International Journal of TESOL Studies (2026)
Vol. 8 (3) 183-203 https://doi.org/10.58304/ijts.260601

Article

Task-induced Involvement, Motivation and Second Language Vocabulary Learners: Replicating and Extending Hulstijn & Laufer (2001)

Paweł Szudarski*
Christine Muir
University of Nottingham, UK

Received: 1 February 2026 / Received in revised form: 8 May 2026 / Accepted: 15 May 2026 / Available online: 17 June 2026

Abstract
This study is a replication and extension of Hulstijn and Laufer (2001) and focuses on English learners' incidental learning of vocabulary in specific task conditions and factors that affect this learning. Using Laufer and Hulstijn's (2001) Involvement Load Hypothesis (ILH) as a basis for examining task effectiveness, we explore EFL learners' vocabulary learning in three task conditions: Reading, Reading Plus Fill-In, and Writing. Beyond the task effects, we extend the original design by examining learners' gains at different levels of lexical mastery, measuring their motivation and task engagement, and including interview-based qualitative findings exploring their experiences and perceptions of participation. Findings indicate that Reading was the least effective condition in terms of vocabulary learning, with Reading Plus and Writing resulting in similar lexical gains, only partially confirming Hulstijn and Laufer (2001). As regards task engagement, there were no statistical differences in participants self-reported scores across groups, but the dynamic nature of their engagement was evident, as was the role of their emotions. Based on these findings, we present a theoretical and methodological discussion of the ILH in terms of its use, interpretation and predictions of task effects on vocabulary learning. We also underline the value of replication research and call for more cross-disciplinary research approaches in this area, particularly in relation to learner-related variables.

Keywords
Replication research, vocabulary learning, task-induced involvement, task engagement, motivation, effectiveness of language tasks, interdisciplinary research

1 Introduction

The Involvement Load Hypothesis (ILH), first proposed by Laufer and Hulstijn (2001), has been an influential framework underpinning a considerable body of vocabulary research over the last two decades (for ILH meta-analyses, see Huang, Wilson & Eslami, 2012; Yanagisawa & Webb, 2021). Approaching

*Corresponding author. Email: pawel.szudarski@nottingham.ac.uk

vocabulary learning through the lens of target language tasks and their effectiveness, the ILH stipulates that task-induced involvement comprises three main components: Search and Evaluation as cognitive factors and Need as a motivational factor. By examining these components and, crucially, assuming that other factors in the learning process are equal (Laufer & Hulstijn, 2001), the ILH has been used to predict learners' success, or lack thereof, in acquiring new knowledge: tasks with higher involvement were hypothesised to lead to higher lexical gains than tasks with lower involvement levels.

Given the amount of ILH-related research published since Laufer and Hulstijn's (2001) article (e.g. Folse, 2006; Kim, 2008; Rott, 2012), their goal of stimulating scholarly discussion around task-induced involvement has certainly been achieved (for a systematic review, see Liu & Reynolds, 2022). Specifically, the conceptualisation of task-induced involvement as a cognitive-motivational mechanism has proved to be a useful and highly productive basis for explorations of vocabulary learning across task and input conditions. One good illustration of this is Nation and Webb's (2011) Technique Feature Analysis (TFA). As a refined version of the original ILH with additional criteria, the TFA was proposed to explore not only incidental but also intentional vocabulary development (Nation & Webb, 2011; Schmitt, in press). Critically, empirical work based on the ILH and TFA frameworks has been important not only theoretically but also pedagogically, with findings used to make classroom recommendations and support teachers in selecting tasks most conducive to vocabulary learning (e.g. Sudsa-Ard, 2023; Nguyen, 2025).

However, despite evidence confirming initial predictions of the ILH (e.g. Hulstijn & Laufer, 2001), subsequent studies have provided only partial support for the claim that higher involvement loads facilitate better learning (e.g. Rott, 2012). Further, empirical studies have tended to integrate focus on the ILH with other learning-related factors (e.g., role of repetition by Folse (2006) or the role of learners' metacognition by Teng and Zhang (2024)), exploring them in concert as impacting lexical gains. Indeed, Yanagisawa and Webb's (2021) meta-analysis of 42 empirical studies, while generally supportive of the ILH, indicated that task-induced involvement explained only between 15.0% and 5.1% of vocabulary learning on immediate and delayed post-tests, respectively. Their analysis demonstrated additionally varying levels of contributions of each component of the model, with Evaluation contributing the most and, significantly, Search not found to contribute to learning.

