1. Background

The EEF has asked the team at Oxford to conduct a rapid evidence assessment (REA) focusing on the evidence on the effects of Modern Foreign Languages (MFL) instruction on wider academic outcomes. In addition to this overarching aim, The EEF has asked the team to gather evidence about effective MFL teaching more generally.

The motives for teaching foreign languages across the globe are diverse; from equipping young people to compete for jobs in a globalised world, through to developing tolerance and understanding of other cultures. But does learning a foreign language also have an impact on wider academic outcomes, and if so, what approaches are most likely to lead to improved learning outcomes in other subjects, as well as achieving proficiency in the given language?

In recent years there has also been an increase in the number of countries providing academic instruction in non-native languages. This is particularly prevalent in non-anglophone countries where there has been a shift from English being taught as a foreign language to English becoming the medium of instruction (EMI) for a number of academic subjects. However, there is limited evidence on the impact of this policy on wider academic outcomes, or evidence on effective implementation. For example, do factors such as the age of introduction, a teacher’s level of proficiency in the language of instruction, or the specific subject in which EMI is deployed, have an impact on student outcomes?

In parallel, countries with high levels of migration have been implementing a range of strategies designed to promote social inclusion and integrate migrant children into the local education system who, due to age or other circumstances, such as having spent their early years in conflict zones with limited access to educational opportunities, may have limited literacy in their home language and no knowledge of the local language. There is evidence that suggests that developing literacy in the child’s home language, before introducing the local language, may have significant benefits for a child’s long-term educational achievement. Are there interventions and approaches that could be implemented by communities and educators at the local level, and supported by government, aimed at strengthening students’ home-language literacy and proficiency as, or before, they enter mainstream education?

The EEF’s overall aim for this project is to understand the impact of foreign language learning and the most effective strategies to achieve language proficiency and positive impact on wider academic attainment. This evidence will be used to inform policy and practice, with the possibility of providing the foundation for further primary research.
2. Objectives

The main objective of this evidence assessment is to understand what is known from the literature about learning a foreign language and its impact on students’ wider academic outcomes.

The specific review objectives are to summarise the evidence on:

- how to effectively teach a foreign language;
- the effect of learning a foreign language on attainment in other academic subjects;
- the effect on second language acquisition and on academic attainment of using a non-native language as the medium of instruction in academic subjects;

and

- to provide practical recommendations on:
  o how to best teach a foreign language
  o how to best teach a foreign language to maximise benefits on wider academic outcomes
  o when and how to introduce a non-native language as the medium of instruction

2.1 Review questions

Primary questions

- What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?
- What is the impact of learning a foreign language on students’ wider academic outcomes?

Secondary questions

- What practitioner skills or programme characteristics contribute to effective language learning among students?
- What is the impact of using a non-native language as the medium of instruction in academic subjects on students’ academic outcomes?
- Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?
- What is the impact of delaying or accelerating the introduction of a new ‘local’ language as a medium of instruction for new arrivals (e.g., refugees, immigrants) who are not yet proficient in their native language?

These aims and objectives are extremely broad, and given the short timescale allotted to complete the REA the team will be relying on previously conducted systematic reviews to provide a trustworthy basis upon which to make the recommendations it generates for the EEF. In brief, the team will locate, appraise and synthesise the findings of published
3. Methods

3.1 Overview

Phase One will consist of a wide-reaching trawl for systematic reviews that address, or may address, the review questions. These will be assessed for relevance and quality. The most relevant and highest quality reviews will become ‘seed’ reviews that form the basis of the REA.

Phase Two will consist of updating the seed reviews by replicating their methods; limiting inclusion to papers published after the original search and to only RCTs and QEDs with a control group and pre- and post-tests. Any new studies meeting these inclusion criteria will be incorporated into the findings of the original reviews. Finally, the findings of this process will be organized thematically across reviews in a narrative synthesis.

Statistical synthesis (meta-analysis) to augment the narrative synthesis will only be conducted when studies of sufficient similarity in terms of intervention, comparator and outcome measure exist in sufficient number across reviews to make this kind of synthesis meaningful, and when data are available to allow these calculations to be made.

