





## EMPIRICAL STUDY

# Effects of Word Definitions on Meaning Recall: A Multisite Intervention in Language-Diverse Second Language English Classrooms

Henrik Gyllstad <sup>a</sup>, Pia Sundqvist <sup>b</sup>, Erica Sandlund <sup>c</sup>,  
and Marie Källkvist <sup>a,d</sup>

<sup>a</sup>Lund University, <sup>b</sup>University of Oslo, <sup>c</sup>Karlstad University, and <sup>d</sup>Linnæus University

**Abstract:** Vocabulary experts recommend first language (L1) translation equivalents for establishing form–meaning mappings for new second language (L2) words, especially for lower proficiency learners. Empirical evidence to date speaks in favor of L1 translation equivalents over L2 meaning definitions, but most studies have investigated bi- rather than multilingual learners. In our study, we investigated instructed English vocabulary learning through an intervention study in six language-diverse secondary school English classrooms in Sweden ( $N = 74$ ) involving three conditions for presentation of word meanings: (a) definitions in the L2 (English), (b) translation equivalents in the shared school and majority language (Swedish), and (c) translation equivalents in the shared school and majority language plus other prior languages among the learners (Swedish and other). Based on overall weighted mean effect sizes and mixed-effects

---

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at <https://oasis-database.org>

We gratefully acknowledge funding from The Swedish Research Council, Project ID 2016–03469. Our gratitude extends to the participating teachers and students as well as to school leaders and staff who contributed to making our research possible. We are also very grateful to the Journal Editor Emma Marsden for exceptionally helpful feedback and recommendations and to five anonymous reviewers for their constructive comments and suggestions during the review process. Any remaining flaws are our own responsibility.

Correspondence concerning this article should be addressed to Henrik Gyllstad, Centre for Languages and Literature, Lund University, Box 201, 221 00 Lund, Sweden.

Email: [henrik.gyllstad@englund.lu.se](mailto:henrik.gyllstad@englund.lu.se)

The handling editor for this article was Emma Marsden.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

modeling, the results showed that conditions that involved L1 translation equivalents yielded higher scores than did target language definitions in immediate posttests with a small effect size but no differences in delayed posttests.

**Keywords** word learning; word meaning; translation equivalents; multilingualism; intervention; EFL teaching

## Introduction

A long-standing issue in instructed second language (L2) acquisition is whether to teach a L2 more or less exclusively through the L2 or to use students' prior language(s) as scaffolding, typically a student's first language (L1). Recent classroom research has suggested that there may be cognitive as well as social benefits for drawing on students' prior language(s) for learning both L2 grammar and vocabulary (e.g., Bartolotti & Marian, 2017; Källkvist, 2013; Shin et al., 2019).

The extent to which the target language is used in L2 English learning relative to L1 use is of particular interest in Swedish education, characterized by the omnipresence and high status of the English language (Hult, 2012) as well as considerable diversity in student prior languages. Young learners' English proficiency levels are generally high, with Swedish adolescents (aged 15–16 years) even scoring on a par with L1 English speakers from Malta, where English is the official language, in the European Survey on Language Competences (European Commission, 2012). Curricula at the national level for compulsory school (Swedish National Agency for Education, 2018) have provided no guidelines regarding the balance of target language versus use of L1 or other shared languages in L2 English classrooms. Instead, educational policy has left language practices up to the professional judgments of individual teachers (Hult, 2017). A questionnaire study of English teachers' self-reported beliefs and classroom practices revealed that 10 out of 139 responding teachers (7.2%) fully agreed with the statement "When I teach English, I use English only" (Sundqvist et al., 2018). In addition, 23 of the teachers (16.5%) fully agreed with the statement "Students learn English best when they stick to only English during English lessons." Similar beliefs are reported in Amir and Musk (2013). Furthermore, target-language use only is required in the annual, mandatory standardized tests of English administered in Grades 6, 9, and 10. Task instructions for the speaking part (all three grade levels) include the following wordings "Be active and speak English all the time" (Swedish National Agency for Education, 2013, p. 2). Counterexamples to Swedish educational policy include the Common European Framework of Reference for Languages (Council of Europe, 2020), which recommends

“extensive use of the target language in the classroom” (p. 30). The American Council on the Teaching of Foreign Languages (n.d.), however, has been more specific, recommending at least 90% use of the target language, with use of the L1 being “reserved for very strategic purposes.”

A rather well-researched domain of L2 learning is vocabulary, with a number of studies (reviewed in the following section) showing that translation equivalents in learners’ L1 often, but not always, have yielded higher levels of word learning than target-language definitions and synonyms. Most studies have, however, involved bi- rather than multilingual learners (e.g., see Laufer & Shmueli, 1997; Lee & Macaro, 2013; Liao, 2006). Growing language diversity in many educational settings across the world calls for attention to classrooms where there is heterogeneity beyond two languages (Baker & Wright, 2021).

Our study addressed this gap by investigating instructed English vocabulary learning through a multisite intervention study, balancing ecological validity and experimental control in language-diverse secondary school English classrooms involving conditions in which word meanings were given as (a) definitions in the target-language English, (b) translation equivalents in the shared school and majority language Swedish, and (c) translation equivalents in the shared school and majority language plus other prior learned languages among the participants, that is, Swedish and other languages. These three learning conditions stemmed from the societal and educational context in Sweden that is characterized by linguistic superdiversity (Blommaert, 2010). Sweden has seen a steady rise in language diversity over the last three decades, and, to date, approximately 26% of school-age students have a migrant, non-Swedish-speaking background (Swedish National Agency for Education, 2021a), making research in this context timely and relevant.

## **Background Literature**

### **The Use of the L1 in L2 Vocabulary Learning**

Even though the ability to define L2 English words through English only is often aspired to in instructed L2 acquisition settings, especially among advanced-level students (Levine, 2003), L2 vocabulary experts have commonly recommended L1 use for learning. Schmitt and Schmitt (2020) stated that even though the use of L1 translation equivalents “is unfashionable in many quarters” (p. 167), it is sensible to draw on these when appropriate to do so, particularly when establishing the initial form–meaning link, especially in the light of the ubiquitous nature of L1 influence during lexical processing. Similarly, Nation (2013) emphasized that studies on L1 use have shown that learning is facilitated by glosses in the L1. Learners have also reported using

L1 glosses as a common strategy in L2 vocabulary learning (Barcroft, 2009) and relying on bilingual dictionaries (Schmitt, 1997).

Furthermore, in the psycholinguistics-oriented literature, a model like the revised hierarchical model (Kroll & Stewart, 1994) predicts that lower L2-proficiency learners will rely on links to L1 translation equivalents to a greater extent than will higher L2-proficiency learners.<sup>1</sup> These predictions have been corroborated in numerous empirical studies, indicating that the L1 is automatically activated during L2 lexical processing in both beginner and advanced learners (e.g., see Carrol et al., 2016; Elston-Güttler & Williams, 2008; Sunderman & Kroll, 2006), a finding that has supported the use of L1 translation equivalents. As a consequence, the observed automatic L1 activation arguably makes it futile to ban or ignore the L1 in L2 learning.

### **Empirical Studies on L1 and L2 Meaning Definitions for L2 Vocabulary Learning**

A number of studies have compared different approaches to meaning definition in L2 vocabulary learning in experimental or quasiexperimental designs. Prince (1996) studied L1 French university-level learners ( $N = 48$ ) of varying L2 English proficiencies by comparing a L1 translation equivalent group to a L2 contextual learning group. On the basis of a set of 44 English nouns in a 20-minute learning phase, Prince found that both lower and higher L2-proficiency groups performed better in the L1 translation equivalent condition, arguing that linking a new word to its translation equivalent(s) is a rapid way of establishing L2 word meaning.

Laufer and Shmueli (1997) studied L1 Hebrew learners of L2 English ( $N = 128$ ), assigned to one of four experimental groups or to a control group. One independent variable involved the presentation mode (words in lists, sentences, texts, and elaborated texts), and a second entailed manipulation of language of presentation (L2 word paraphrasing and L1 translations). Laufer and Shmueli tested 20 low-frequency target words; the results showed that L1 glosses proved more beneficial for retention than L2 glosses. The authors argued that L1 glosses allowed full attention to the new L2 word compared to the L2 glosses, where longer paraphrases may have made it difficult for learners to focus on the target word. However, design weaknesses included no test of learners' prior target word knowledge, and the researchers used a translation test to gauge vocabulary knowledge, causing bias toward one of the experimental conditions.<sup>2</sup>

Investigating L1 French university students ( $N = 191$ ), Hummel (2010) hypothesized that exposure to translation equivalents and active translation

involve deeper, more elaborated processing, facilitating retention. Hummel randomly assigned participants to one of three tasks: translating L1 sentences to L2, the L2 sentences to the L1, and an exposure and copy (L2 only) condition. Hummel found significant target word retention in all three conditions, with rote-copying being the best, and no difference between the two translation conditions. Hummel suggested that the poorer results for translation were due to cognitive processing overload. Thus, Hummel's study is an example where no positive effect for L1 use was found. However, the focus seemed to have been translation of longer sentences rather than L1 translation equivalents of individual words.

In a series of studies, Macaro and colleagues have investigated L1 versus target-language use. Tian and Macaro (2012) studied Chinese L2 English university students ( $N = 80$ ), focusing on teacher codeswitching and its effect on vocabulary learning during listening comprehension. They employed codeswitching versus target-language-only conditions and incorporated intentional and incidental word learning into the design. Tian and Macaro's results showed that intentional word learning and teacher codeswitching were superior compared to target-language-only information, whereas they found no effect of proficiency. In Lee and Macaro (2013), two groups, Grade 6 students ( $n = 443$ ) and L1 Korean university students ( $n = 286$ ) of L2 English, participated in a classroom study. Following two week-long instructional sessions, both groups benefitted from teacher L1 use compared to L2-only use when tested both on word form recall and form recognition, with young learners benefitting more. Lee and Macaro concluded that L2 proficiency level was a contributing variable. Zhao and Macaro (2016) investigated teacher L1 use compared to teacher L2-only explanations and learners' vocabulary uptake in two experimental groups ( $n = 50$  in each group). Compared to a control group ( $n = 48$ ), L1 Chinese learners reached higher vocabulary gains in the L1-use condition. Zhao and Macaro suggested that the results were due to a difference in how learners retrieve lexical information, with direct links between L2 words and L1 translation equivalents yielding processing ease compared to more complex target-language-only explanations.

Joyce (2018) investigated the knowledge of academic vocabulary among L1 Japanese undergraduates ( $N = 48$ ) in Japan, whose proficiency ranged from false beginners to upper-intermediate-level English learners. They were assigned to one of two experimental groups for a 10-week-long treatment. Group 1 studied 100 target words using English definitions in List A and another 100 words using Japanese translation equivalents in List B. Group 2 studied counterbalanced lists of the same words, where List A featured Japanese translation

equivalents, and List B featured English definitions. Joyce tested the participants' receptive vocabulary knowledge through multiple-choice pretests and posttests. Joyce found no difference for study language as a main effect but an interaction between study language and testing language, with higher scores when these two conditions matched.

Studies have also compared L1 use and pictures. Lotto and De Groot (1998) compared two learning methods for word learning in an unfamiliar language (L1 translation equivalents or pictures) and tested two further variables (cognateness and frequency). Adult L1 Dutch psychology students ( $N = 56$ ) with no prior Italian knowledge took part, with participants receiving either Dutch words or pictures for a set of 80 Italian words. In both conditions, the participants were instructed to name and type an Italian word (i.e., form recall). L1 translation equivalents led to higher levels of learning than did the picture condition. Lotto and De Groot concluded that L1 translations are commonly used by learners, and that their results might have been due to habitual effects.