As Laufer (2020, p. 359) herself has reflected, comparing tasks in terms of involvement load "is not simple at all". Three important issues merit attention. The first relates to specific vocabulary tasks chosen, the operationalisation of ILH core components, and varying empirical research designs. The second pertains to the unique nature of individual learners, as situated in diverse learning contexts, and the *dynamic* interrelationships between person, context and other factors over time (Ushioda, 2009). Muir and Szudarski (2025) discuss these issues at length, focusing on the operationalisation of the Need/motivation component within the extant ILH research and reflecting on the confounding interpretation of findings. Muir and Szudarski (2025) further called for future studies to more centrally position language learners, with the aim of recognising that their motivation, emotions, and other psychological factors also affect vocabulary learning beyond and in tandem with task-induced involvement. Schmitt (in press) reflects similarly on the various iterations of the ILH and TFA frameworks, which he considers useful pedagogically (e.g., when developing teaching materials to facilitate vocabulary learning), but as capturing only one part of the learning process. Foregrounded here, thus, is the third key issue meriting attention; namely, that empirical studies have tended to equate the ILH with, or misunderstand it as, a broader model of vocabulary development rather than as a framework that captures *task-related* variance in learners' vocabulary development.

Responding to these points, rather than seeking to expand the ILH framework, the current replication study adopts Hulstijn and Laufer's (2001) original design with supplementary extension elements. Specifically, we felt that a close replication of Hulstijn and Laufer's findings was needed first to see 1) whether their results could be directly replicated in a new context, and 2) to establish a reference point for future developments in this research strand by, for instance, examining proposals such as the New TFA (Muir & Szudarski, 2025) or the ILH Plus (Yanagisawa & Webb, 2022).

Further, the present study is a response to the call for more replication studies in SLA (McManus, 2024; 2026). Doing so here facilitates the purposeful revisitation of prior ILH-based claims, allowing us through a replication design to reconsider, refine, and extend findings and methodological decisions (McManus, 2024) as related to ILH-inspired studies.

Finally, the current replication straddles an interdisciplinary perspective across the fields of vocabulary studies and language learning psychology, contributing not only to the body of ILH research but also extending it methodologically. Our extension elements include measures of learners' reported motivation and task engagement, and, rare in the field of vocabulary studies, a complimentary qualitative dataset. As such, our aims are to more broadly foreground the role of language learners, and to underline the importance of this variable in examining task-induced involvement and vocabulary learning.

2 Literature Review

In this literature review, we first briefly introduce Need, Search and Evaluation as the three core components of the ILH. Next, we overview summary findings of ILH research, beginning with Hulstijn and Laufer (2001), introducing its methodological details alongside the design of our own. Finally, in outlining the rationale for the current study, we draw on reviewed findings to underline the importance of replication research in this area.

2.1 Need, Search and Evaluation

In Laufer and Hulstijn's (2001) original articulation of the ILH, language learners' task-induced involvement consisted of three main elements: Need, Search and Evaluation. Evaluation and Search were cognitive dimensions of the framework, with the former referring to learners' choices of target language words and their meanings as required by the task, and the latter to learners finding the right meanings or forms of words. Need, as the motivational component, concerned the source of a learner's behaviour or intention to understand and use new words in specific language learning tasks. The ILH postulates that, when other factors are equal, words processed with a higher involvement load are retained better than words processed with lower involvement.

2.2 Summary of ILH findings

Hulstijn and Laufer (2001) was the first empirical examination of the theoretical claims of Laufer and Hulstijn (2001). With EFL participants from Israel and the Netherlands, the study tested the relative effectiveness of three experimental tasks: reading comprehension, reading comprehension plus fill-in-the-blanks, and writing a composition. These three tasks represented increasing involvement loads and therefore were predicted to differently support incidental vocabulary learning. Findings confirmed that the lexical gains of participants who completed the writing task significantly outperformed those in the other two conditions. However, a significant difference between the reading comprehension and the reading plus conditions (with the latter returning higher mean scores) was found only in Israel. Collectively, Hulstijn and Laufer (2001) was perceived as broadly supportive of the ILH's proposition that task-induced involvement predicted the effectiveness of specific tasks in facilitating vocabulary learning.

A large body of research has since further interrogated this proposition. However, methodological differences between individual studies have presented difficulties in making direct comparisons, this including, as already mentioned, adding new factors to research designs and, at times, treating the ILH as a broader model of vocabulary learning. Folse (2006), for instance, working with ESL learners of English in the US, compared the effectiveness of vocabulary learning in three task conditions (one fill-in-the-blanks exercise, three fill-in-the-blanks exercises, and one original-sentence-writing exercise) and included also time-on-task as an additional factor. With learners' vocabulary tested by means of the

vocabulary knowledge scale (Paribakht & Wesche, 1997), results indicated that the second condition (three fill-in exercises), was significantly more effective than the other two, pointing to the importance of multiple encounters and repeated engagement with target words. As regards time-on-task as a further potential factor affecting learning, a post-hoc analysis of students with equivalent amounts of time-on-task revealed that writing was largely as effective as completing three fill-in tasks containing the target items. This led Folse (2006: 288) to conclude that "writing original sentences is neither an effective nor efficient written exercise for students to do when the goal is L2 vocabulary growth and retention". The practical and theoretical implication of these findings is thus that time-on-task should be considered alongside the ILH's conceptualisation of task-induced involvement when evaluating specific tasks in relation to vocabulary learning. Gao et al. (2024) offer more recent findings further supporting this relationship.