3.2 Phase 1 – Locating and appraising existing research syntheses on the topics of the REA

Phase 1 will entail a wide-reaching trawl for systematic reviews that focus on the effects of MFL instruction. The purpose of this trawl will be to identify as many systematic reviews as possible that address, or may address, the substantive questions driving the REA. This search will therefore be sensitive rather than precise (Higgins & Green 2011). That is, it is likely to return many irrelevant reviews as a proportion of the whole, but as a function of this, the strategy helps to ensure that fewer relevant reviews are missed. Reviews returned by the trawl will be assessed for relevance and quality. The relevance of each review to the review questions will be assessed against the inclusion/exclusion criteria itemized in section 3.2.2. The quality of each review will be appraised using an adapted version of the CASP Checklist (CASP-UK 2019). The highest quality reviews that are also most relevant to the aims and objectives of the REA will be selected for inclusion in. Details of the selection process are itemized in section 3.2.4.

This trawl will consist of two main approaches: 1) an electronic search of bibliographic databases and 2) asking colleagues, members of professional language teaching networks, and learned associations to tell us of any systematic reviews which might be eligible for consideration in the REA.

The online search will be conducted using the following bibliographic databases:

- Web of Science
- Education Collection (including ERIC)
Several searches will be conducted, with slightly different search terms applied in each, to account for the different review questions driving the REA.

These are:

**RQ1. What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?**

**RQ3. What practitioner skills or programme characteristics contribute to effective language learning among students?**

<table>
<thead>
<tr>
<th>Type of publication:</th>
<th>systematic OR meta-analysis OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR rapid evidence assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context:</strong></td>
<td>second language* OR foreign language* OR modern language* OR additional language* OR L2 OR MFL</td>
</tr>
<tr>
<td>AND</td>
<td>teach* OR learn* OR instruc* OR pedagog* OR acqui* OR train* OR study* OR educat* OR intervention*</td>
</tr>
<tr>
<td>AND</td>
<td>effect* OR outcome* OR improv* OR develop* OR attain* OR gain* OR increas* OR grow* OR succe* OR competenc* OR develop* OR scor* OR grad* OR result*</td>
</tr>
</tbody>
</table>

**RQ2. What is the impact of learning a foreign language on students’ wider academic outcomes?**

<table>
<thead>
<tr>
<th>Type of publication:</th>
<th>systematic OR meta-analysis OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR rapid evidence assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context:</strong></td>
<td>second language* OR foreign language* OR modern language* OR additional language* OR L2 OR MFL</td>
</tr>
<tr>
<td>AND</td>
<td>teach* OR learn* OR instruc* OR pedagog* OR acqui* OR train* OR study* OR educat*</td>
</tr>
<tr>
<td>AND</td>
<td>effect* OR affect* OR impact* OR influence* OR improv* OR promot* OR benefit*</td>
</tr>
<tr>
<td>AND</td>
<td>academic OR achiev* OR outcome* OR skill* OR literac* OR read* OR competenc* OR metacognit* OR meta-cognit* OR metalinguistic OR meta-linguistic OR problem-solv* OR develop* OR scor* OR grad* OR result* OR subject* OR motivat* OR exam* OR set* OR band* OR GCSE* OR GRE* OR level* OR SAT* OR baccalaureate</td>
</tr>
</tbody>
</table>
RQ4. What is the impact of using a non-native language as the medium of instruction in academic subjects on students’ academic outcomes?

RQ5. Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?

<table>
<thead>
<tr>
<th>Type of publication:</th>
<th>systematic OR meta-analysis OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR rapid evidence assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context:</td>
<td>EMI OR CLIL OR medium instruct* OR medium educat* OR content and language OR content-and-language OR content-based OR bilingual educat* OR bilingual program* OR bilingual programme* OR language immersion OR dual language OR dual-language OR content-based education OR content-based education OR L2MI</td>
</tr>
<tr>
<td>Outcomes:</td>
<td>effect* OR outcome* OR achiev* OR improv* OR develop* OR attain* OR gain* OR increas* OR grow* OR succe* OR competenc*</td>
</tr>
</tbody>
</table>

RQ6. What is the impact of delaying or accelerating the introduction of a new ‘local’ language as a medium of instruction for new arrivals (e.g. refugees, immigrants) who are not yet proficient in their native language?