Some studies have involved educational contexts with more than two languages. In two studies, Hopp et al. (2018) and Hopp et al. (2019) investigated the contribution of minority and majority languages to early L2 English learning in Germany. The L1 was a better individual predictor for third language (L3) vocabulary than L2 vocabulary, suggesting a pivotal role of the L1 lexicon for conceptual knowledge and further learning (Hopp et al., 2018). Hopp et al. (2019) found that bilingual benefits emerged for vocabulary when controlling for socioeconomic variables at the school level. Hirosh and Degani (2021) studied learning vocabulary in a L3. In a between-groups design ( $N = 59$ ), with language of instruction as the primary independent variable, one group was given the task of learning a set of 55 new L3 German words through their L1 Hebrew and another group the task of learning the same L3 words through their L2 English. Hirosh and Degani predicted that the L1 condition would give rise to more learning through more cognitive resources being available as well as more accumulated experience. The results showed that the L1 learning condition yielded better learning, except for cognates, which were learned equally well in both conditions.

Finally, a recent intervention study by Busse et al. (2020) featured primary school (approximately 8.5 years old) students of English ( $N = 42$ ) in two classes in a German school, many of low socioeconomic background and with a low English literacy level. One class acted as an experimental group with a multilingual approach and the other as a control group receiving regular teaching. The intervention comprised a pretest, posttest, and delayed posttest, and a five 45-minute English lesson treatment over 3 weeks. The treatment entailed

activities asking students how prior languages could be seen as treasures. They were asked to translate words into their prior (i.e., L1 or early-learned) languages, using memory games with German word cards and word cards in their prior languages. The researchers assessed multilingual ideal and English ideal self-aspirations, affect, and vocabulary learning. The multilingual approach group made considerably larger learning gains than the regular teaching group on productive and receptive English vocabulary across the three measurement points.

### **Recent Meta-Analyses on L1 and L2 Meaning Definitions for L2 Vocabulary Learning**

In a meta-analysis, Lee and Lee (2022) pooled 14 studies investigating teachers' verbal lexical explanation for vocabulary learning in a L2. The meta-analysis included data from 3,304 learners. The study showed first, and perhaps not surprisingly, that the use of teachers' explanations was more effective than was the absence of such explanations. More importantly, however, L1 explanations led to more vocabulary knowledge than did L2 explanations, both at immediate (effect size  $d = 0.59$ ) and delayed (effect size  $d = 0.28$ ) posttests.

In another meta-analysis, Yanagisawa et al. (2020) investigated the overall effects of glossing on L2 vocabulary learning from reading on the basis of 42 studies comprising 3,802 participants. Glossing in a L2 situation refers to the provision of meaning indication through either L1 translation equivalents, target-language (L2) synonyms, or shorter L2 definitions. Previous studies have generated inconsistent results, with Ko (2012) and Yoshii (2006) finding no difference between L1 and L2 glossing, whereas Xu (2010) reported better results for L1 glosses compared to L2 glosses. In the meta-analysis, Yanagisawa et al. (2020) investigated the influence of five potential moderator variables: gloss format type, language, mode, text characteristics, and learner characteristics. Glossed reading led to more learning than nonglossed reading. Most relevant for our study was that Yanagisawa et al. found L1 glossing to be more beneficial than L2 glossing. They observed no interactions for glossing language and proficiency but did find that learning gains were moderated by proficiency.

### **Summary of Results From Previous Research Informing the Present Study**

Taking stock of the above literature review, although the previous studies differed in study design, participant L1 backgrounds, education level(s), domain targeted (spoken or written), and type of word knowledge tested (meaning

recognition vs. meaning recall), more often than not the use of L1 translation equivalents for word meaning definition seemed to yield a higher vocabulary learning pattern than did L2 definitions, but with some exceptions. In addition, a number of recent meta-analyses have indicated more vocabulary learning for L1 translation equivalents than for L2 meaning definitions. Many studies have suggested that a higher cognitive load imposed by L2 definitions lies behind the higher learning scores following L1 translation equivalents relative to those following L2 definitions. Other explanations have included habitual effects, that is, learners do better at what they are accustomed to, and also when the learning condition matches the way word knowledge is tested. In addition, classroom interventions comparing learners' drawing on prior languages compared to a focus on L2-only use have resulted in better L2-vocabulary scores for prior language use.

Some studies that we have reviewed had clear design weaknesses such as no pretest of the targeted vocabulary or translation used as the way to gauge learning, causing a bias toward one of the manipulated conditions. The mixed results regarding the effects of proficiency motivated us to include a measure of proficiency in our study. Finally, few quantitative studies have been conducted in multilingual secondary school level classrooms. Our study aimed to address these observed gaps.

### **The Present Study**

In this study, we researched short- and long-term instructed vocabulary learning through three week-long classroom interventions involving three conditions in which the presentation of English word meanings was given in the target language, in the shared school language (i.e., the majority society language) and in the shared school language plus any other prior language in learners' repertoires. The conditions for word meaning presentations were thus: (a) English as the target language, (b) Swedish as the school language, and (c) Swedish as the school language, plus students' additional prior languages.

In all conditions, the teacher used English as the medium of instruction in all lessons, but with the following modifications. In the English condition, vocabulary-learning materials presented English target words with meaning definitions in English, and students were encouraged to use only English. In the Swedish condition, the vocabulary-learning materials juxtaposed the English target vocabulary with Swedish translation equivalents, and the teacher used Swedish to provide the meaning of English target words. Finally, the Swedish and other languages condition provided materials in which English target words were presented with translation equivalents in Swedish and in all



other prior languages represented among the students in each class, and the students were encouraged to draw on their own set of prior languages. In this context, Swedish was the school language and the society majority language and the L1 for students who were exposed to Swedish from birth. For students who had migrated and encountered Swedish later, both Swedish and English were additional languages. With this research design, we sought to address the following research questions:

1. What are the relative effects of presenting vocabulary meaning definitions through (a) target-language (English), (b) school-language (Swedish), and (c) school language (Swedish) plus heritage languages on students' word meaning recall knowledge?
2. To what extent are effects moderated by students' language background, English proficiency, and English and Swedish school subject grades?

## Method

### Participants

Students from six intact classes from four secondary schools located in urban areas in two parts of Sweden participated in the study (see Table 1).<sup>3</sup> Thus, we used a multisite design in line with suggestions for beneficial approaches for instructed L2 acquisition research (Moranski & Ziegler, 2021). This allowed for data collection that yielded greater sample size while keeping conditions as similar as possible. Greater sample size yields greater statistical power, in turn increasing the probability of capturing existing statistically significant differences where those exist.

All participants (aged 14–16 years) were enrolled in Grade 9 English as an additional language (a mandatory subject) when the three-week intervention took place. The total number of learners in the classes was initially 127, but the final number used in our analyses was 74 learners (52% girls, 48% boys). We recruited schools and teachers using convenience sampling, aiming at multilingual L2 classes that varied in terms of mean overall grades and socioeconomic status. Table 1 shows that a school with a higher proportion of students who risked not being admitted to upper-secondary school compared to the mean in Sweden had a socioeconomic index higher than 100, but a school with a lower proportion had an index lower than 100. Thus, the lower the index, the more likely that the school had students who would be eligible to move on to upper-secondary school (Statistics Sweden, 2019; Swedish National Agency for Education, 2021b, data from 2019/2020). A multilingual class was operationalized as including at least five students who used Swedish and one or more

**Table 1** Overview of participating schools, teachers, and classes

School	Region	Municipality	Teacher	Classes	Overall grade <sup>a</sup>	Students' migration background <sup>b</sup>	Caregivers' tertiary education <sup>c</sup>	Noneligible students <sup>d</sup>	Socioeconomic index <sup>e</sup>
1	1	City	Jill	1A	174	69.0	37.0	29.8	213.9
2	1	Town	Anita	2A, 2B	232	34.0	28.0	28.8	206.4
3	2	Major city	Unni	3A, 3B	246	73.0	48.0	19.0	136.2
4	2	City	Kajsa	4A	261	31.0	81.0	12.4	88.6
Total	2	4	4	6	—	—	—	—	—

*Note.* Teacher names are pseudonyms.

<sup>a</sup>Mean overall grades; mean grade for Sweden = 221 for the school year 2016/2017 (Swedish National Agency for Education). The overall grade is the sum of the 16 highest grades in a student's final school leaving report in Grade 9, or the sum of 17 grades if a student has studied a modern language (typically French, German, or Spanish) as part of the Language Option. The letter grades (A–F) are transformed: A = 20, B = 17.5, C = 15, D = 12.5, E = 10, and F = 0. Thus, the highest possible overall grade is 340.

<sup>b</sup>Expressed as percentage, defined as born abroad or born in Sweden with both parents born abroad.

<sup>c</sup>Expressed as percentage, defined as having studied at least one semester at university level (Swedish: *eftergymnasial utbildning*).

<sup>d</sup>Expressed as percentage of students not eligible to apply to a national program in upper-secondary school.

<sup>e</sup>Calculated annually (for each school) to establish the size of a government subsidy that should contribute to equivalence in schools. The index is based on predictions about the expected proportion of students who will not be eligible for admission to upper-secondary school upon leaving Grade 9. The socioeconomic index builds on data about gender, caregivers' educational background, caregivers' income, year of immigration (when applicable), family, siblings, and the socioeconomic status of the neighborhood (Statistics Sweden, 2019).

languages, for example, Arabic, Finnish, or Somali, in their everyday life, usually in the home. Use of these different languages varied and was self-reported in a questionnaire (see Appendix S1 in the Supporting Information online). Thus, we used a nonprobability sampling method, but the sample nevertheless shared characteristics with the target population, that is, multilingual lower-secondary school students of L2 English in Sweden (compared with Sundqvist et al., 2021).

### *The Six Intervention Classes*

The English teachers consented to observations of their lessons in Grades 7 and 8 multilingual English classrooms and to a three-week intervention in Grade 9. Preintervention observations served to collect ethnographic data and build trust with the students and teachers. Two classes were profile classes. Class 3B was a content and language integrated learning class, with students receiving English-medium instruction in three subjects (history, art, and sports) other than English since Grade 7. Furthermore, Class 3A was a fast-track English class, studying three years of lower-secondary school English courses (Grades 7–9) in two years' time (in Grades 7–8), and studying upper-secondary school English in Grade 9. For this reason, the Class 3A intervention had to take place in Grade 8, that is, when Class 3A was studying Grade 9 English. On balance, all the participating classes shared enough characteristics to render their inclusion in the study justifiable.

### *Language Background*

We elicited data on participants' language background using a questionnaire (see Appendix S1 in the Supporting Information online). We created the independent variable language background on the basis of the questionnaire data, and it aligned theoretically with our operationalization of multilingual student. In the final sample of 74 participants, 19 students (25.7%) had a L1 Swedish background; 38 students (51.4%) were simultaneous bilinguals of Swedish and another language or were born abroad but moved to Sweden with an age of onset of Swedish before age 3 years; and 17 students (23.0%) were successive multilinguals, that is, they had a L1 other than Swedish, with both Swedish and English as L2s or were multilinguals born abroad with an age of onset of Swedish at age 3 years or later (see Table 2).