Kim's (2008) study also considered Hulstijn and Laufer's (2001) design and provided only a partial replication of their findings. Across two experiments with ESL learners, and with time-on-task controlled, the first found that the writing task yielded significantly higher vocabulary scores than the reading and gap-fill conditions. Interestingly, the gap-fill group did not differ significantly from the reading group, mirroring the findings of the L1-Dutch sample in Hulstijn and Laufer (2001). Kim's (2008) second experiment compared writing a composition and writing sentences as different tasks with the same level of task-induced involvement and found both to be similarly effective in terms of enhancing lexical learning, here aligning with ILH predictions.

There have also been efforts to employ the ILH for examining vocabulary learning in languages other than English. Keating (2008), for instance, conducted a conceptual replication in a study of beginner-level learners of Spanish completing three tasks: reading comprehension, reading comprehension plus target word suppliance, and sentence writing. Methodologically, learners' lexical knowledge was tested through translations: a passive Spanish-English recall and an active English-Spanish recall. Time-on-task was also considered, with learners' scores on the translation tests converted to words learned per minute. Results confirmed that the best retention of vocabulary occurred in the sentence writing task, followed by reading plus gap-fill and reading comprehension. However, when time-on-task was taken into account, no task effects were evident, as the advantage of the reading plus gap-fill and sentence writing over reading comprehension did not hold. Thus, while broadly aligning with the predictions of the ILH, these findings, too, point to time-on-task as an important factor to consider in tandem.

One final study of relevance is Rott's (2012) approximate replication of Hulstijn and Laufer (2001), with German as the target language. The study compared the effectiveness of reading plus gloss, reading plus fill-in and an essay-writing task, measuring participants' vocabulary learning at both a receptive and productive level. Results indicated that the ILH was confirmed only partially: the writing condition led to significantly more lexical knowledge than the two reading plus conditions, but there was no difference between the effectiveness of the fill-in and gloss conditions. Interestingly, the same pattern of results was obtained in relation to both receptive and productive aspects of lexical mastery.

By way of summary, based on this literature review several important points are evident. First, it is fair to say that the large body of empirical work around the ILH is illustrative of its importance and impact. However, mixed findings also suggest that further developments are needed, one of which, as we argue, is to study the ILH's task orientation alongside a more explicit focus on learner (e.g. motivation, task engagement, existing language repertoire) as well as context of learning (e.g. cultural factors, educational level and environment). Not necessarily seeing these factors as belonging to the ILH itself, we regard them as key contributors to language acquisition more broadly.

Such an approach might facilitate the widening of empirical perspectives, leading investigations of vocabulary and language learning to adopt a broader lens theoretically and methodologically. Recognition of the complexity of the SLA process has been present since the inception of the ILH, captured by Laufer and Hulstijn's (2001) important, although often overlooked, caveat of 'other things being equal' in the original formulation of their hypothesis. Our replication aims to bring this key aspect

to the fore, highlighting the importance of the multiplicity of contributing factors, including word-, learner- and context-related, and thus more explicitly recognising the original conceptualisation of the ILH. To achieve this, the design adopted in the current study has sought to reflect a more interdisciplinary approach.

Second, the findings from the reviewed research suggest that caution is needed when interpreting the results of individual studies and making broader generalisations. This is especially true when direct comparisons of results are attempted, and when the ILH framework has been treated, inadvertently perhaps, as a larger model of vocabulary learning. Finally, and perhaps most importantly in the context of the present study, we see the above as collectively pointing to the need for more replication research around the ILH, particularly when findings are to inform pedagogical decisions.

3 Current Replication and Extension of Hulstijn & Laufer (2001)

3.1 Study design

Table 1 outlines full methodological details of Hulstijn and Laufer (2001) alongside the design of the current study, conducted with learners of English in Poland. In the table, we summarise study features, with the following narrative outlining our rationale for the extension added. The prose offers more detail concerning our decision-making rationales. Crucially, in designing this replication, our decisions were informed not only by the reviewed ILH-related literature but also by email correspondence with the authors of the original study. For instance, our correspondence with Batia Laufer resulted in stressing the focus of the ILH on task-related determinants of vocabulary learning and consequently treating the hypothesis as such, rather than as a broader model of vocabulary development encompassing all factors.