<table>
<thead>
<tr>
<th>Type of publication:</th>
<th>systematic OR meta-analysis OR state-of-the-art OR review OR analy* OR survey* OR synthesi* OR rapid evidence assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context:</td>
<td>additional language OR EAL OR English language learner* OR new arrival* OR refugee* OR immigrant* OR minority-language* OR language-minorit* OR language minorit* OR linguistic minorit* OR multicultural OR mother tongue* OR native language* OR heritage language*</td>
</tr>
<tr>
<td>Outcome:</td>
<td>effect* OR affect* OR impact* OR influence* OR improv* OR promot* OR benefit*</td>
</tr>
</tbody>
</table>

In all cases, the search terms in our category of type of publication will be limited to appearances in the document title only. All other search terms will be limited to appearances anywhere but in the main text (i.e. in title, abstract, tags, and so on). Searches will be limited to documents published after 1999.

In addition to the electronic search, we will email members of professional networks in the field of language education via their membership listserves (or, in the case of colleagues known personally to us, via their personal email addresses) asking them to provide information about any systematic reviews (or other types of research syntheses) addressing the focus areas of the REA. The text of our email can be seen in Appendix 1.
This email will be sent to the membership listserves of:
- The Research in Primary Languages (RIPL) Network http://www.ripl.uk/network/
- The British Association of Applied Linguistics (BAAL) https://www.baal.org.uk
- The Association Internationale de Linguistique Appliquée (AILA) https://aila.info
- The National Centre for Excellence for Language Pedagogy (NCELP) https://ncelp.org/

Finally, the online libraries of the Campbell Collaboration (campbellcollaboration.org) and EPPI-Centre (eppi.ioe.ac.uk) - both commissioners and publishers of systematic reviews in education - will be hand-searched for potentially eligible reviews.

The bibliographic information returned by these searches will be entered into a shared Excel spreadsheet, at which point any duplicate entries will be removed.

3.2.1 Longlisting

On the basis of the information contained in their abstracts, structured summaries or executive summaries, reviews will be assessed for relevance against the inclusion/exclusion criteria below. Any review that cannot be excluded on the basis of the information contained therein will be forwarded for full text screening.

3.2.2 Inclusion/exclusion criteria

Include 1: The document is a systematic review.
Exclude 1: The document is not a systematic review or does not refer to systematic methods used to locate and appraise the literature included in the review. 
Rationale: Reviews in which the methods adopted for locating and appraising the literature included in the review are not reported cannot be updated on their own terms because those terms are not stated.

Include 2: The review has relevance either (a) to teaching FL in school settings, i.e. it is engaged with pedagogy and classroom practice, and the population of interest is school-aged (4-18 years), and it has relevance to teaching non-dominant languages or (b) to settings where the second language (L2) of the learners is used as the medium of instruction.
Exclude 2: The review does not have relevance either (a) to teaching FL in school settings, i.e. it is wholly or mainly focused on, for example, incidental language learning, study abroad, or teaching the language of the majority in the context in which the individual studies were conducted, or (b) to L2 medium of instruction. Or it has nothing at all to do with FL teaching.

Include 3: The review reports on the (relative) effectiveness of approaches to teaching foreign languages.
Exclude 3: The review does not report on the (relative) effectiveness of alternative approaches to teaching foreign languages. e.g. it is a review of: theory, research methodology, perceptions about or attitudes towards foreign language instruction, etc.
Include 4: The review reports on substantive educational outcomes such as test scores or exam pass rates, including if these are not directly related to FL outcomes, such as attainment in other academic subjects.

Exclude 4: The review reports only on non-educational outcomes such as student satisfaction, or is purely descriptive.

Two members of the review team will independently screen each abstract and, on the basis of the above criteria, decide either to exclude the review or to forward it for full text screening. Decisions between pairs of reviewers will be checked for consistency. Any disagreements will be resolved through discussion. If discussion between the two reviewers does not resolve the disagreement, a third reviewer will mediate.

It is expected that not all reviews will address all of the review questions driving this REA, and that some reviews will extend their reach beyond just these questions. Therefore, each review marked for inclusion will be tagged with the review question or questions it helps to address.