## **Ethics**

The study was part of the MultiLingual Spaces project funded by the Swedish Research Council (Reg. no. 2016–03469) and underwent ethical review. We

**Table 2** The language background and English proficiency scores of the students in the six intervention classes

Class	L1 Swedish background		Bilingual background		Other language background		English proficiency score <sup>a</sup>	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>M</i>	<i>SD</i>
1A ( <i>n</i> = 7)	0	0.0	5	71.0	2	29.0	21.86	4.41
2A ( <i>n</i> = 10)	5	50.0	0	0.0	5	50.0	19.70	6.45
2B ( <i>n</i> = 14)	7	50.0	3	21.4	4	28.6	20.50	5.83
3A ( <i>n</i> = 17)	0	0.0	15	88.2	2	11.8	26.88	2.67
3B ( <i>n</i> = 16)	1	6.2	11	68.8	4	25.0	22.93	3.35
4A ( <i>n</i> = 10)	6	60.0	4	40.0	0	0.0	25.10	4.33
Total ( <i>N</i> = 74)	19	25.7	38	51.4	17	23.0	22.83	4.50

*Note.* The rounded percentages do not always sum to 100%.

<sup>a</sup>Maximum score = 32.

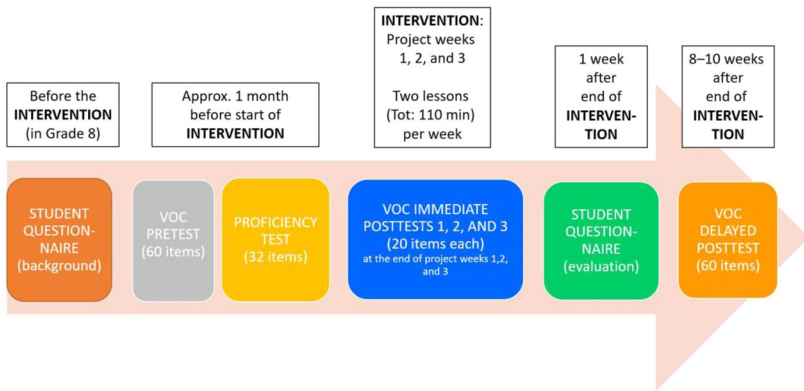
collected written informed consent from teachers, students, caregivers, and school leaders, and we provided information in parent meetings. We informed the participants that no data from the study would be shared with their teachers, but the participants themselves received their individual test scores postintervention.

### Design of Intervention

We conducted a quasiexperimental, mixed within- and between-subjects intervention study, featuring a pretest administered 1 month prior to the intervention, immediate posttests administered at the end of the last lesson of a week, and a delayed posttest administered 8–10 weeks after the intervention (10 weeks after Week 1, 9 weeks after Week 2, and 8 weeks after Week 3; see Appendix S2 in the Supporting Information online for all test materials). The three conditions defined above—target language (English), school language only (Swedish), and Swedish plus all students' L1s (Swedish and any other languages)—were used as levels of an independent variable condition, together with an independent variable test with three levels (pretest, immediate posttest, and delayed posttest).

### *Balancing Experimental Control and Ecological Validity*

We sought to strike a balance between an ecologically valid context and desirable levels of experimental control. This can be challenging due to difficulties in controlling for extraneous independent variables (Baker & Wright,



**Figure 1** The principal design of the intervention study.

2021; Hulstijn, 1997). Yet, there are calls for research “in real classrooms with real learners” (Spada, 2005, p. 330; see also Leung & Valdés, 2019). Marsden (2007) and Spada (2019) have discussed such challenges, addressing researchers’ dilemma of attempting a tradeoff between maintaining control of the procedures while also reducing experiment artificiality (see also Sato & Loewen, 2019, for a discussion of ecological validity). Thus, we aimed for a balance by conducting our study in a classroom setting, following carefully planned procedures, using detailed lesson plans, counterbalancing treatment order, and documenting classroom events through audio and video recordings. The regular classroom teacher was present, prepared to take care of any unexpected events in the classroom such as IT problems or student discipline so that the guest teacher (researcher) could remain focused on following the protocol for the lesson. Altogether, we aimed to maximize treatment fidelity, emphasizing consistent and uniform procedures across the classrooms.

### *Design*

Each of the six classes participated over three weeks either right before or after the Christmas break (2018/2019 and 2019/2020; see Figure 1). In conjunction with administering the vocabulary pretest and the English proficiency test, the researchers informed the participants about the upcoming intervention. The research team consisted of the four authors, divided into teams of two—one team for each region—where one researcher served as the guest teacher during the intervention and the other was responsible for audio- and video-recording the lessons.

**Table 3** Counterbalanced design of conditions

Class	Condition		
	Intervention Week 1	Intervention Week 2	Intervention Week 3
	Lessons 1 & 2 (110 min)	Lessons 3 & 4 (110 min)	Lessons 5 & 6 (110 min)
1A	E	S	SO
2A	SO	E	S
2B	S	E	SO
3A	E	SO	S
3B	S	SO	E
4A	SO	S	E

*Note.* In the teaching/learning materials in Appendix S8 of the Supporting Information online, the English (E) treatment is referred to as A, the Swedish (S) treatment as B, and the Swedish and other languages (SO) treatment as C.

The team worked in close collaboration throughout all interventions, sharing progress updates and experiences and discussing any questions or issues regularly to make the interventions run smoothly and similarly. To control for condition order, we adopted a counterbalancing design (see Table 3). This ensured that the word sets featuring in the three project weeks were the same, but the different classes encountered the (same) words under different conditions. For example, Class 1A worked with the learning material and pertinent 20 words (12 infrequent and eight frequent) in Week 1 through the English condition, whereas Class 2A encountered the same 20 words in Week 1 through the Swedish and other languages condition, and Class 2B encountered the same 20 words in Week 1 through the Swedish condition.

### *Materials*

In Sweden, teachers are free to choose teaching materials provided that the materials are in line with the curriculum. In order to make sure the materials were new to all classes, we based teaching on texts from the *VOICES in Time 3* textbook used in Norway (Brevik, 2008). We selected texts from chapters under the theme of Freedom, complemented with an authentic text about Nadine Gordimer (see Table 4). We selected target words using words found in these texts that we complemented with a small number of additional words.

We analyzed the texts in terms of difficulty using the Flesch Reading Ease Score (FRES; Flesch, 1949), a readability formula first used in 1948 and frequently used today (Alderson, 2000). The formula is based on mean

**Table 4** Texts used in the project and their Flesch Reading Ease Scores (FRES)

Week	Text	Source	FRES
1	<i>Martin Luther King – Free at Last</i>	Textbook	76.55
2	<i>Goodbye Bafana + The Rise of a Nation</i>	Textbook	72.28
3	<i>Nadine Gordimer<sup>a</sup> + The Moment Before the Gun Went Off</i>	Authentic + textbook	70.53

*Note.* Textbook = *VOICES in Time 3* (Brevik, 2008). <sup>a</sup>Based on texts retrieved from [https://en.wikipedia.org/wiki/Nadine\\_Gordimer](https://en.wikipedia.org/wiki/Nadine_Gordimer) and <https://themanbookerprize.com/fiction>.

sentence length and the mean number of syllables per word (see Appendix S3 in the Supporting Information online). The higher the FRES, the easier a text is to read, and texts with FRES scores of 60–70 are expected to be easily read by English native-speaking Grade 8 and 9 students, whereas those with FRES scores of 70–80 are expected to be easy to read for Grade 7 students. We adapted the texts to be in the range of 70–80, deemed suitable for participants who were all nonnative speakers of English. The final texts were produced in Word and audio-recorded for classroom use (see Table 4).

To answer Research Question 1 about the effect of three intervention treatments on vocabulary learning outcomes, it was essential to include words that would be unknown to the participants. Using nonce words would have been unethical because ethics clearance required intervention lessons to conform with the syllabus for English. We aimed for the chosen sets of target words to have similar item facility scores, mean frequencies, and mean word lengths (reference tool: Nation, 2012, 14K list). We piloted 90 infrequent words in a nonproject school in Grade 9 with 60 students. Analyzing these data, we identified items that were largely unknown (preferably with a mean facility score of  $< .15$ , that is, a very low mean proportion of pilot participants who knew a word; see Appendix S4 in the Supporting Information online). We considered nouns, verbs, and adjectives, reflecting their proportions in the English language, and then decided on the target items (see Appendix S4 in the Supporting Information online). We used *t* tests to compare the infrequent words for the intervention weeks. The comparison revealed that the words from the weeks were not significantly different from one another (Week 1 vs. Week 2,  $p = .152$ ; Week 1 vs. Week 3,  $p = .866$ ; Week 2 vs. Week 3,  $p = .200$ ).

Considering the age of participants, lesson time (110 min/week), and length of treatment, we included 20 target words per week, of which 12 were infrequent and therefore less likely to be known, whereas the remaining

eight were frequent, and therefore likely to be known by the participants. We included the eight frequent words in order not to discourage students when sitting the tests and to serve as a check that they were taking the tests seriously. We considered 20 words per week to be suitable based on suggestions in Schmitt and Schmitt (2020) and the research team's extensive secondary school teaching experience. Thus, in total 60 target words (36 infrequent + 24 frequent) were included, but our focus in the analysis and presentation of results was on the infrequent words.

The intention was to measure L2 vocabulary knowledge through meaning recall, that is, supplying the meaning when prompted by a L2 word form (Schmitt, 2010). In order to avoid a bias toward any of the languages known by participants, something that would have invalidated the assumed link between our treatment conditions and the word learning performance of the participants, the students were allowed to use any language when doing the tests. We decided to use a question/answer format that allowed for eight answer modes: supplying a word or an explanation in Swedish, English, or another language (six modes), or supplying a drawing or displaying vocabulary knowledge in any other way (two modes). We also chose to use dual-language instructions (English and the school language Swedish, see Figure 2). Finally, we consulted an international vocabulary expert on test formats (N. Schmitt, personal communication), asking for feedback on our suggested test approach.

We scored the vocabulary tests in a two-step process. First, two of the researchers scored answers to the test items as 0, 1, or 2 points. We gave a partial credit score of 1 when the answer conveyed semantic features close to those of the target words but lacked some precision. As a second step, several months later, we reviewed all of the scoring and compared scores across individual researchers to achieve consistency.

### *Classroom Procedures*

Structurally, there were two similar lessons each week of the intervention (see Table 5). The total weekly teaching time was 110 min (330 min over the three-week period). The first lesson introduced the Freedom theme. Each language condition was enforced by a laminated card called Rules of Engagement, placed on students' desks (every lesson) as a way to remind them of the current treatment (Figure 3).

Before we introduced the weekly theme, every student received a color-coded folder that included handouts of the week's text, word list, and vocabulary and text comprehension activities (Activities in Table 5). Although the texts were identical across all treatments (all content in English), the



**PRETEST**

# VOCABULARY VERSION 1

Name: ..... Class: .....

Please write your name and class above

*Vänligen skriv ditt namn och klass ovan*

- Please answer all questions to the best of your ability! For each question, there are two options.
- If you do not know a word, tick "I don't know this word" and move on to the next question.
- If you know a word, or if you think you know a word, please tick "I (think I) know this word". Then please show the meaning of the word by writing
  - a **translation** in Swedish or another language you know, or
  - an **explanation** in Swedish, English, or another language you know, or
  - a **synonym** in English.
- If you find it difficult to explain in words, you can **draw** the meaning of the word.

• *Vänligen svara på alla frågorna så gott du kan! För varje fråga har du två val.*

• *Om du inte kan ett ord, kryssa för "Jag kan inte detta ord" och gå vidare till nästa fråga.*

• *Om du kan ett ord, eller om du tror att du kan ett ord, vänligen kryssa för "Jag (tror att jag) kan detta ord".*

*Visa sedan att du förstår ordets betydelse till exempel genom*

- *att översätta det till svenska eller ett annat språk du kan, eller*
- *att skriva en kort förklaring på svenska, engelska, eller ett annat språk du kan, eller*
- *att skriva en synonym på engelska.*

• *Om du tycker det är svårt att förklara med ord kan du rita en teckning som visar på ordets betydelse.*

**Examples**

**a flute**

I don't know this word

I (think I) know this word: .....