Time on task. In both studies, time-on-task is regarded as an inherent property of each of the three tasks. This is an important methodological consideration, as already discussed in the literature review. In our design, and as time-on-task is also important pedagogically, we followed Rott's (2012) replication design and shortened the teaching times devoted to each experimental task in order to 1) better represent typical class times given to such tasks, and 2) reflect participants' high proficiencies (see Participants section/VLT scores).

Experimental text & topic. The topic of the text (fake news) was comparable to the original but updated with an aim to increase relevance and interest for current participants. Written by the second author (an L1-English user), both the text and accompanying comprehension questions were piloted with another L1-English user and an English-Polish bilingual. Replicating Hulstijn and Laufer, the text required participants to process the target items to correctly answer the comprehension questions (Reading & Reading Plus conditions).

Extension elements. While replicating Hulstijn and Laufer's design as closely as possible, our intention was also to increase methodological rigour and incorporate several measures of relevant learner-related factors. Inclusion of these additional measures sought to maximise the validity of findings without compromising the design or introducing additional variance (for a discussion of methodological implications in designing and interpreting replication studies in SLA, see Szudarski & Mikołajczak, 2023). Finally, beyond the vocabulary, task engagement and motivation measures, our extension included also interviews with three participants, allowing for qualitative analysis of learners' experiences and perceptions of participation in the study.

Based on the above, and following McManus's (2024) categorization of replication research, we treat our study as a conceptual replication and extension of Hulstijn and Laufer (2001), with changes to the experimental text and time-on-task as replication elements and five additions as extension elements: meaning recognition test of target vocabulary; Vocabulary Levels Test; survey of learners' task engagement; survey of learners' motivational profiles; qualitative interviews with participants.

Table 1

Study Designs of Hulstijn & Laufer (2001) and Current Replication

Feature of study	Methodological modifications or additions	Hulstijn & Laufer (2001)	Current replication
Experimental conditions – three tasks	No	• **Reading**: Reading & 10 multiple choice comprehension questions. Target items highlighted in the text & glossed in Hebrew/Dutch & English • **Reading Plus**: Mirroring the Reading task, but with target words omitted from text and given to students (glossed in Hebrew/Dutch & English) with 5 distractors. Students first required to fill in the blanks before answering 10 multiple choice comprehension questions. • **Writing**: Students asked to write a letter using the same ten target items, here given as a list with L1 translations and in example sentences • Six intact groups of learners (three groups in each country) randomly assigned to each of the three conditions in each country	• Mirroring of task designs: Reading, Reading Plus, Writing • Minor adjustments accounting for participants' first languages: Reading & Reading Plus task words glossed in Polish and English, Writing task translations into Polish with example sentences in English (see Supplementary Materials) • Participants in four intact classes (two of these were small/both assigned to the same condition)
Target items	No	• Ten low-frequency items unlikely to be known by participants: *inflammatory, to curb, grist, morally derelict, deeply ingrained, wrath, to sanitise, privy to, not one whit, rigmarole*	• Mirrored target items & distractors, with a minor modification to one from *grist* to *grist to the mill* (word → phrase) to better reflect current usage
Experimental text & topic	Yes – modification	• A 621-word letter to the editor of a magazine in response to a recently passed bill in 1980s British Parliament negatively perceived by the writer as state censorship of the media and TV. Taken from the national reading comprehension exam system in the Netherlands.	• A blog post of 636 words 'written by a teenager' in response to an article about fake news, language use and negativity online, with the length and frequency of words in the text controlled for to attain appropriate lexical coverage.
Time-on-task	Yes – modification	• **Reading**: 40-45 minutes, • **Reading Plus**: 50-55 minutes • **Writing**: 70-80 minutes	• **Reading**: 15-20 minutes • **Reading Plus**: 20-25 minutes • **Writing**: 35 minutes

Feature of study	Methodological modifications or additions	Hulstijn & Laufer (2001)	Current replication
Extension elements	Yes – addition	N/A	• Meaning recognition test (immediate & delayed) • Vocabulary Levels Test/ VLT (Schmitt et al., 2001) • Survey of participants' motivation (Papi & Khajavy, 2021) • Survey of participants' task engagement (Zare & Derakhshan, 2024) • Qualitative interviews

3.2 Research questions

The study pursued the following research questions:

RQ1a. What is the effect of three language tasks with different involvement loads (Reading; Reading Plus; Writing) on participants' incidental learning of English vocabulary?

RQ1b. Does this effect vary depending on the aspect of lexical mastery considered (meaning recall vs. meaning recognition) and testing time (immediate vs. delayed)?

RQ2. How, if at all, does participants' reported task engagement vary across the three experimental conditions?

RQ3. How do participants reflect on the experimental tasks and on the broader experience of participating in this study?

3.3 Participants

Fifty-nine EFL university-level students participated in the current study. All were recruited from a Polish university and had Polish (n = 54) or Ukrainian (n = 5) as their L1. These students represented different levels of study, from both undergraduate and postgraduate courses, and all majored in philology/modern languages or linguistics. All were thus learning not only English but other foreign languages as well (including e.g. Russian, Spanish, Italian) and therefore should be considered multilingual.