Following this first round of screening, the full texts of reviews that could not be excluded on the basis of the information contained in their abstracts will be screened against the same inclusion/exclusion criteria. The full texts of each review will be screened by two reviewers. Reviews that meet all of the inclusion criteria at this stage will be moved into the shortlisting process. Any disagreements between reviewers will be resolved through discussion. If discussion between the two reviewers does not resolve the disagreement, a third reviewer will mediate.

3.2.3 Shortlisting

Reviews not excluded at the longlisting stage will be assessed for quality against each item on an adapted version of the CASP checklist (Appendix 2). These will also be screened by two members of the review team. First, the decision to include taken at the longlisting stage will be confirmed by reading the full text against the inclusion/exclusion criteria above. Any reviews that, on closer inspection of the full text, fail to meet any of the four inclusion criteria above will be excluded at this point.

Once confirmed as relevant, the quality of each review will be assessed using the CASP checklist (CASP-UK 2019). The CASP Checklist helps reviewers assesses the quality of the review in general methodological terms. It includes judging the appropriateness of the methods used in the review, the likelihood that the review was exhaustive relative to its stated aims, the quality of the literature informing the review (including the risk of bias or weight of evidence assessments of individual studies), and the quality and appropriateness of the synthesis. For each item on the checklist reviewers will assign a score of between 0 and 5, 0 being ‘does not meet minimum expectations for this item’ and 5 being ‘meets expectations to the highest level for this item’. Judgements for each item on the CASP Checklist for each review will be recorded in an Excel spreadsheet shared amongst the team. Team members will also make an overall judgment for each review they have assessed, based on the focus of the review and its quality as assessed using CASP, indicating whether it is of sufficient relevance and of high enough quality to be considered for
inclusion in the REA. This too will be reported using a five-point scale. This process will generate a hierarchy of quality and appropriateness of all potential seed reviews.

### 3.2.4 Selecting reviews to form the bases of the REA

Reviews that have been assessed as both relevant and minimally methodologically sound will be considered together through discussion among the team, with the objective of selecting the highest quality and most relevant of these reviews for inclusion as the bases for the REA. We have termed these ‘seed reviews’.

Selection will be informed in the first instance by the thematic focuses of each review. That is, all reviews addressing Review Question 1 will be considered as one corpus. All reviews addressing Review Question 2 will be considered as another corpus, and so on. Within each corpus we will take into account how closely each review addresses the REA’s review questions, and what score was obtained during the quality assessment. We will adopt a “best evidence” (Slavin 1986) stance with regard to quality. That is, there will be no cut-off score, beneath which reviews will not be considered for inclusion. Rather, the best available evidence in each corpus will be selected. To illustrate, if two reviews address the same or similar themes, yet one has been assessed as 5 and the other has been assessed as 4, then the review scoring 5 will be preferred over the review scoring 4. If, however, the highest quality review addressing a particular theme scores 3, it will be selected as a seed review and its methodological shortcomings will be clearly described in the write up. Should more than one review addressing the same theme(s) score equally on the quality assessment measure, the more recent of the reviews will be preferred.

From each of these corpora we will select the review or reviews which, individually or together, address our review questions most comprehensively. The scope of the review questions for this REA is very broad. It is possible that some reviews will be as broad in their scope, but equally that some reviews will be tightly focused on just one element within that scope. The discussion among the team will determine whether more than one seed review need to be combined to adequately address each review question, or whether a single review is sufficient. The final decision will be made on the basis of the expert judgement of the team.

### 3.3 Phase 2 - Working from the seed reviews to prepare a synthesis of updated syntheses

#### 3.3.1 Overview

Seed reviews will be distributed among the team, and pairs of reviewers will work together to update each one. On completion of the update, findings from all updated seed reviews will be combined and presented thematically. For example, in respect of the first primary review question (What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?), evaluations of alternative ways to teach vocabulary will be presented together, evaluations of alternative ways to teach reading will be presented together, and so on.
3.3.2 Updating seed reviews

The review team will be divided into pairs with each pair taking responsibility for updating at least one of the seed reviews, or one set of seed reviews should the decision be taken to update multiple reviews addressing the same review question. The seed reviews will be updated on the basis of the methodology used in the original. That is, wherever possible, the same search terms and same search engines/bibliographic databases will be used to locate primary research published since the date of the original search (where stated) or the date of publication of the seed review (where the search date is not stated). The same inclusion/exclusion criteria as used in the seed reviews will be applied, with the following exceptions:

- Only electronic databases will be searched – no handsearching will be conducted, as the time constraints on this REA are prohibitively short to allow this to happen.
- Only RCTs and QEDs are considered for inclusion.
- Only studies published since the original search/publication date of the seed review will be considered for inclusion.
- For older reviews, the review team will consider whether additional search terms should be added to reflect changes in terminology since their publication. For example, reviews published in or before 1994 will not include the term CLIL (content and language integrated learning) as it was not coined until 1995.

3.3.3 Data management

The updated search for each seed review will be conducted by the team’s research assistant, who will apply the search terms to the relevant databases, on a review by review basis. Then, the returned records will be uploaded to Rayyan, a web-based application for collaborative abstract screening in the preparation of systematic reviews (Ouzzani et al. 2016). A separate Rayyan database will be created for each seed review.

3.3.4 Selection process

The titles and abstracts returned by the database search will be split between the two reviewers in each pair, each of whom will assess them for relevance against the seed review’s original inclusion/exclusion criteria and any updated criteria. Titles and abstracts that cannot be excluded on the basis of the information contained in their titles and abstracts will be included for full text review. Records which can be ruled out on the basis of the information contained in the titles and abstracts will be tagged as ‘exclude’ and the reason(s) for exclusion noted.

---

1 We operationalize QEDs as formal comparisons in which alternative teaching approaches or conditions are evaluated against each other (i.e. a treatment and comparator/control). By way of illustration, we will include non-equivalent groups designs, matched-pairs designs, and regression discontinuity designs. We will exclude single group pre-post designs, case studies, ethnographies, and cross-sectional designs. Methodological shortcomings will be reflected in the weight of evidence assessments for new studies in the update.
3.3.5 Quality assurance – Abstract screening
The first 10% of the titles and abstracts assigned to the review pairs will be reviewed by both members of the pair. The decisions for this 10% will be compared for consistency in applying the inclusion/exclusion criteria. Quality will be considered ‘assured’ if agreement is 90% or higher. If agreement is lower than 90% the reviewers will resolve disagreements through discussion, then repeat the process for the second 10% of the titles and abstracts. This process will continue iteratively until agreement is 90% or higher.

3.3.6 Full text screening
Following abstract screening, the full text of potentially relevant articles identified in the abstract screening process will be obtained. These will be divided by the review pair responsible for the update and screened against the inclusion/exclusion criteria of the original reviews and any updated criteria. Studies that do not meet all inclusion criteria will be excluded at this point and the reason(s) for exclusion noted.

3.3.7 Quality assurance – Full text screening
The same quality assurance process as for title and abstract screening will be followed for full text screening.

3.3.8 Data extraction
Once the final body of updated literature has been screened, the reports marked for inclusion for each seed review will be read closely and data will be extracted. Data will include information about the PICO (participants, intervention, comparator and outcome), the setting, the results, the study design, and the bottom-line conclusions. See Appendix 3, Data Extraction Sheet, for full details.

3.3.9 Trustworthiness appraisal of individual studies
Trustworthiness of individual studies identified in the update will be determined using Gorard’s sieve, a tool for assessing the trustworthiness of research findings (Gorard 2014). As with abstract and full text screening, the pair responsible for updating each seed review will assess the newly identified studies for trustworthiness. A sample of 10% of the studies will be appraised by both reviewers in the pair and checked for consistency. An agreement rate of 90% will be considered sufficiently close to allow for the remainder of the studies to be appraised independently. If agreement is lower than 90%, disagreements will be resolved through discussion, then another 10% of the studies will be double appraised. The process will be repeated iteratively until a 90% agreement rate is reached.

We acknowledge that there may be differences in the methods by which trustworthiness assessments have been made across seed reviews, and between the original seed reviews and our updates. Given the timeframe allotted to complete this rapid evidence assessment, it is not possible to reassess the trustworthiness of the individual studies included in the seed reviews so that all ratings conform to a common scale. We will work under the assumption that trustworthiness assessments are broadly valid across different seed
reviews, and make this position (and its attendant limitations) clear in the methods section of the report.