**b house**

I don't know this word

I (think I) know this word: ...*hus / a building where you live...*

Figure 2 Vocabulary test instructions and question/answer format.

accompanying word lists differed according to condition (English, Swedish, and Swedish and other languages; see Appendices S5, S6, and S7 in the Supporting Information online for all wordlists). For the English condition (coded blue), the list contained 20 target items juxtaposed with definitions in English from the *Longman Dictionary of Contemporary English Online* (<https://www.ldoceonline.com>). For the Swedish condition (coded green), the same target items were listed but were matched with Swedish translation equivalents from the bilingual English-Swedish dictionary from

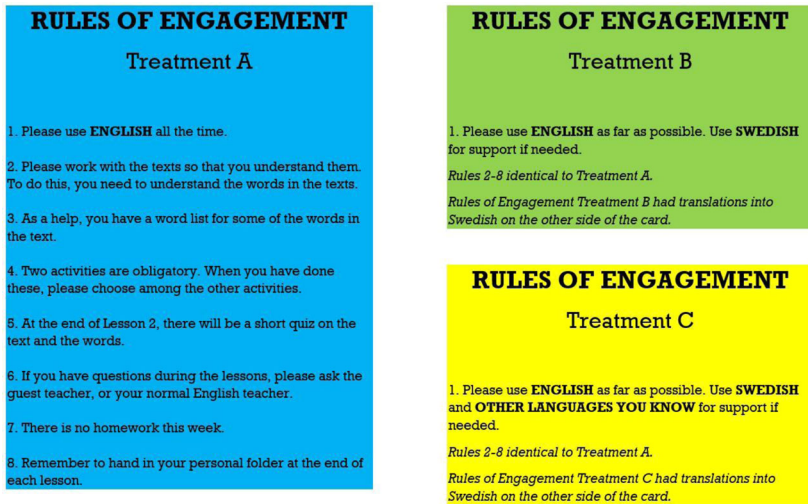
**Table 5** Lesson overview

Weekly lesson	Intervention Week 1 Lessons 1 & 2	Intervention Week 2 Lessons 3 & 4	Intervention Week 3 Lessons 5 & 6
Lesson 1	<ul style="list-style-type: none"> <li>• Intro. to intervention theme: Freedom</li> <li>• Week 1 theme: Martin Luther King, Jr.</li> <li>• Listening/reading Week 1 text + plenum teaching</li> <li>• Target vocabulary (1–20)</li> <li>• Activities</li> </ul>	<ul style="list-style-type: none"> <li>• Week 2 theme: Nelson Mandela and South Africa</li> <li>• Listening/reading Week 2 text + plenum teaching</li> <li>• Target vocabulary (21–40)</li> <li>• Activities</li> </ul>	<ul style="list-style-type: none"> <li>• Week 3 theme: Nadine Gordimer and South Africa</li> <li>• Listening/reading Week 3 text + plenum teaching</li> <li>• Target vocabulary (41–60)</li> <li>• Activities</li> </ul>
Lesson 2	<ul style="list-style-type: none"> <li>• Kahoot</li> <li>• Activities</li> <li>• Quiz (IMP1)</li> </ul>	<ul style="list-style-type: none"> <li>• Kahoot</li> <li>• Activities</li> <li>• Quiz (IMP2)</li> </ul>	<ul style="list-style-type: none"> <li>• Kahoot</li> <li>• Speaking activity</li> <li>• Activities</li> <li>• Quiz (IMP3)</li> </ul>

*Note.* In Intervention Week 1, all classes worked with one text on Martin Luther King, Jr. and the same set of words but under different conditions from those shown in Table 3. The same procedures applied for Week 2 and Week 3, that is, the same texts but different conditions. IMP = immediate posttest.

*Nationalencyklopedin.* For the Swedish and other languages condition (coded yellow), the target items were similarly listed, but juxtaposed with translation equivalents in the 31 different languages reported by the participants in the questionnaire (see Appendix S1 in the Supporting Information online).<sup>4</sup> The participants received lists of those languages represented in their particular class. The participants’ proficiency levels in these languages in all likelihood varied but were not assessed.

The purpose of the Activities handout was to engage students in intentional vocabulary learning tasks and in text comprehension. The activities handout for the English condition employed only English, whereas the corresponding handouts for the Swedish condition and the Swedish and other languages condition employed Swedish. Although translation equivalents in the 31 languages were provided in the word lists, all these translation equivalents were not used in the activities. This meant that the Swedish and other languages condition



**Figure 3** Color-coded Rules of Engagement cards.

handout was in Swedish, just as was the activities handout for the Swedish condition. The instructions for the activities differed, however, in line with each condition (for examples, see Appendix S8 in the Supporting Information online). The first two activities each week were obligatory, namely, Matching and Word Cards.

At the end of each lesson, the participants indicated which activity they had completed by ticking boxes on the cover sheet of the activity booklet. They were encouraged to work in pairs or small groups, which most students did. A few preferred to work alone and were allowed to do so. No homework was assigned, but all materials were available via the learning management platform, mainly so that absent students could catch up. Students' folders were collected after each lesson and were kept in a classroom cupboard.

The structure for each lesson was transformed into lesson plans by a member of the research team who had 10 years' experience of teaching English to this age group. Plans included time estimates, content details, and meta comments regarding instructions (see Appendix S9 in the Supporting Information online). The researchers acting as guest teachers used the same plan for all classes, and the pedagogical content was similar across intervention classrooms. We prepared specific listening/reading scripts to assure that the guest teachers engaged with each text in a similar fashion (texts and scripts cannot be made available for copyright reasons). Following listening and

reading, the focus was on the target vocabulary (choral speaking and repetition) before the students started engaging with the activities. The second lesson each week started with a Kahoot game (<https://kahoot.com/schools-u>) we had prepared, which included 20 multiple-choice items, each a target word with four pictures given. The students then worked with the activities until it was time for the quiz (i.e., the immediate posttest plus three reading comprehension questions; see Appendix S1 in the Supporting Information online). To facilitate student–student communication (in line with the syllabus for English) in Lesson 6, there was a speaking activity in small groups on the topic “What languages do you use outside school?” (see Appendix S11 in the Supporting Information online).

Overall, then, the procedures were controlled. Each target lexical item was repeated at least five times (listening/reading text, plenum teaching of word list, matching, word cards, Kahoot). For individual students, the number of repetitions was probably higher because they worked with the word cards in several rounds and engaged with more activities than the two obligatory ones.

### *Treatment Fidelity*

In intervention research, treatment fidelity has been defined as “the strategies that monitor and enhance the accuracy and consistency of an intervention to ensure it is implemented as planned and that each component is delivered in a comparable manner to all study participants over time” (Smith et al., 2007, p. 121). Given “the messy environment of teaching practice” (Marsden, 2007, p. 566), establishing treatment fidelity to the conditions was paramount. We achieved this by the guest teachers’ adhering to the lesson plans, to the scripts for teaching (see Appendix S9 in the Supporting Information online), and to the Rules of Engagement cards, identical across all classes. To assess treatment fidelity postintervention, two of the researchers listened to the audio recordings of all 36 intervention lessons, checking fidelity to the three treatments. The observation points during the fidelity-check were: (a) teacher vocabulary explanations, (b) teacher language use, and (c) verbal reminders of the applicable Rules of Engagement. Although the treatment fidelity check was, by necessity, qualitative in nature, we found no deviations from the treatment conditions for vocabulary explanations and vocabulary teaching in our thorough review of each recording.

### **Data Analysis**

We focused our analyses on the 36 ( $3 \times 12$ ) infrequent words and based the analyses on data from 74 participants who attended all six intervention lessons

and completed all tests (see Figure 1). We computed descriptive statistics with confidence intervals and effects sizes following suggestions by Plonsky (2015). In addition, we used mixed-effects modeling for data analysis; mixed-effects modeling provides many advantages over traditional analysis of variance, such as accounting for random variation in items and participants in one analysis, being robust against violations of homoscedasticity and sphericity, combining both random and fixed effects in the same analysis, and importantly, providing the capacity for nested random effects designs (Linck & Cunnings, 2015). We analyzed the data using the lme4 package (Version 1.1-28; Bates et al., 2015) in the R statistical software (R Core Team, 2021) together with the RStudio application (RStudio Team, 2022). We set alpha at .05 for all the analyses in our study.

## Results

### Class-Based Results

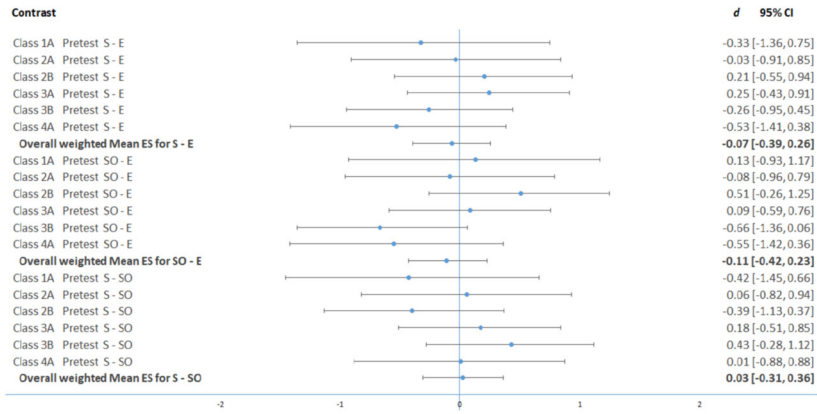
We found Cronbach's alpha reliability coefficients of internal consistency for the vocabulary test scores from the 74 participants to be satisfactory, with a range of .84–.92 for the infrequent words: pretest infrequent words ( $k = 36$ ),  $\alpha = .86$ ; immediate posttest Week 1 infrequent words ( $k = 12$ ),  $\alpha = .84$ ; immediate posttest Week 2 infrequent words ( $k = 12$ ),  $\alpha = .84$ ; immediate posttest Week 3 infrequent words ( $k = 12$ ),  $\alpha = .87$ ; and delayed posttest infrequent words ( $k = 36$ ),  $\alpha = .92$ .<sup>5</sup> These values were the same as or higher than the benchmark median reliability of .82 provided by Plonsky and Derrick (2016) and above the median of .79 for instruments in classroom settings. This was important because reliability coefficients at these levels indicated that the meaning–recall vocabulary knowledge construct was measured in a consistent way, which is a prerequisite for validity.