To control for any L1 effects on vocabulary learning, L1-Ukrainian students were excluded from statistical analysis, as were students absent from the delayed post-test. This resulted in the final sample of 48 participants: Reading condition = 16, Reading Plus = 15, Writing = 17. One L1-Ukrainian participant (Reading condition; Interviewee 1) volunteered to participate in an interview, as did two L1-Polish participants (one each from Reading and Reading Plus conditions; Interviewees 2 & 3 respectively). Regrettably, no participants volunteered from the Writing condition.

Table 2 outlines the vocabulary and motivation profiles of all participants taken forward for quantitative analysis. Using the VLT 2K-5K levels as a proxy for proficiency, an ANOVA confirmed no significant differences between experimental groups. Similar proficiency levels were deemed necessary ahead of the main statistical analysis. Participants were familiar with English words at the level of 5K, thus confirming their advanced level and mirroring the level of participants in Hulstijn and Laufer (2001).

A similar pattern was also found for participants' motivational profiles. A series of ANOVAs did not indicate significant differences in any of the four future self-guides reported by participants across experimental groups (see Table 3). Looking at these descriptively, learners reported strong

ideal L2 selves/own, followed by an *ideal L2 self/other*, indicating a clear promotion focus. This is perhaps unsurprising considering the status of participants enrolled in tertiary linguistics programmes. We considered this important to explore, because different self-guides have been found to predict qualitatively different motivated behaviours and emotions. For example, an *ideal L2 self/own* has been found to positively predict an 'eager strategic inclination' in learners' behaviour (Papi et al., 2019), and to predict L2 enjoyment positively and negatively predict L2 anxiety (Papi & Khajavy, 2021). We return later in the paper to continue discussion of learner emotions.

Table 2

Vocabulary and Motivation Profiles of Participants Broken Down by Experimental Task (ns = non-significant)

Measure	Experimental group	Mean	SD	ANOVA
Vocabulary: Total VLT (max score = 90)		**84.31**	**6.12**	F (2, 45) = 1.30, p = .282 ns
	Reading	82.31	7.48	
	Reading Plus	85.40	3.70	
	Writing	85.23	6.29	
Motivation: Ideal L2 self/other		**3.62**	**.99**	F (2, 45) = 2.05, p = .141 ns
	Reading	3.83	.79	
	Reading Plus	3.82	.84	
	Writing	3.24	1.20	
Motivation: Ought to L2 self/Own		**3.17**	**1.11**	F (2, 45) = .059, p = .942 ns
	Reading	3.25	1.37	
	Reading Plus	3.12	1.01	
	Writing	3.15	1.00	
Motivation: Ideal L2 self/own		**4.43**	**.61**	F (2, 45) = .182, p = .834 ns
	Reading	4.45	.75	
	Reading Plus	4.49	.49	
	Writing	4.36	.58	
Motivation: Ought to L2 self/other		**2.00**	**.99**	F (2, 45) = .765, p = .471 ns
	Reading	2.23	1.01	
	Reading Plus	1.97	1.06	
	Writing	1.81	.92	

3.4 Instruments & data analysis

Table 3 presents study instruments and approaches to data analysis. In the following, we more fully introduce our extension elements and rationales for inclusion.

Table 3

Study Instruments & Data Analysis

Instrument	Details	Analysis
Schmitt et al.'s (2001) Vocabulary Levels Test (VLT)	• Three frequency levels included (2K, 3K & 5K, with academic vocabulary being beyond the focus of the study	• ANOVA conducted to explore participants' vocabulary size across the three experimental groups
Motivation questionnaire (See Supplementary Materials)	• Items taken from Papi & Khajavy (2021). • Scale reliabilities (Cronbach alpha): Ideal other: .804 / Ideal own: .821 / Ought other: .903 / Ought own: .837	• ANOVAs conducted to explore participants' motivation across the three experimental groups
Task engagement questionnaire (See Supplementary Materials)	• Items taken from Zare & Derakhshan (2024) • Scale reliabilities (Cronbach alpha): behavioural .578 / emotional .879 / cognitive .717; excluding item "*During the task, I repeated the contents and asked myself questions about them*" / agentive .817; excluding the item "*During the task, I expressed my preferences and opinions*" / social .739	• Kruskal-Wallis tests used to interrogate differences in participant ratings of engagement across experimental groups • Wilcoxon Signed Ranks tests used to explore patterns of engagement within experimental groups
Immediate & delayed post-tests (See Supplementary Materials)	• Meaning recall: Participants required to provide either a Polish translation or English explanation of the 10 target items & 15 distractors • Meaning recognition: 5-item multiple choice questions, including an 'I don't know' option, for each of the 10 target items + 15 distractors (words & phrases)	• Participants awarded one point for each correct answer • Two-way (3x2) mixed ANOVA, task type as between-subjects variable/time as within-subjects variable
Post-treatment check of learners' pre-knowledge of target items (See Supplementary Materials)	• Mirroring Hulstijn and Laufer's design, participants given a list of target words and asked to indicate which they knew in advance of the study	• Where participants indicated they had known an item prior to the experiment, a score of zero was awarded for the item on all immediate & delayed vocabulary post-tests
Qualitative interviews	• Semi-structured, individual interviews exploring participants' experiences of participation • Conducted in English by the first author, online via Microsoft Teams, three weeks after completion of post-test. Audio- or video-recorded dependent on participant preference. 45-60 minutes.	• Reflexive thematic analysis (Braun et al., 2023)