3.3.10 Synthesis

The review team has been tasked with addressing a series of questions that represent enormous scope and room for interpretation. Review Question 1, for example, asks ‘What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?’. The first part of this question implies an exploration into all possible approaches to teaching foreign languages that have been promulgated and tested. The second part refers to effectiveness. Measures of effectiveness used in individual studies are likely to vary considerably depending on the type of intervention, the outcomes of interest, the population of interest, and the nature of the comparator. Both parts of this question, therefore, invite the likelihood of substantial heterogeneity between studies. This heterogeneity may be in terms of participants, settings, interventions, comparators, outcome measures, and so on. Moreover, some of the review questions ask for information that is best addressed through qualitative approaches to the data. For example, Review Question 5 asks ‘Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?’. In addressing this question, it is more likely that useful information will be generated by reporting the descriptive findings of process evaluations in the primary research, rather than relying on bottom line findings based on quantitative data. As a result of this expected heterogeneity across studies and the nature of the review questions, we will follow guidance proposed by Popay et al. (2006) for narrative synthesis of heterogenous literature.

In brief, this approach to a narrative synthesis means we will generate textual descriptions of the primary research (both pre-existing based on information in the seed reviews and newly included studies) and organize these thematically. Themes will be informed by the nature of each review question, and will also emerge inductively in the process of preparing the REA. Indicatively, themes may include teaching approaches, language domains, populations, settings, study designs, and the nature of the results being reported (e.g. different outcomes, or different implementation factors). The weight of the evidence based on the trustworthiness appraisals in the seed reviews and conducted by the team on newly included research will be reported and discussed. Key features of individual studies will be tabulated for visual comparison on a theme by theme basis. Finally, where sufficient data are reported in the primary studies, we will create a common rubric to help understand how findings across individual studies compare to each other by calculating effect sizes with confidence intervals for comparison interventions. In addition, forest plots of the effect sizes and confidence intervals of individual studies within each theme will be constructed to provide a visual representation of the findings of the literature.

3.3.11 Calculating effect sizes

In the event that sufficient data is available to calculate effect sizes, they will be estimated using Hedge’s $g$, which is more robust against heterogeneity in the sample sizes and variances between the comparison groups than the more commonly used Cohen’s $d$ (Durlak, 2009). The $g$-statistic is calculated by dividing the differences in group means by the sample size weighted pooled standard deviations (Hedges & Olkin, 1985):
\[ g = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \text{ where } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \]

Since Hedge’s \( g \) is biased upwards for small samples \( (n < 50) \), the following correction will be applied (ibid.):

\[ g_c = g \times j \text{ where } j = \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) \]

The corrected \( g \)-values will be reported with their 95% confidence intervals, the upper and lower bounds of which will be calculated according to the following formula (ibid.):

\[ CI_{g_c} = g_c \pm (1.96 \times SE_{g_c}) \]

where

\[ SE_{g_c} = SE_g \times j = \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{g^2}{2(n_1 + n_2)} \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right)} \]

### Synthesising effect sizes

If the nature of the identified studies does allow for statistical synthesis, the following steps will be taken (as outlined by Turner & Bernard, 2006):

1. **Calculating simple averages of effect sizes within studies (if necessary)**
   Any studies which compare the same groups on multiple outcome measures violate the assumption of statistical independence that underlies meta-analyses. In these cases, a simple averages of the effect sizes and their standard errors will be calculated.

2. **Calculating weighted averages of effect sizes across studies**
   Since studies with larger samples provide more precise estimates, effect sizes will be weighted according to their standard error when calculating the average effect size across different studies. That is, each effect size will first be multiplied by the inverse of its variance to yield the weighted effect size:

   \[ g_w = g_c \times IVW \text{ where } IVW = \frac{1}{(SE_{g_c})^2} \]

   Subsequently, the average weighted effect size can be calculated by dividing the sum of the weighted effect sizes by the sum of the inverse variances:

   \[ \overline{g_w} = \frac{\sum_{i=1}^{n} g_w}{\sum_{i=1}^{n} IVW} \]