Table 6 reports the mean scores, standard deviations, and 95% confidence intervals for the pretest, immediate posttests, and delayed posttest for the infrequent words in all classes by conditions. Pretest scores were low for all classes, but as a general trend, we observed descriptively sizeable gains on the immediate posttests. Scores on the delayed posttest were distinctly lower than on the immediate posttests, except for Class 1A. On the immediate posttests, Classes 1A, 3B, and 4A had somewhat lower mean scores for the English condition compared to the Swedish and the Swedish and other languages conditions, whereas there was little difference between the conditions for Classes 2A, 2B, and 3A, especially for Classes 2A and 2B. For the scores on the delayed posttest, the same trends are apparent in Table 6, but they were less pronounced, and for Class 3B the means were close for the English and Swedish and other

**Table 6** Descriptive statistics for raw vocabulary scores for the 12 infrequent target words on the pretest and the immediate and delayed posttests by class and condition

Class	n	Condition	Pretest			Immediate posttest			Delayed posttest		
			M (SD)	95% CI		M (SD)	95% CI		M (SD)	95% CI	
1A	7	E	1.43 (1.90)	[0.02, 2.84]		12.43 (5.44)	[8.40, 16.46]		9.14 (8.73)	[2.68, 15.61]	
		S	0.86 (1.57)	[-0.31, 2.02]		17.43 (6.60)	[12.54, 22.32]		9.71 (8.83)	[3.18, 16.25]	
		SO	1.71 (2.36)	[-0.03, 3.46]		16.71 (7.34)	[11.28, 22.15]		8.29 (7.72)	[2.57, 14.00]	
2A	10	E	1.50 (3.41)	[-0.61, 3.61]		14.90 (6.54)	[10.85, 18.95]		2.70 (1.83)	[1.57, 3.83]	
		S	1.40 (2.67)	[-0.26, 3.06]		14.80 (7.13)	[10.38, 19.22]		2.40 (2.95)	[0.57, 4.23]	
		SO	1.20 (3.79)	[-1.15, 3.55]		13.90 (8.35)	[8.73, 19.07]		2.10 (2.02)	[0.84, 3.36]	
2B	14	E	0.50 (0.85)	[0.05, 0.95]		10.36 (7.25)	[6.56, 14.15]		2.79 (3.56)	[0.92, 4.65]	
		S	0.71 (1.20)	[0.08, 1.35]		11.86 (7.56)	[7.90, 15.82]		2.71 (2.37)	[1.47, 3.95]	
		SO	1.57 (2.85)	[0.08, 3.06]		12.14 (7.43)	[8.25, 16.04]		3.14 (3.21)	[1.46, 4.82]	
3A	17	E	2.71 (2.80)	[1.37, 4.04]		16.18 (3.56)	[14.49, 17.87]		6.53 (3.86)	[4.70, 8.36]	
		S	3.41 (2.90)	[2.04, 4.79]		18.00 (4.61)	[15.81, 20.19]		8.88 (5.10)	[6.46, 11.31]	
		SO	2.94 (2.41)	[1.80, 4.09]		16.12 (5.25)	[13.62, 18.62]		6.59 (3.10)	[5.11, 8.06]	
3B	16	E	1.81 (2.40)	[0.64, 2.99]		9.88 (6.14)	[6.87, 12.88]		3.00 (2.73)	[1.66, 4.34]	
		S	1.25 (1.95)	[0.29, 2.21]		15.13 (6.32)	[12.03, 18.22]		6.06 (4.20)	[4.00, 8.12]	
		SO	0.56 (1.15)	[0.00, 1.13]		15.63 (5.58)	[12.89, 18.36]		2.88 (2.70)	[1.55, 4.20]	
4A	10	E	2.50 (2.88)	[0.72, 4.28]		10.60 (5.72)	[7.06, 14.14]		5.30 (2.50)	[3.75, 6.85]	
		S	1.20 (1.93)	[0.00, 2.40]		16.10 (5.63)	[12.61, 19.59]		6.30 (5.52)	[2.88, 9.72]	
		SO	1.20 (1.69)	[0.15, 2.25]		19.00 (5.21)	[15.77, 22.23]		7.60 (4.81)	[4.62, 10.58]	
All	74	E	1.78 (2.53)	[1.20, 2.37]		12.43 (6.20)	[11.00, 13.90]		4.62 (4.37)	[3.61, 5.63]	
		S	1.62 (2.33)	[1.08, 2.16]		15.47 (6.43)	[14.00, 17.00]		5.96 (5.35)	[4.72, 7.20]	
		SO	1.58 (2.50)	[1.00, 2.16]		15.40 (6.56)	[13.90, 16.90]		4.82 (4.36)	[3.81, 5.83]	

Note. Maximum score = 24. Scores on each item were 0, 1, or 2. E = English; S = Swedish; SO = Swedish and other languages.

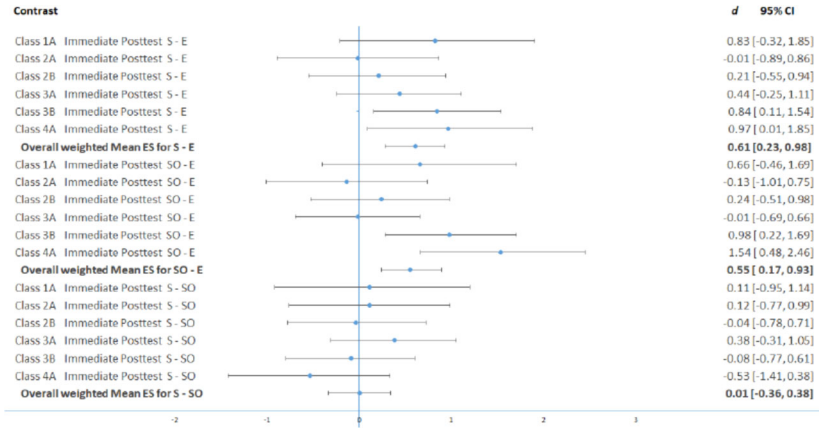


**Figure 4** Pretest forest plot of effect sizes across participating classes for the effect of pairwise condition contrasts on vocabulary scores (12 infrequent words). For each comparison, the figure reports mean difference and plots raw effect sizes with 95% confidence intervals by class. The overall weighted mean effect for each comparison is also plotted with its 95% confidence intervals (in boldface). ES = effect size; S = Swedish; E = English; SO = Swedish and other languages.

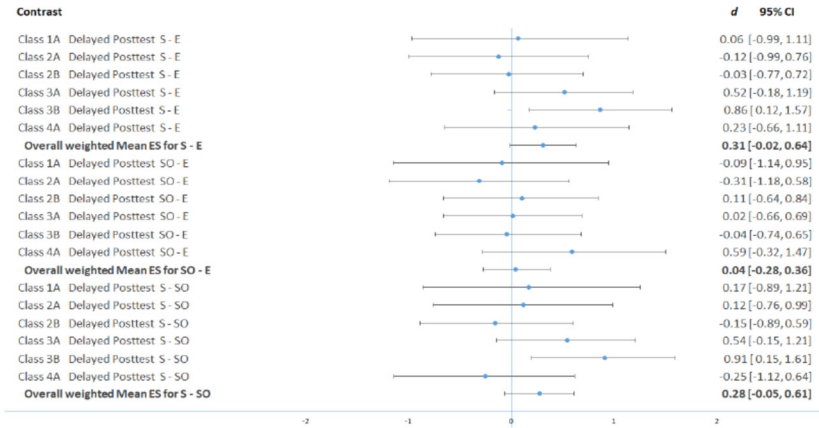
languages conditions. The delayed posttest performance by Class 2A was distinctly lower than that of most other classes and was somewhat puzzling.

**Effect Sizes Across Classes**

To measure the general effect of each of the three experimental conditions compared to one another, we calculated effect sizes through Cohen’s *d* and its 95% confidence interval for each class for the three test times: pretest, immediate posttest, and delayed posttest. Subsequently, we also calculated the overall weighted mean effect size for these site-specific effect sizes based on inverse variance, which included the precision of the effect estimates. We have presented the results of these calculations as forest plots in Figures 4, 5, and 6. We followed Plonsky and Oswald’s (2014) field averages for effect sizes for within-group comparisons in L2 research, where  $0.60 \leq d < 1.00$  suggests a small effect,  $1.00 \leq d < 1.40$  indicates a medium effect, and  $d \geq 1.40$  implies a large effect. For the confidence intervals of the effect sizes, we interpreted those entirely on the positive side of zero as showing a statistically significant positive effect, whereas we deemed those entirely on the negative side of zero as showing a statistically significant negative effect. We interpreted effect sizes with 95% confidence intervals that did not include zero as indicating significant effects (Cumming & Finch, 2005).



**Figure 5** Immediate posttest forest plot of effect sizes across participating classes for the effect of pairwise condition contrasts on vocabulary scores (12 infrequent words). For each comparison, the figure reports mean difference and plots raw effect sizes with 95% confidence intervals by class. The overall weighted mean effect for each comparison is also plotted with its 95% confidence intervals (in bold). ES = effect size; S = Swedish; E = English; SO = Swedish and other languages.



**Figure 6** Delayed posttest forest plot of effect sizes across participating classes for the effect of pairwise condition contrasts on vocabulary scores (12 infrequent words). For each comparison, the figure reports mean difference and plots raw effect sizes with 95% confidence intervals by class. The overall weighted mean effect for each comparison is also plotted with its 95% confidence intervals (in bold). ES = effect size; S = Swedish; E = English; SO = Condition Swedish and other languages.



Figure 4 shows the overall mean effect sizes for the differences between the three conditions at pretest were  $-0.07$  (between Swedish and English),  $-0.10$  (between Swedish and other languages and English), and  $0.03$  (between Swedish and Swedish and other languages), all minimal or negligible effects, with confidence intervals that included zero, suggesting no mean differences. Figure 5 shows that the overall mean effect sizes for the differences between the three conditions at the immediate posttests were  $0.60$  (between Swedish and English, a small effect),  $0.55$  (Swedish and other languages and English, a small to very small effect), and  $0.01$  (Swedish and Swedish and other languages, a negligible effect). Importantly, the confidence intervals for Swedish versus English and for Swedish and other languages versus English did not include zero, which we interpreted as the population mean residing within the confidence interval range in those two cases. Figure 6 shows that the overall mean effect sizes for the differences between the three conditions at the delayed posttest were  $0.31$  (between Swedish and English, a very small effect),  $0.04$  (between Swedish and other languages and English, a negligible effect), and  $0.28$  (between Swedish and Swedish and other languages, a very small effect). All the confidence intervals included zero, suggesting no mean differences. In sum, there was a small overall weighted mean effect size of the Swedish condition versus the English condition in the immediate posttest across the six classes at the four schools and an effect size approaching a small overall weighted mean effect ( $0.55$ ) of the Swedish and other languages condition versus the English condition likewise in the immediate posttest, but no other overall effect sizes reached the threshold for a small effect ( $\geq .60$ ).

### **Mixed-Effects Models for Condition Across Pretest, Immediate Posttest, and Delayed Posttest**

Next, we computed mixed-effects models to include several covariates. For these, test items were the unit of analysis rather than the mean scores from the individual students on the tests, thus increasing statistical power by raising the likelihood of detecting effects should they exist and allowing us to check whether there were differences between the words in each condition despite our efforts to make the word sets as similar as possible. Because we had scored our items as 0, 1, or 2 (i.e., partial credit scoring), we employed a standard linear model (rather than a binomial logistic analysis). In the model, we analyzed score differences for the 12 infrequent words from each of the three weeks for the variable test (three levels: pretest, immediate posttests, and delayed posttest), and condition (three levels: English, Swedish, and Swedish and other languages). Thus, the main fixed effects were test and condition,

and their interaction. For random effects, we specified random hierarchical relations (nested) for the Participant  $\times$  Class  $\times$  School interaction and for the Item  $\times$  Test  $\times$  Condition interaction, all with random intercepts. Even though we had aimed at having sets of items to be as similar as possible, we wanted to control for any variance that item nested in test and condition could yield. Further specification of random slopes led to failures to converge. Because target language proficiency had been a variable in previous studies that we had reviewed, we added English proficiency as a covariate as well as English grades (six levels) to see if they differed. We also included Swedish grades (six levels) as a covariate because Swedish was the school language and we could not rule out a potential effect. We centered all the added covariates. The model yielded 7,992 observations.

As a first step, we compared the model containing language background as a predictor covariate to a model without this covariate. An analysis of variance showed that the Akaike information criterion values were identical, so in the name of parsimony, we excluded the language background covariate (see Appendix S12 in the Supporting Information online for descriptive results with comments). Table 7 shows the full output for the model.<sup>6</sup> The reference categories (Intercept) were English for the variable condition and pretest for the variable test. There were no significant main effects for condition, but we found main effects for test, specifically for the levels immediate posttest and delayed posttest. Releveling showed no significant effect comparing these two testing times,  $t(316.65) = -0.02, p = .981$ . We also found a significant effect for English proficiency. This effect indicated that the participants with higher levels of proficiency scored higher on the meaning recall tests. There were also two-way interaction effects for the variables condition and test, specifically for the levels Swedish and Swedish and other languages and immediate posttest. Due to the observed interactions indicating that scores varied for test (pretest, immediate posttest, delayed posttest) and condition (English, Swedish, Swedish and other languages), we computed a series of post hoc pairwise comparison tests with the emmeans package in R (Lenth et al., 2021), using Tukey adjustments for multiple comparisons. The emmeans procedure uses least square means (i.e., group means adjusted for means of other variables in the model).