Vocabulary tests. (1) Hulstijn and Laufer's (2001) original design did not include measures of language proficiency. Including the VLT as a proxy for English proficiency allowed us to ensure parity across experimental groups. (2) Using multiple tests of vocabulary knowledge and tapping into its different aspects is now an established practice in vocabulary research (Laufer & Goldstein, 2004; Schmitt, 2010), facilitating richer and more detailed descriptions of language learning. The original design with a measure of meaning recall was therefore extended, following e.g. Rott (2012), with a meaning recognition test.

Motivation questionnaire. The importance of motivation to vocabulary learning is long recognised (e.g. Tseng & Schmitt, 2008). Conceptualising motivation as future self-guides (Dörnyei, 2005), here we looked to Papi et al.'s (2019) 2x2 model in which four future self-guides differ along two dimensions. The first of these captures regulatory distinctions, with the *ideal L2 self* having a promotion focus (representing L2 attributes that a learner seeks to achieve) and *ought to L2 self* a prevention focus (a desire to avoid negative consequences). The second dimension captures differing standpoints: *self* and *other*. For example, an *ideal L2 self/other* represents what we believe our parents or teachers (*other*) might want us to achieve (promotion), and an *ought to L2 self/own* represents L2 attributes we believe we must possess (*own*) in order to avoid negative consequences (prevention; see the Supplementary Materials for full instrument). Mirroring the use of the VLT as a proxy for proficiency, this questionnaire allowed us to explore any differences in participants' reported motivations across experimental groups.

Task engagement questionnaire. This inclusion was rooted in the desire for a complimentary measure that invited participants' own (self-reported) perceptions of the task to more centrally foreground them in the research process. The notion of *engagement* (Fredricks et al., 2004) has become increasingly influential in language learning studies, being used, for example, to examine relationships between different tasks and learners' target language vocabulary development (e.g. Hiver & Dao, 2025). We drew on the notion of *task engagement* (Philp & Duchesne, 2016) to 1) explore participants' reported behavioural, emotional, cognitive, agentic and social engagement across experimental tasks (Zare & Derakhshan, 2024; see the Supplementary Materials C for full instrument) and 2) to investigate how, if at all, the differing involvement loads required of each task might be reflected in participants' self-reported engagement.

Interviews. The majority of research on the ILH, and indeed in the field of vocabulary studies more broadly, is quantitative in nature (see e.g. Coxhead, 2025). Interviews allowed us to represent participants' voices, creating space for generating novel insights into our methodological decisions and learners' experiences of participation.

3.5 Procedures

Ethical approval for the study was gained from the University of Nottingham alongside formal agreement from the hosting university in Poland. Data was collected in person by the first author during participants' regular university tuition over two classes (each 90 minutes) spaced a week apart. The first class began with introduction to the study, after which all students gave informed consent. In this first class, participants completed a treatment and immediate post-test. In the second class, they completed the delayed post-test, VLT, a test checking pre-knowledge of target items, and the motivation and engagement questionnaires. Data collection in the second class ended in advance of the lesson, so the first author facilitated a discussion about conducting empirical research, important for this cohort as they looked ahead in future years to completing their own empirical dissertations. After this initial period of in-person data collection, participants were contacted by the first author via email and invited to take part in a follow-up interview. Three participants volunteered to participate; these were arranged and held online, approximately three weeks later.

4 Results

RQ1a: What is the effect of three language tasks with different involvement loads (Reading; Reading Plus; Writing) on participants' incidental learning of English vocabulary?

Participants' vocabulary learning was measured at two levels of lexical mastery: *meaning recall* and *meaning recognition*.