3. **Assessing observed variability in effect sizes pooled across studies**
Finally, a 95% confidence interval will be calculated for the average weighted effect size using the following method:

\[ CI_{\bar{g}_w} = \bar{g}_w \pm (1.96 \times SE_{\bar{g}_w}) \]

where

\[ SE_{\bar{g}_w} = \sqrt{\frac{1}{\sum_{i=1}^{n} IVW}} \]

### 3.3.12 Making recommendations

Recommendations for practice and policy will be made on a theme by theme basis and will be informed by the nature and direction of the findings of the body of literature informing each theme. Inevitably, given the expected heterogeneity of the literature this will rely on a vote counting approach. While vote counting has been criticized for its potential to introduce bias, we accept Popay et al.’s (2006) contention that it can be a useful tool in a narrative synthesis, provided care is taken to account for the trustworthiness appraisal of individual studies, which reports on factors considered important to assessing the weight of the evidence; study size, design and so on. Recommendations will draw on the overall direction of findings in each theme and any important caveats and exceptions to general trends will be highlighted and discussed.

### 4. Timeline

- Phase 1 – September 1 – November 1st
- Phase 2 – November 4 – December 6th
- First draft of report – December 6th – 20th
- Liaise with EEF re: feedback on first draft w/c January 13th 2020
- Final draft ready by January 31st, 2020

### References


Appendices
Appendix 1 – Text of email to professional networks in language education

Dear colleagues,

The Applied Linguistics Research Group at the Department of Education at Oxford is currently working on a Rapid Evidence Assessment, funded by the Education Endowment Foundation, of research into languages in education and foreign language teaching and learning.

Our specific review questions are:

- What approaches to teaching a foreign language have been used, and what is the evidence on their effectiveness?
- What is the impact of learning a foreign language on students’ wider academic outcomes?
- What practitioner skills or programme characteristics contribute to effective language learning among students?
- What is the impact of using a non-native language as the medium of instruction in academic subjects on students’ academic outcomes?
- Are there implementation factors that lead to a positive impact on attainment of using a non-native language as the medium of instruction?
- What is the impact of delaying or accelerating the introduction of a new ‘local’ language as a medium of instruction for new arrivals (e.g., refugees, immigrants) who are not yet proficient in their native language?

To that end I am emailing to ask members of [insert name of organisation] to email to us details of any research of which they are aware, either published via traditional channels or constituting ‘grey literature’ (e.g., working papers; PhD theses), that fits the following criteria:

Research which was
- Published / distributed / released in or after 2000
AND which is a
- Systematic review;
- Meta-analysis;
- Narrative review
- State-of-the-art article/review; or
- Rapid evidence assessment
AND which focuses on
- Language(s) in education;
- Second / foreign / additional language learning; and/or
- Second / foreign / additional language teaching

In addition, if you know of any other research (i.e. stand-alone studies) that speaks directly to our review questions, we would be grateful for details of these. Please kindly email references (or weblinks, full text documents etc.) of any relevant studies to Henriette Arndt (henriette.arndt@education.ox.ac.uk) at your earliest convenience, and thank you for your help!
Appendix 2: CASP appraisal tool for Systematic Reviews

Adapted CASP Checklist for assessing the quality of Systematic Reviews

1. Reference
HINT: Include full citation info for the paper being assessed

2. Does the review address a question of relevance to the objectives of the REA?
[0=Not at all | 5=yes, completely]

HINT: If the review does not address a question of relevance to the aims of the REA, stop here!

3. Did the review address a clearly focused question?
[0=Not at all | 5=yes, very clearly focused]

HINT: An issue can be ‘focused’ in terms of
• the population studied
• the intervention given
• the outcome considered

4. Did the authors look for the right type of papers?
[0=Not at all | 5=yes, included research addressed the review questions precisely and were of the highest quality methodological designs for the purpose]

HINT: ‘The right type of papers’ would
• address the review’s question
• have an appropriate study design (RCTs or QEDs as per the EEF’s guidelines)

5. Are the search methods used to identify the relevant studies described in enough detail to permit replication?
[0=No, there is not enough information to replicate the search | 5=yes, replicating the search based on the information in the review would be very straightforward]

HINT: Descriptions should include
• Search date
• Databases used
• Search strategy
• Search terms

6. Do all important, relevant studies seem to be included?
[0=No, information sources are not reported at all | 5=all information sources are reported and appear to be exhaustive]

HINT: Look for
• which bibliographic databases were used
• follow-up from reference lists
• personal contact with experts
• unpublished as well as published studies
• non-English language studies

7. Did the review’s authors do enough to assess quality of the included studies?
[0=No risk of bias or Weight of Evidence assessment procedures are reported | 5=Risk of bias or Weight of Evidence assessments conform to the highest standards]

HINT: The authors need to consider the rigour of the studies they have identified. Lack of rigour may affect the studies’ results (“All that glisters is not gold”: Merchant of Venice – Act II, Scene 7).