Table 8 shows the pairwise comparisons grouped by contrasts between conditions within each test (see Appendix S12 in the Supporting Information online for all comparisons). There were no significant differences between the condition means in the pretest nor in the delayed posttest. However, in the immediate posttest differences were significant between the English and the Swedish conditions as well as between the English and the Swedish and

**Table 7** Mixed-effects model for scores on the 12 infrequent words on pretest, immediate posttest, and delayed posttest ( $N = 74$ )

Parameter	<i>b</i>	95% CI	<i>SE</i>	<i>t</i>	<i>df</i>	<i>p</i> ( $>  t $ )
<b>Fixed effects</b>						
(Intercept)	0.16	[0.04, 0.27]	0.06	2.58	368.04	.010
Swedish	-0.03	[-0.18, 0.12]	0.08	-0.37	314.93	.713
Swedish and other	-0.01	[-0.16, 0.14]	0.08	-0.10	314.80	.917
Immediate posttest	0.89	[0.74, 1.04]	0.08	11.49	312.93	< .001
Delayed posttest	0.24	[0.09, 0.39]	0.08	3.07	312.93	.002
Grade Swedish	0.02	[-0.03, 0.07]	0.03	0.79	75.27	.432
Grade English	0.05	[-0.01, 0.11]	0.03	1.52	79.31	.132
Proficiency English	0.02	[0.01, 0.03]	0.01	3.68	67.93	< .001
Swedish × Immediate posttest	0.28	[0.07, 0.49]	0.11	2.54	314.77	.011
Swedish and Other × Immediate posttest	0.26	[0.04, 0.47]	0.11	2.34	314.63	.020
<b>Random effects</b>						
Item × Test × Condition	0.09	0.30				
Participant × Class × School	0.05	0.21				
Residual	0.40	0.63				

Variance *SD*

*Note.* Condition: pretest, immediate posttest, and delayed posttest; reference categories: English for the variable condition, pretest for the variable test; estimation method: restricted maximum likelihood.  $R^2_{\text{marginal}} = .31$ ;  $R^2_{\text{conditional}} = .49$ .

**Table 8** Results of post hoc pairwise tests of estimated marginal mean scores on pretest, immediate posttest, and delayed posttest for the English (E), Swedish (S), and Swedish and other languages (SO) conditions ( $N = 74$ )

Contrast	<i>b</i>	95% CI	<i>SE</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
Pretest							
E vs. S	0.03	[-0.21, 0.27]	0.08	0.37	313	1.000	0.10
E vs. SO	0.01	[-0.23, 0.25]	0.08	0.10	313	1.000	0.03
S vs. SO	-0.02	[-0.26, 0.22]	0.08	-0.26	315	1.000	-0.07
Immediate posttest							
E vs. S	-0.25	[-0.49, -0.01]	0.08	-3.23	313	.037	-0.90
E vs. SO	-0.25	[-0.49, -0.01]	0.08	-3.20	313	.039	-0.89
S vs. SO	0.01	[-0.24, 0.25]	0.08	0.02	315	1.000	0.01
Delayed posttest							
E vs. S	-0.11	[-0.36, 0.13]	0.08	-1.45	313	.875	-0.41
E vs. SO	-0.02	[-0.27, 0.22]	0.08	-0.32	313	1.000	-0.09
S vs. SO	0.09	[-0.16, 0.33]	0.08	1.13	315	.969	0.32

other languages conditions. The Cohen's  $d$  effects sizes for the lower English condition means were both small to medium ( $-0.90$  and  $-0.89$ ). These differences matched those shown in Figure 2 for the overall weighted mean, even though the effect sizes were a bit lower in Figure 2 than in Table 8.

Table 9 shows the pairwise comparisons grouped by contrasts between tests within each condition (see Appendix S12 in the Supporting Information online for all comparisons). All the conditions but one—the English condition in the pretest compared to the delayed posttest—were significantly different from each other. The effect sizes for the differences between the pretest and the immediate posttests were large for all three conditions. For the differences between the pretest and the delayed posttests, the contrasts for the Swedish condition had a medium effect size ( $d = 1.36$ ), whereas the contrasts for the Swedish and other languages condition had a small effect size ( $d = 0.97$ ). In contrasts for the immediate posttest compared to the delayed posttest, the observed differences for the conditions were all associated with large effect sizes:  $d = -2.35$  for the English condition,  $d = -2.84$  for the Swedish condition, and  $d = -3.15$  for the Swedish and other languages condition.

In a final analysis, we looked at the participants' scores in response to the frequent words included in the intervention material to keep participants motivated by feeling that there were words that they knew and also as a way

**Table 9** Results of post hoc pairwise tests of estimated marginal mean scores on pretest, immediate posttest, and delayed posttest for the English, Swedish, and Swedish and other languages conditions ( $N = 74$ )

Contrast	<i>b</i>	95% CI	<i>SE</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>
<b>English</b>							
Pretest vs. Immediate posttest	-0.89	[-1.14, -0.65]	0.08	-11.49	311	< .001	-3.20
Pretest vs. Delayed posttest	-0.24	[-0.48, 0.00]	0.08	-3.07	311	.058	-0.86
Immediate posttest vs. Delayed posttest	0.65	[0.41, 0.90]	0.08	8.42	311	< .001	2.35
<b>Swedish</b>							
Pretest vs. Immediate posttest	-1.17	[-1.42, -0.93]	0.08	-15.04	315	< .001	-4.20
Pretest vs. Delayed posttest	-0.38	[-0.62, -0.14]	0.08	-4.88	315	< .001	-1.36
Immediate posttest vs. Delayed posttest	0.79	[0.55, 1.04]	0.08	10.16	315	< .001	2.84
<b>Swedish and other languages</b>							
Pretest vs. Immediate posttest	-1.15	[-1.39, -0.91]	0.08	-14.76	315	< .001	-4.12
Pretest vs. Delayed posttest	-0.27	[-0.52, -0.03]	0.08	-3.49	315	.016	-0.97
Immediate posttest vs. Delayed posttest	0.88	[0.63, 1.12]	0.08	11.27	315	< .001	3.15

of checking that they had taken the tests seriously (see Appendix S13 in the Supporting Information online for descriptive statistics for these data). The maximum score on each of the immediate posttest was 16, and scores were generally high, so that the frequent words appeared to have served their intended purpose.

## Discussion

In this multisite study, we investigated instructed vocabulary learning in multilingual, English as an additional language classrooms in Sweden. Specifically, we carried out an intervention in six secondary school English classrooms in four schools. In the intervention, we manipulated the provided meanings of target English words as a condition with three levels: English definitions, Swedish translation equivalents, and translation equivalents in Swedish and other languages. We used a pretest–treatment–immediate posttest–delayed posttest design and analyzed data through descriptive statistical analysis, employing overall weighted mean effect sizes and their 95% confidence intervals as well as mixed-effects modeling with nested random variables.

### Research Question 1

Research Question 1 asked what the relative effects were of the conditions of target-language English, school-language Swedish, and multilingual Swedish and other languages on students' word meaning recall knowledge of meaning definitions from vocabulary learning. Previous research has, with a few exceptions, predominantly found favorable outcomes for L1 translation equivalents over L2 definitions in L2 word learning, but few studies have investigated contexts involving learners with more than two languages.

The descriptive statistics (means, Cohen's  $d$  effect sizes, and their confidence intervals) from the six classes for the differences of word meaning recall mean scores between conditions showed that the overall weighted mean effect sizes for the pretest scores (Figure 4) for our three conditions were minuscule ( $-0.07$ ,  $-0.10$ , and  $0.03$ ) at baseline. At immediate posttest, however, there were differences with a small effect size ( $d = 0.60$ ) for the overall weighted mean effect between the English condition and the Swedish condition, and a small to very small effect size ( $d = 0.55$ ) for the overall weighted mean effect between the English condition and the Swedish and other languages condition. We observed no difference contrasting the Swedish and the Swedish and other languages conditions ( $d = 0.01$ ). The subsequent mixed effects model analysis corroborated these descriptive findings. The interaction between test and condition and the post hoc pairwise comparisons reported in Tables 7, 8, and 9 showed that performances were higher on the Swedish and Swedish and other languages conditions compared to the English condition for the immediate posttests. We observed Cohen  $d$  effect sizes of  $0.90$  and  $0.89$ , respectively, which were thus both small effects.

These results align with previous studies where a L1 translation equivalent condition was found to yield higher scores than a L2 meaning definition condition (Hirosh & Degani, 2021; Laufer & Shmueli, 1997; Lee & Macaro,

2013; Zhao & Macaro, 2016). The results are also in line with the meta-analysis reported in Lee and Lee (2022) on teachers' verbal lexical explanations, where L1 explanations yielded more vocabulary knowledge than L2 explanations ( $d = 0.59$ ), and with Yanagisawa et al.'s (2020) meta-analysis of studies on the effect of glossing on vocabulary learning from reading in a L2, where L1 glossing trumped L2 glossing. In another meta-analysis, Kim et al. (2020) reported similar results on glossing. Yanagisawa et al.'s (2020) conclusion that their findings "indicate that unknown target words are more easily learned in glosses with L1 translations compared to L2 definitions or synonyms in general" (p. 431) matched the outcome in our study. Yanagisawa et al. also made a reference to the predictions of the revised hierarchical model (Kroll & Stewart, 1994), which we think is a relevant observation. However, we need to be cautious about claiming similarity with Yanagisawa et al.'s study because our study did not use glosses per se, and because ours was not a study of incidental vocabulary learning from reading. The glossing type in Yanagisawa et al.'s study that was closest in type to our use was glossary, which only featured in three out of 89 studies that they compared for glossing type.

How can the L1 translation equivalent advantage partially observed in our data be explained? Compared to target language definitions, a translation equivalent is a fast link to word meaning, taking less time both to process and to read compared to L2 meaning definitions. As we stated previously, although L2 meaning definition use is reportedly popular, especially with advanced learners, vocabulary scholars champion L1 use for L2 word learning (Nation, 2013; Schmitt & Schmitt, 2020), arguing that translation equivalents expedite creation of initial form–meaning links. Laufer and Shmueli (1997) argued that the superiority of L1 glosses in their study was due to maximum attention put on the new L2 word "since the L1 equivalent is fully familiar to the learner and consists of only one word" (p. 103). Learners in our study echoed similar preferences in the posttreatment questionnaire: "It was good to get the translation into Swedish as this helped making the connection stronger than when getting it in English" (Learner 444112); "It's been fun because I got to learn many new English words and in addition there were translations into my own language" (Learner 335262). Another learner wrote: "If there are only explanations in English it gets harder to remember the words and what explanations fit to which word" (Learner 444132). This is interesting in the light of Hummel's (2010) study where higher scores were reported for the rote copy condition that involved exposure to the L2 English target word and their French translation equivalents, compared to the more contextually rich conditions asking participants to translate whole sentences including the L2

target word. Hummel argued that translating whole sentences including the L2 target word entailed cognitive processing overload, distracting from the L2 target word and its L1 equivalent. This may have been at play also for our participants.

