should, within reason, continue to be seen as a catalyst for language development. However, it must be noted that the extent to which sojourning benefits language development appears dependent on linguistic outcome. Given the methodological quality of the included studies, the domains of oral fluency and general proficiency demonstrated the largest gains which could be directly attributed to the learning context. Given the paucity of studies that met the full inclusion criteria, the generalisability of our conclusions is undoubtedly weakened but provides further insights into the impact of SA on language change. As aforementioned in the introduction, for many, the goal of SA is to improve their linguistic competencies, with this review demonstrating that a short stay (~1 semester) is sufficient if this goal relates to the domain of fluency. In sum, the review has systematically evaluated the existing literature in relation to the extent to which sojourning facilitates language acquisition, both in Europe and beyond. In doing so, it is hoped that this review can better inform practitioners and students alike to set realistic expectations of linguistic change abroad and continue to question whether SA programmes achieve what they ultimately set out to achieve.

CONFLICT OF INTEREST

The authors of this paper either work or have attended the institution of the journal editorial board.

ETHICAL APPROVAL

Ethics approval was obtained by the Department for Education at Durham University and adheres to the 2018 BERA Ethical Guidelines.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ENDNOTE

[1] Effect sizes given are taken from the original publication.

[2] Effect size was calculated by corresponding author.

REFERENCES

*These are studies which were included in the in-depth review.

Anderson, J. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–406. https://doi.org/10.1037/0033-295X.89.4.369

Avello, P. (2014). *L2 phonological development in speech production during study abroad* (unpublished doctoral dissertation). Universitat Pompeu Fabra.

Avello, P., Mora, J., & Pérez-Vidal, C. (2012). Perception of FA by non-native listeners in a study abroad context. *Research in Language*, *10*(1), 63–78. https://doi.org/10.2478/v10015-011-0050-9

Bädstuber, T., & Ecke, P. (2009). Student expectations, motivations, target language use, and perceived learning progress in a summer study abroad program in Germany. *Teaching German*, *42*(1), 41–49.

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Variables affecting L2 gains during study abroad. *Foreign Language Annals*, *47*(3), 464–486. https://doi.org/10.1111/flan.12093

Barron, A. (2019). Using corpus-linguistic methods to track longitudinal development: Routine apologies in the study abroad context. *Journal of Pragmatics*, *146*, 87–105. https://doi.org/10.1016/j.pragma.2018.08.015

Beattie, J., Valls-Ferrer, M., & Pérez-Vidal, C. (2014). Listening performance and onset level in formal instruction and study abroad. In C. Pérez-Vidal (Ed.), *Language acquisition in study abroad and formal instruction contexts* (pp. 87–110). John Benjamins.

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, *368*, 1–6. https://doi.org/10.1136/bmj.l6890

Coleman, J. A. (1998). Language learning and study abroad: The European perspective. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *4*(2), 167–203. https://doi.org/10.36366/frontiers.v4i1.67

Coleman, J. A. (2015). Social circles during residence abroad: What students do, and who with. *Social Interaction, Identity and Language Learning during Residence Abroad*, *4*, 33–52.

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

Edmonds, A., & Gudmestad, A. (2018). Gender marking in written L2 French: Before, during, and after residence abroad. *Study Abroad Research in Second Language Acquisition and International Education*, *3*, 59–82. https://doi.org/10.1075/sar.16018.edm

Field, A. (2013). *Discovering statistics using SPSS* (4th ed.). SAGE Publications.

Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). John Benjamins.

Grey, S. (2018). Quantitative approaches for study abroad research. In C. Sanz, & A. Morales-Front (Eds.), *The Routledge Handbook of Study Abroad Research and Practice* (pp. 48–57). New York: Routledge.

*Hessel, G. (2016). *The impact of participation in ERASMUS study abroad in the UK on students" overall English language proficiency, self-efficacy, English use anxiety and self-motivation to continue learning English: A mixed methods investigation* (unpublished doctoral dissertation). Oxford University.

Hessel, G. (2017). A new take on individual differences in L2 proficiency gain during study abroad. *System*, *66*, 39–55. https://doi.org/10.1016/j.system.2017.03.004

*Hessel, G., & Vanderplank, R. (2018). What difference does it make?: Examining English proficiency gain as an outcome of participation in ERASMUS study abroad programmes in the UK. *Study Abroad Research in Second Language Acquisition and International Education*, *3*(2), 191–219. https://doi.org/10.1075/sar.16020.hes

Howard, M. (2005). Second Language Acquisition in a study abroad context: A comparative investigation of the effects of study abroad and foreign language instruction on the L2 learner's grammatical development. Investigations in instructed Second Language Acquisition. *Studies on Language Acquisition*, *25*, 495–530.

Howard, M. (2008). Morphosyntactic development in the expression of modality: The subjunctive in French L2 acquisition. *Canadian Journal of Applied Linguistics*, *11*(3), 171–192.