Table 4

Descriptive Statistics for Two Tests of Target Vocabulary (max score = 10 in both cases)

		Meaning recall		Meaning recognition	
Condition	**N**	**Immediate Post-test M (SD)**	**Delayed Post-test M (SD)**	**Immediate Post-test M (SD)**	**Delayed Post-test M (SD)**
Reading (R)	16	0.8 (0.9)	1.2 (1.3)	3.5 (2.3)	3.8 (2.7)
Reading Plus (RP)	15	3.7 (1.8)	3.6 (2.1)	5.7 (1.8)	5.2 (2.0)
Writing (W)	17	3.8 (1.6)	3.0 (1.5)	5.5 (1.8)	5.2 (1.6)

The mean scores suggest that the RP and W conditions led to better results compared to R, this being the case at both levels of mastery. Mirroring the analysis of Hulstijn and Laufer (2001), a two-way (3x2) mixed ANOVA was conducted, with task type as the between-subjects variable (R, RP & W) and time as the within-subjects variable (immediate & delayed; for descriptives see Table 4). Mauchly's test indicated that the assumption of sphericity was violated, so degrees of freedom were adjusted using the greenhouse-Geisser correction.

As regards *meaning recall*, the two-way mixed ANOVA revealed a significant effect of task type ($F(2, 45) = 16.8$, $p < .001$, eta = .16), no effect of time ($F(1, 45) = 7.14$, $p = .402$, eta = 4.27), and a significant interaction between task type and time ($F(2, 45) = 4.95$, $p = .011$, eta = .18), suggesting that changes in participants' scores over time depended on the task completed. As confirmed by post-hoc (Tukey) tests of learners' meaning recall scores, R was significantly lower than RP and W. This aligns with Hulstijn and Laufer (2001), where R was also the least effective task. However, we found no statistical difference in effectiveness between RP and W.

As for *meaning recognition*, the two way-mixed ANOVA analysis showed neither a significant effect of time ($F(1, 45) = .54$, $p = .464$, eta = .12) nor an interaction between task type and time ($F(2, 45) = 1.12$, $p = .336$, eta = .47). Crucially, however, there was a statistical effect of task type ($F(1, 45) = 4.43$, $p < .001$, eta = .47), with post-hoc (Tukey) tests mirroring the results of the recall test: R was statistically less effective than RP and W, with no statistical difference between the latter. Thus, RP and W tasks seemed to have been equally effective in enhancing learners' meaning recognition scores, although only on the immediate post-test and not extending to the delayed session.

RQ1b: Does this effect vary depending on the aspect of lexical mastery considered (meaning recall vs. meaning recognition) and testing time (immediate vs. delayed)?

Building on the significant interaction between task type and time, as well as the presence of two vocabulary tests, learners' results were further compared in terms of recall vs. recognition scores (see Table 5). Given the respective difficulty of these tests, meaning recognition scores were expected to be higher than those for meaning recall. Indeed, when compared statistically via Wilcoxon Signed Rank

tests, participants' recognition scores were significantly higher than their recall scores (large effect sizes), regardless of task condition.

Table 5

Meaning Recall vs. Meaning Recognition Results at Immediate and Delayed Post-test

	Immediate post-test: Recall vs. Recognition	**Delayed post-test: Recall vs. Recognition**
Reading (R)	0.8 (0.9)	1.2 (1.3)
	3.5 (2.3)	3.8 (2.7)
	z = -3.07, p = .002, r = .54 *	* z = -3.33, p < .001, r = .58 *
Reading Plus (RP)	3.7 (1.8)	3.6 (2.1)
	5.7 (1.8)	5.2 (2.0)
	z = -3.27, p = .001, r = .60 *	z = -2.47, p = .013, r = .45 *
Writing (W)	3.8 (1.6)	3.0 (1.5)
	5.5 (1.8)	5.2 (1.6)
	z = -3.22, p = .001, r = .55 *	z = -3.44, p < .001, r = .59 *

Note: * indicates statistical significance

Thus, depending on the aspect of knowledge measured, as well as the timing of testing, our analyses revealed important variation in learners' task-related vocabulary performance. This includes, unexpectedly, R's delayed post-test scores being higher compared to the immediate test, pointing to potential testing effects. Overall, such results suggest a complex set of relationships that go beyond task effects, including most notably the impact of measurement tools on the amount of learning shown.

RQ2: How, if at all, does participants' reported task engagement vary across three experimental conditions?

Data to answer RQ2 came from the Task Engagement Questionnaire. A summary of descriptive statistics can be found in Table 7, with the results broken down by engagement dimension and experimental task.

Seeking to examine learners' reported engagement across the three tasks, participants' mean scores on each dimension were submitted to a series of Kruskal-Wallis tests (non-parametric due to non-normal data distribution; see Table 7). No statistical differences were observed between the three groups: participants reported similar levels of engagement along each dimension across all experimental groups.