8. If the results of the review have been combined, was it reasonable to do so?
[0=Statistical syntheses do not take into account differences or similarities between studies | 5=any statistical syntheses undertaken are entirely appropriate and well conducted]

HINT: Consider whether
• results were similar from study to study
• results of all the included studies are clearly displayed
• results of different studies are similar
• reasons for any variations in results are discussed

9. Are the overall results of the review clear and precise?
[0=the overall results of the review are not reported or extremely difficult to understand | 5=the overall results of the review are very clearly reported and precise]

HINT:
• Consider whether the ‘bottom line’ results are clearly presented, and in what way.
• Are the results expressed precisely (e.g. numerically, if appropriate)?
• Are the stated conclusions supported by the data presented?

10. How precise are the results?
[0=results are not reported, or no statistical information is provided at all | 5=where appropriate, statistics are reported precisely using confidence intervals]
HINT: Look at the confidence intervals, if given

11. Can the results be applied to the current REA?
[0=the results do not address any of the review questions driving the REA | 5=the results are exactly relevant to some or all of the review questions driving the REA]

HINT:
• Which review question(s) do they address?
• How strong is the evidence?
• Are there any mediators to consider?

12. Were all important outcomes considered?
[0=there appear to be significant omissions | 5=all relevant information appears to have been thoroughly covered in the review]

HINT: Consider whether there is other information you would like to have seen

13. Is this review of high enough quality and sufficient relevance to inform the REA?
[0=not at all on both counts | 5=the review is exactly relevant to the aims of the REA and was conducted to the highest standards for systematic reviews]
# Appendix 3: Data Extraction Sheet

<table>
<thead>
<tr>
<th>Item</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design.</strong> Describe the design of the study (e.g. RCT, matched comparison, RDD etc.)</td>
<td></td>
</tr>
<tr>
<td><strong>Participants.</strong> Describe in as much detail as is given in the report who took part in the study. Include, for example, age, gender, socio-economic status, L1 etc.</td>
<td></td>
</tr>
<tr>
<td><strong>Intervention.</strong> Describe the treatment intervention being evaluated in as much detail as necessary/available to understand what the treatment intervention was.</td>
<td></td>
</tr>
<tr>
<td><em><em>Comparator</em>.</em>* Describe the intervention against which the treatment is being compared, in as much detail as necessary/available to understand what the comparison intervention was.</td>
<td></td>
</tr>
<tr>
<td><strong>Outcomes.</strong> Describe the outcome measures used in the study.</td>
<td></td>
</tr>
<tr>
<td><strong>Results.</strong> Look for effect sizes and confidence intervals, or means and standard deviations, or raw scores. Report the results of statistical tests, if used. If the results are narrative, summarise them here. If results were disaggregated by participant characteristics (e.g. moderator analyses were included in the report) report them here, or append them to the bottom of this sheet.</td>
<td></td>
</tr>
<tr>
<td><strong>Factors contributing to success.</strong> Record here any reporting in the study that provides information (or hypotheses) about what contributed to the success of the interventions (assuming they were successful).</td>
<td></td>
</tr>
<tr>
<td><strong>Limitations.</strong> Did the authors identify limitations to their study that may compromise the trustworthiness of the findings? This might</td>
<td></td>
</tr>
</tbody>
</table>
include attrition, fidelity to the intervention, disruption to the schedule etc. If you have noticed anything that may have compromised the trustworthiness of the findings, which the authors have left unremarked on, note it here.

**Bottom line conclusion.** Summarise the bottom line finding of the study. This can be a verbatim excerpt from the text or in the reviewer’s own words.