Ife, A., Vives Boix, G., & Meara, P. (2000). The impact of study abroad on the vocabulary development of different proficiency groups. *Spanish Applied Linguistics*, *4*(1), 55–84.

*Jochum, C. J. (2014). Measuring the effects of a semester abroad on students' oral proficiency gains: A comparison of at-home and study abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *24*, 93–104. https://doi.org/10.36366/frontiers.v24i1.338

Juan-Garau, M. (2014). Oral accuracy growth after formal instruction and study abroad. In C. Pérez-Vidal (Ed.), *Second language acquisition in study abroad and formal instruction contexts* (pp. 87–110). John Benjamins.

Kim, R. I., & Goldstein, S. B. (2005). Intercultural attitudes predict favorable study abroad expectations of U.S. college students. *Journal of Studies in International Education*, *9*(3), 265–278. https://doi.org/10.1177/1028315305277684

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

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

*Li, L. (2014). Language proficiency, reading development, and learning context. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *24*, 73–92. https://doi.org/10.36366/frontiers.v24i1.337

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*(3), 189–215. https://doi.org/10.1080/14790718.2010.550297

*Llanes, À., & Muñoz, C. (2013). Age effects in a study abroad context: Children and adults studying abroad and at home. *Language Learning*, *63*(1), 63–90. https://doi.org/10.1111/j.1467-9922.2012.00731.x

Llanes, À., Tragant, E., & Serrano, R. (2012). The role of individual differences in a study abroad experience: The case of Erasmus students. *International Journal of Multilingualism*, *9*(3), 318–342. https://doi.org/10.1080/14790718.2011.620614

Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness: Special Issue on Systematic Reviews*, *4*(3), 445–455. https://doi.org/10.1080/19439342.2012.711342

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2020). A longitudinal study of advanced learners' linguistic development before, during, and after study abroad. *Applied Linguistics*, *42*(1), 136–163. https://doi.org/10.1093/applin/amaa003

Mendelson, V. G. (2004). "Hindsight is 20/20:" Student perceptions of language learning and the study abroad experience. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *10*, 43–63. https://doi.org/10.36366/frontiers.v10i1.132

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL. Institut Voor Togepaste Linguistik*, *107-108*, 17–34. https://doi.org/10.1075/itl.107-108.02mil

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone students abroad: Identity, social relationships, and language learning*. Routledge.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097.

Pérez-Vidal, C. (2014). Study abroad and formal instruction contrasted. In C. Pérez-Vidal (Ed.), *Second language acquisition in study abroad and formal instruction contexts* (pp. 17–58). John Benjamins.

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

Rees, J., & Klapper, J. (2007). Analysing and evaluating the linguistic benefit of residence abroad for UK foreign language students. *Assessment and Evaluation in Higher Education*, *32*(3), 331–353. https://doi.org/10.1080/02602930600801860

Regan, V. (1995). The acquisition of sociolinguistic native speech norms. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 245–267). John Benjamins.

Rehner, K., & Mougeon, R. (2003). The effect of educational input on the development of sociolinguistic competence by French immersion students: The case of expressions of consequence in spoken French. *Journal of Educational Thought*, *37*(3), 259–281.

Sanz, C., & Morales-Front, A. (2018). Introduction: Issues in study abroad research and practice. In C. Sanz, & A. Morales-Front (Eds.), *The Routledge handbook of study abroad and practice* (pp. 1–17). Routledge.

*Segalowitz, N., & Freed, B. (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.

*Serrano, R., Llanes, A., & 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*(2), 133–143.

Serrano, R., Tragant, E., & Llanes, À. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review*, *68*(2), 138–163. https://doi.org/10.3138/cmlr.68.2.138

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, *349*, g7647.

Thomas, J., Brunton, J., & Graziosi, S. (2010). *EPPI-Reviewer 4: Software for research synthesis. EPPI-Centre Software*. Social Science Research Unit, UCL Institute of Education.

Tullock, B., & Ortega, L. (2018). Fluency and multilingualism in study abroad: Lessons from a scoping review. *System*, *71*, 7–21. https://doi.org/10.1016/j.system.2017.09.019

Valls-Ferrer, M., & Mora, J. C. (2014). The role of onset level on L2 perceptual phonological development after formal instruction and study abroad. In C. Pérez-Vidal (Ed.), *Second language acquisition in study abroad and formal instruction contexts* (pp. 111–136). Amsterdam: John Benjamins.

Wilkinson, S. (1998). Study abroad from the participants' perspective: A challenge to common beliefs 1. *Foreign Language Annals*, *31*(1), 23–39.

Yang, J. S. (2016). The effectiveness of study-abroad on second language learning: A meta-analysis. *Canadian Modern Language Review*, *72*(1), 66–94. https://doi.org/10.3138/cmlr.2344

SUPPORTING INFORMATION

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

How to cite this article: Moore, I., Torgerson, C., & Beckmann, N. (2021). Systematic review measuring the efficacy of study abroad in undergraduate language learners on linguistic proficiency gains. *Review of Education*, 9, e3306. https://doi.org/10.1002/rev3.3306