Turning to the delayed posttest results, relative to the immediate posttest performance, we observed lower scores overall, as we reported in Table 6. The analysis of overall weighted mean effect sizes (Figure 6) yielded  $d$  values of 0.31 (English vs. Swedish), 0.04 (English vs. Swedish and other languages), and 0.28 (Swedish vs. Swedish and other languages), and indicated that none of them reached even a small effect size, and the 95% confidence intervals all included zero. Low scores on a delayed posttest were reported by Zhao and Macaro (2016), who had allocated just one week between their immediate and delayed posttests. The studies in Lee and Lee's (2022) meta-analysis ranged between one week up to four and a half weeks. These are considerably shorter intervals than in our study, where we administered the delayed posttest 8–10 weeks after the immediate posttest, an interval that is considerably longer than what Lee and Lee (2022) reported. For most of our learners, this time span was evidently too long to retain a moderate level of knowledge of the infrequent target vocabulary. The vocabulary learning literature has unanimously treated repetition as crucial (Webb & Nation, 2017) but has noted considerable variation in how many repetitions have been claimed to be needed for long-term memory retention. A distinction in this regard has been made between massed learning and spaced learning (Nakata, 2015), where massed learning implies learning concentrated into a single session, whereas spaced learning implies multiple learning episodes distributed over longer time periods, which has been considered superior for learning (e.g., Ellis, 1995; Nakata & Suzuki, 2019). The low scores in the delayed posttest could have been due to the type of spacing of repetitions. Although the words in our intervention were not processed in a single session, learning happened in two lessons over just a week and was, therefore, more akin to massed than to spaced learning. It must also be remembered that meaning recall is an advanced type of ability, and even though our scoring procedure gave partial credit, the participants did not demonstrate comparable scores as they did in the immediate posttests.

One result that deserves a brief comment is the Class 2A performance on the delayed posttest. Although scoring relatively similarly to other classes on the immediate posttest, the scores on the delayed test dropped conspicuously. One explanation from the researchers present in Class 2A was that students were keen on getting to watch some of the video footage from the intervention, and many students seemingly rushed through the test so as to start watching



video uptakes sooner. We thus believe that their performance, to some extent, did not accurately reflect their capacity.

### **Research Question 2**

We turn now to Research Question 2, which asked to what extent the meaning definition condition affected the participants' meaning recall knowledge and was moderated by English proficiency, English and Swedish grades as school subjects, and language background. Our study included a measure of English proficiency (scores on a 32-item multiple-choice cloze test) and also the participants' grades in their English and Swedish school subjects. Based on the mixed-effects model (see Table 7), neither English nor Swedish grades were significant covariates, but English proficiency was,  $t(67.93) = 3.68, p < .001$ . This means that the higher the participants' scores on the proficiency test, the higher their word meaning recall scores. Our literature review showed that previous research has produced inconclusive results on this topic, with no proficiency effect found in Tian and Macaro's (2012) study on teacher codeswitching into L1 versus L2 English definitions for L1 Chinese students. Similarly, in Yanagisawa et al.'s (2020) meta-analysis, L2 proficiency was not observed to moderate gloss language (L1 or L2). Lee and Macaro (2013), however, observed that the L1 Korean Grade 6 students benefitted more from teacher L1 use than did university-level students in their study, inferring that proficiency was a variable. It is somewhat surprising in our study that the grade in English as a school subject did not also come out as a significant covariate because it stands to reason that a language grade shares considerable variance with a proficiency test of that same language.

We incorporated the remaining covariate, language background, in our modeling because our study included learners with diverse linguistic backgrounds. We categorized the participants into three groups: L1 Swedish learners, bilingual learners, and multilingual learners, on the basis of their reported language histories in the student questionnaire. The inclusion of this covariate in an initial mixed-effects model rendered no significant effect, and as we reported in the Results section, a comparison between that model and a model without this covariate yielded identical Akaike information criterion values. We have reported the descriptive statistics for the vocabulary scores laid out in terms of language background rather than classes in Appendix S13 in the Supporting Information online, and refer the reader to further comments there. Irrespective of its role as a covariate, the observed results merit a discussion related to language background and linguistic diversity in the classes. Two classes whose scores followed similar patterns on the immediate posttest were

Classes 3B and 4A, with higher scores on the Swedish and Swedish and other languages conditions and a significantly lower score on the English condition words. What was striking was how different these classes were in terms of students' language profiles. In Class 4A, 60% of the students classified as L1 Swedish learners compared to only 6% in Class 3B. We predicted that learners of this category would perform similarly on the Swedish and Swedish and other languages treatments because both featured Swedish translation equivalents, and this is what we saw in Class 4A. The proportion of learners classified as bilingual learners (simultaneous bilinguals of Swedish and a heritage language) was 40% (four out of 10) in Class 4A, compared to 69% (11 out of 19) for Class 3B. Despite different profiles, we observed similar scores. Furthermore, if we had compared the results and language background proportions in Class 1A, where no L1 Swedish category students existed and that had 71% bilingual learners and 29% in the multilingual category, we would perhaps have expected this class to score higher in the Swedish and other languages and English treatments than in the Swedish treatment. However, their highest scores were in the Swedish and Swedish and other languages treatments in the immediate posttest and quite similar across the three conditions in the delayed posttest. We acknowledge that the class sample sizes were small, and caution is needed when taking stock of these data on a class-by-class level.

Another variable to consider is that our participants' responses to the different interventions may have been influenced by the approach used by their regular English teacher. Of our four intervention teachers (in six classes), three followed a Swedish approach, whereas the fourth, Anita (pseudonym), used the English approach. It is therefore interesting to see how Anita's two classes (2A and 2B) had comparatively higher scores on the English condition. It cannot be ruled out that the regular teacher's approach had an impact on the results for these two classes. We may also note that these two classes scored the lowest in the delayed posttest compared to the other classes.

Although we counterbalanced the order of the treatments to minimize such order effects (see Table 3), there was no counterbalancing in place for the lessons. Classes 3B and 4A scored the lowest on the English condition (their last condition), whereas Class 1A students, who also scored the lowest on the English treatment, had experienced this condition first. The remaining three classes scored relatively similarly across conditions. In sum, we suggest that no systematic influence of the order of treatments prevailed. Motivation may also have affected the results. Because lessons in the three intervention weeks followed the same structure, gradually worse performance over the three weeks may have occurred as the participants became fatigued with the same pattern.

On the other hand, the same structure may have benefitted students who prefer knowing what is to come each lesson. A slight practice effect could also have been at play through students' realizing how they would be tested at the end of each week.

### **Limitations and Future Directions**

In our classroom-based study, we attempted to balance ecological validity with desirable levels of experimental control, and we need to acknowledge the contextual complexity of the classroom and concomitant threats to generalizability (Baker & Wright, 2021). In terms of counterbalancing, the condition order was different for the classes, but not the order of the lessons. Furthermore, the intervention was relatively short, and there was participant attrition due to students who were sometimes absent from class and who, consequently, were excluded from the data analysis. In addition, the participants were not asked specifically about their level of literacy in their heritage language(s). These limitations also form the basis for future directions. Longer interventions would be welcome to provide empirical evidence for more long-term effects of different vocabulary learning conditions. For multilingual learners who are proficient in a heritage language, it would be desirable to assess their proficiency level rather than rely on self-reported data. In the future, designing studies to control also for order effects for the lesson content would be desirable to the extent that this is possible in similar classroom studies.

### **Conclusion**

We premised our multisite intervention study on the current situation in Sweden of learning L2 English in secondary school, that is, where there is a high degree of language background diversity among students and of variation among teachers as to the degree that they make use of Swedish and other languages in their teaching. Our results showed a mixed pattern where English word meaning recall in the immediate posttests by and large was higher in conditions where Swedish (the majority language of school and society) and other L1 translation equivalents were featured in the word meaning definition materials used in the intervention. This is in line with previous studies that have reported better results for L1 translation equivalent(s) compared to target-language synonyms and definitions. However, we observed no statistically significant differences in the delayed posttest administered 8–10 weeks after the intervention, which underscores the need for systematic repetition of target vocabulary for longer-term retention.

Final revised version accepted 12 July 2022

## Open Research Badges



This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at <http://www.iris-database.org>. All proprietary materials have been precisely identified in the manuscript.

## Notes

- 1 Rice and Tokowicz (2020) have recently suggested a modification to the original revised hierarchical model called revised hierarchical model-repetition elaboration retrieval, with new mechanisms introduced in the model in the form of repetition, elaboration, and retrieval that entrench the associative connections between L1 and L2 lexical representations.
- 2 A reviewer pointed out that Laufer and Shmueli (1997) used two different sets of 10 items for L2 word paraphrasing and L1 translations conditions and argued that the results might therefore simply be attributable to possible differences in item difficulty rather than to the independent variable (L2 word paraphrasing vs. L1 translations) per se.
- 3 The locations of the schools in different types of municipalities largely resembled the stratified random sampling of Swedish schools reported in Appendix 2 in Sundqvist et al. (2021). Thus, despite using a nonprobability sampling method in this study, it was reasonable to claim that our sample shared many characteristics with the target population.
- 4 The 31 languages were: Albanian, Arabic, Armenian, Bosnian, Croatian, Danish, Dari, Dutch, Farsi, French, German, Hindi, Italian, Chinese, Korean, Kurdish (Kurmanji), Kurdish (Sorani), Macedonian, Norwegian, Pashto, Polish, Portuguese, Punjabi, Russian, Somali, Spanish, Swahili, Tigrinya, Turkish, Ukrainian, and Urdu.
- 5 For the frequent words, the reliability coefficients were: pretest frequent words ( $k = 24$ ),  $\alpha = .90$ ; immediate posttest Weeks 1–3 frequent words ( $k = 24$ ),  $\alpha = .67$ ; delayed posttest frequent words ( $k = 24$ ),  $\alpha = .94$ . The somewhat lower value for the immediate posttests likely stemmed from lack of variance due to scores being at ceiling.
- 6 The model specifications in R were: `MODEL = lmer(score ~ condition * test + (1|participant:class:school) + (1|itemid:test:condition) + Grade_Swedish + Grade_English + Proficiency_English, data = Intervention_infrequentwords, na.action = na.exclude)`.

## References

- Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511732935>
- American Council on the Teaching of Foreign Languages. (n.d.). *ACTFL proficiency guidelines*. <https://www.actfl.org/resources/guiding-principles-language-learning/target-language>
- Amir, A., & Musk, N. (2013). Language policing: Micro-level language policy-in-progress in the foreign language classroom. *Classroom Discourse, 4*(2), 151–167. <https://doi.org/10.1080/19463014.2013.783500>
- Baker, C., & Wright, W. E. (2021). *Foundations of bilingual education and bilingualism* (7th ed.). Multilingual Matters.
- Barcroft, J. (2009). Strategies and performance in intentional L2 vocabulary learning. *Language Awareness, 18*(1), 74–89. <https://doi.org/10.1080/09658410802557535>
- Bartolotti, J., & Marian, V. (2017). Bilinguals' existing languages benefit vocabulary learning in a third language. *Language Learning, 67*(1), 110–140. <https://doi.org/10.1111/lang.12200>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511845307>
- Brevik, L. M. (2008). *Engelsk for ungdomstrinnet. VOICES in Time 3*. Cappelen Damm.
- Busse, V., Cenoz, J., Dalmann, N., & Rogge, F. (2020). Addressing linguistic diversity in the language classroom in a resource-oriented way: An intervention study with primary school children. *Language Learning, 70*(2), 382–419. <https://doi.org/10.1111/lang.12382>
- Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in translation: The influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition, 38*(3), 403–443. <https://doi.org/10.1017/S0272263115000492>
- Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. <http://www.coe.int/lang-cefr>
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 170. <https://doi.org/10.1037/0003-066X.60.2.170>
- Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning, 8*(2–3), 103–128. <https://doi.org/10.1080/0958822940080202>
- Elston-Güttler, K. E., & Williams, J. N. (2008). First language polysemy affects second language meaning interpretation: Evidence for activation of first language

- concepts during second language reading. *Second Language Research*, 24(2), 167–187. <https://doi.org/10.1177/0267658307086300>
- European Commission (2012). *First European survey on language competences: Final report*. Directorate-General for Education, Youth, Sport and Culture, Publications Office. <https://data.europa.eu/doi/10.2766/34160>
- Flesch, R. F. (1949). *The art of readable writing*. Harper.
- Hirosh, Z., & Degani, T. (2021). Novel word learning among bilinguals can be better through the (dominant) first language than through the second language. *Language Learning*, 71(4), 1044–1084. <https://doi.org/10.1111/lang.12457>
- Hopp, H., Kieseier, T., Vogelbacher, M., & Thoma, D. (2018). L1 effects in the early L3 acquisition of vocabulary and grammar. In A. Bonnet & P. Siemund (Eds.), *Foreign language education in multilingual classrooms* (pp. 305–330). John Benjamins. <https://doi.org/10.1075/hsl.7.14hop>
- Hopp, H., Vogelbacher, M., Kieseier, T., & Thoma, D. (2019). Bilingual advantages in early foreign language learning: Effects of the minority and the majority language. *Learning and Instruction*, 61, 99–110. <https://doi.org/10.1016/j.learninstruc.2019.02.001>
- Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, 19(2), 131–143. <https://doi.org/10.1017/S0272263197002015>
- Hult, F. M. (2012). English as a transcultural language in Swedish policy and practice. *TESOL Quarterly*, 46(2), 230–257. <https://doi.org/10.1002/tesq.19>
- Hult, F. M. (2017). More than a lingua franca: Functions of English in a globalised educational language policy. *Language, Culture and Curriculum*, 30(3), 265–282. <https://doi.org/10.1080/07908318.2017.1321008>
- Hummel, K. M. (2010). Translation and short-term L2 vocabulary retention: Hindrance or help? *Language Teaching Research*, 14(1), 61–74. <https://doi.org/10.1177/1362168809346497>
- Joyce, P. (2018). L2 vocabulary learning and testing: The use of L1 translation versus L2 definition. *The Language Learning Journal*, 46(3), 217–227. <https://doi.org/10.1080/09571736.2015.1028088>
- Kim, H. S., Lee, J. H., & Lee, H. (2020). The relative effects of L1 and L2 glosses on L2 learning: A meta-analysis. *Language Teaching Research*, Advance online publication. <https://doi.org/10.1177/1362168820981394>
- Ko, M. H. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly*, 46(1), 56–79. <https://doi.org/10.1002/tesq.3>
- Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2), 149–174. <https://doi.org/10.1006/jmla.1994.1008>

- Källkvist, M. (2013). Languageing in translation tasks used in a university setting: Particular potential for student agency? *The Modern Language Journal*, 97(1), 217–238. <https://doi.org/10.1111/j.1540-4781.2013.01430.x>
- Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, 28(1), 89–108. <https://doi.org/10.1177/003368829702800106>
- Lee, J. H., & Lee, H. (2022). Teachers' verbal lexical explanation for second language vocabulary learning: A meta-analysis. *Language Learning*, 72(2), 576–612. <https://doi.org/10.1111/lang.12493>
- Lee, J. H., & Macaro, E. (2013). Investigating age in the use of L1 or English-only instruction: Vocabulary acquisition by Korean EFL learners. *The Modern Language Journal*, 97(4), 887–901. <https://doi.org/10.1111/j.1540-4781.2013.12044.x>
- Lenth, R. V., Buurkner, P., Herve, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2021). *Estimated marginal means aka least-square means* (Version 1.7.3) [Computer software]. <https://cran.r-project.org/web/packages/emmeans/index.html>
- Leung, C., & Valdés, G. (2019). Translanguaging and the transdisciplinary framework for language teaching and learning in a multilingual world. *The Modern Language Journal*, 103(2), 348–370. <https://doi.org/10.1111/modl.12568>
- Levine, G. S. (2003). Student and instructor beliefs and attitudes about target language use, first language use, and anxiety: Report of a questionnaire study. *The Modern Language Journal*, 87(3), 343–364. <https://doi.org/10.1111/1540-4781.00194>
- Liao, P. (2006). EFL learners' beliefs about and strategy use of translation in English learning. *RELC Journal*, 37(2), 191–215. <https://doi.org/10.1177/0033688206067428>
- Linck, J. A., & Cummings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, 65(S1), 185–207. <https://doi.org/10.1111/lang.12117>
- Lotto, L., & De Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, 48(1), 31–69. <https://doi.org/10.1111/1467-9922.00032>
- Marsden, E. (2007). Can educational experiments both test a theory and inform practice? *British Educational Research Journal*, 33(4), 565–588. <https://doi.org/10.1080/01411920701434094>
- Moranski, K., & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning*, 71(1), 204–242. <https://doi.org/10.1111/lang.12434>
- Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, 37(4), 677–711. <https://doi.org/10.1017/S0272263114000825>

- Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, 41(2), 287–311. <https://doi.org/10.1017/S0272263118000219>
- Nation, P. (2012). *Vocabulary size test instructions and description*. <https://www.victoria.ac.nz/lals/about/staff/paul-nation>
- Nation, P. (2013). *Learning vocabulary in another language* (2 ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139858656>
- Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics, and effect sizes: A “back-to-basics” approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). Routledge.
- Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, 100(2), 538–553. <http://doi.org/10.1111/modl.12335>
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Prince, P. (1996). Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The Modern Language Journal*, 80(4), 478–493. <https://doi.org/10.1111/j.1540-4781.1996.tb05468.x>
- R Core Team. (2021). R: A language and environment for statistical computing (Version 4.1.2) [Computer software]. R Foundation for Statistical Computing. <http://www.r-project.org>
- Rice, C., & Tokowicz, N. (2020). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition*, 42(2), 439–470. <https://doi.org/10.1017/S0272263119000500>
- RStudio Team. (2022). RStudio: Integrated development for R (Version 2021.09.2.382). RStudio, PBC. <http://www.rstudio.com>
- Sato, M., & Loewen, S. (2019). Methodological strengths, challenges, and joys of classroom-based quasi-experimental research. In R. M. DeKeyser & G. P. Botana (Eds.), *Doing SLA research with implications for the classroom: Reconciling methodological demands and pedagogical applicability* (pp. 31–54). John Benjamins. <https://doi.org/10.1075/llt.52.03sat>
- Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 199–227). Cambridge University Press.
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan. <https://doi.org/10.1057/9780230293977>
- Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching* (2 ed.). Cambridge University Press.
- Shin, J., Dixon, L. Q., & Choi, Y. (2019). An updated review on use of L1 in foreign language classrooms. *Journal of Multilingual and Multicultural Development*, 41(5), 406–419. <https://doi.org/10.1080/01434632.2019.1684928>



- Smith, S. W., Daunic, A. P., & Taylor, G. G. (2007). Treatment fidelity in applied educational research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education & Treatment of Children*, 30(4), 121–134. <https://doi.org/10.1353/etc.2007.0033>
- Spada, N. (2005). Conditions and challenges in developing school-based SLA research programs. *The Modern Language Journal*, 89(3), 328–338. <https://doi.org/10.1111/j.1540-4781.2005.00308.x>
- Spada, N. (2019). Methodological rigor and pedagogical relevance. In R. M. DeKeyser & G. P. Botana (Eds.), *Doing SLA research with implications for the classroom: Reconciling methodological demands and pedagogical applicability* (pp. 201–215). John Benjamins. <https://doi.org/10.1075/llt.52.10spa>
- Statistics Sweden. (2019). *Resursfördelningsmodell för fördelning av statsbidrag till huvudmän HT19* [Resource-allocation model for the distribution of state funding to responsible school organization]. Statistics Sweden. <https://scb.se/en>
- Sunderman, G., & Kroll, J. F. (2006). First language activation during second language lexical processing: An investigation of lexical form, meaning, and grammatical class. *Studies in Second Language Acquisition*, 28(3), 387–422. <https://doi.org/10.1017/S0272263106060177>
- Sundqvist, P., Gyllstad, H., Källkvist, M., & Sandlund, E. (2021). Mapping teacher beliefs and practices about multilingualism: The development of the MultiBAP questionnaire. In P. Juvonen & M. Källkvist (Eds.), *Pedagogical translanguaging: Theoretical, methodological and empirical perspectives* (pp. 56–75). Multilingual Matters. <https://zenodo.org/record/5269102#>. YUMVHp0zaUm
- Sundqvist, P., Källkvist, M., Gyllstad, H., & Sandlund, E. (2018, February 1–3). *Language practices and ideologies among English teachers in Sweden* [Paper presentation]. Language, Identity and Education in Multilingual Contexts (LIEMC18) Conference, Dublin, Ireland.
- Swedish National Agency for Education (2013). *English. Ämnesprov, läsår 2012/2013. Lärarinformation inklusive bedömningsanvisningar till Delprov A. Årskurs 9* [English. National test in 2012/2013. Information to teachers, including guidelines for assessment for Test Component A. Grade 9]. Stockholm: Swedish National Agency for Education.
- Swedish National Agency for Education (2018). *Curriculum for the compulsory school, preschool class and school-age educare (Revised 2018)*. <https://www.skolverket.se/publikationsserier/styrdokument/2018/curriculum-for-the-compulsory-school-preschool-class-and-school-age-educare-revised-2018>
- Swedish National Agency for Education (2021a). *Elever och skolenheter i grundskolan läsåret 2020/2021* [Students and schools in compulsory school 2020/2021]. <https://www.skolverket.se/getFile?file=7920>
- Swedish National Agency for Education (2021b). *Statsbidrag för stärkt likvärdighet och kunskapsutveckling* [State funding for enhanced equivalence and knowledge

- development]. <https://www.skolverket.se/download/18.70f8d1a017495c3cb5913b0/1603700418873/Lista%20%C3%B6ver%20skolors%20socioekonomiska%20index%202021.pdf>
- Tian, L., & Macaro, E. (2012). Comparing the effect of teacher codeswitching with English-only explanations on the vocabulary acquisition of Chinese university students: A lexical focus-on-form study. *Language Teaching Research, 16*(3), 367–391. <https://doi.org/10.1177/1362168812436909>
- Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.
- Xu, X. (2010). The effects of glosses on incidental vocabulary acquisition in reading. *Journal of Language Teaching and Research, 1*(2), 117–120. <https://doi.org/10.4304/jltr.1.2.117-120>
- Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition, 42*(2), 411–438. <https://doi.org/10.1017/S0272263119000688>
- Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology, 10*(3), 85–101. <https://doi.org/10.10125/44076>
- Zhao, T., & Macaro, E. (2016). What works better for the learning of concrete and abstract words: Teachers' L1 use or L2-only explanations? *International Journal of Applied Linguistics, 26*(1), 75–98. <https://doi.org/10.1111/ijal.12080>

## Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

### Accessible Summary

**Appendix S1.** Student Questionnaire – Background.

**Appendix S2.** All Test Materials for the Intervention.

**Appendix S3.** Flesch Reading Ease Score.

**Appendix S4.** Target Words: Infrequent and Frequent.

**Appendix S5.** Word Lists Intervention Week 1.

**Appendix S6.** Word Lists Intervention Week 2.

**Appendix S7.** Word Lists Intervention Week 3.

**Appendix S8.** Activities Handout.

**Appendix S9.** Plans for Teaching Lessons 1–6.

**Appendix S10.** Speaking Activity Project Week 3.

**Appendix S11.** Student Questionnaire – Post Intervention.

**Appendix S12.** Descriptive Statistics for Language Background Groups.

**Appendix S13.** Descriptive Statistics for Scores on the Frequent Words.