Next, we looked to explore patterns of engagement reported within each experimental group. Using a series of non-parametric Wilcoxon Signed Ranks tests (see Table 8), this analysis was conducted separately for R, RP, and W task conditions. Analysis revealed the same pattern in each of the three experimental groups. Specifically, participants' reported levels of agentic and social engagement were significantly lower than their self-reported levels of behavioural, emotional and cognitive engagement. There were no differences reported between participants' agentic and social engagement, nor their behavioural, emotional and cognitive engagement (see the Supplementary Materials for full questionnaire items and further definitional clarification of each of these aspects of engagement).

Table 7

Descriptive Statistics for Task Engagement in Three Experimental Groups

Scale		M	SD	Kruskal-Wallis tests
Behavioural engagement (BE)		**3.74**	**.54**	$x^2(2, 48) = .324$, p = .851 ns
Task	**Reading**	3.81	.54	
	Reading Plus	3.73	.68	
	Writing	3.68	.41	
Emotional engagement (EE)		**3.60**	**.71**	$x^2(2, 48) = 2.47$, p = .29 ns
	Reading	3.86	.56	
	Reading Plus	3.56	.72	
	Writing	3.39	.79	
Cognitive engagement (CE)		**3.77**	**.72**	$x^2(2, 48) = .46$, p = .793 ns
	Reading	3.84	.68	
	Reading Plus	3.78	.74	
	Writing	3.69	.78	
Agentive engagement (AE)		**2.07**	**.79**	$x^2(2, 48) = 3.39$, p = .183 ns
	Reading	2.31	.71	
	Reading Plus	1.95	.95	
	Writing	1.95	.68	
Social engagement (SE)		**1.67**	**.65**	$x^2(2, 48) = 1.92$, p = .384 ns
	Reading	1.78	.70	
	Reading Plus	1.57	.80	
	Writing	1.63	.45	

Table 8

Five Dimensions of Participants' Task Engagement across Three Experimental Groups

	BE vs. EE	**BE vs. CE**	**BE vs. AE**	**BE vs. SE**	**CE vs. EE**	**AE vs. EE**	**EE vs. SE**	**AE vs. CE**	**CE vs. SE**
Reading (R)	z = -.44 p = .659 ns	z = .75 p = .452 ns	z = -3.5 **p < .001** *	z = -3.4 **p < .001** *	z = -.15 p = .877 ns	z = -3.47 **p < .001** *	z = -3.5 **p < .001** *	z = -3.4 **p < .001** *	z = -3.5 **p < .001** *
Reading Plus (RP)	z = -.91 p = .365 ns	z = -.08 p = .932 ns	z = -3.40 **p < .001** *	z = -3.41 **p < .001** *	z = -.55 p = .582 ns	z = -3.35 **p < .001** *	z = -3.42 **p < .001** *	z = -3.3 **p < .001** *	z = -3.4 **p < .001** *
Writing (W)	z = -1.43 p = .153 ns	z = -.19 p = .850 ns	z = -3.62 **p < .001** *	z = -3.63 **p < .001** *	z = -1.58 p = .115 ns	z = -3.52 **p < .001** *	z = -3.52 **p < .001** *	z = -3.64 **p < .001** *	z = -3.62 **p < .001** *

Note: results based on Wilcoxon Signed Ranks Test; * = statistically significant

RQ3: How do participants reflect on the experimental tasks and on the broader experience of participating in study?

Three thematic areas of interest were identified. The first of these related to interviewees' overall positive evaluations of participation. For all three interviewees, it was the first time they were participating in this type of research study and they were able to articulate clear rationales for taking part, ranging from curiosity "*to see what it is like because since we have to do our master's degree dissertation and conduct a study*" (Int 3), anticipated enjoyment "*I actually, I think it would be fun, just to be fun*" (Int 1) and more altruistic motives "*I just like being helpful. I mean, if somebody's doing a study and I can be helpful in any way, I'm, I'm not going to refuse, am I?*" (Int 3). Importantly, all three interviewees perceived having "*learned something*" (Int 3), reporting their perceptions of the vocabulary tests/study participation as educational. This is highly relevant both in terms of understanding the impact of testing effects and, critically, ethical uses of participants' class time for research purposes.

Interviewee 1, however, reasonably questioned the usefulness of target items; while with one item ('wrath') she consciously sought out the word's origins and definition, "*with others, like, I thought maybe I didn't need them to, to be learned for me, like I didn't learn them, or studied or be or, or was interested in them*". This raises an interesting methodological question about the extent to which participants' subjective appraisals of vocabulary items, for example driven by perceptions of usefulness, may affect the intensity of their efforts across different items. Or, as here, how this may lead to a consciously uneven allocation of attentional resources. Support for this can be found in Coxhead (2025), whose qualitative data similarly suggested eclectic reasoning from participants when describing their individual assessment of and engagement with researcher-chosen target vocabulary.

The second thematic area concerned participants' emotional experiences of participation in the study. Interviewee 2 described feelings of embarrassment after the immediate post-test, having "*just read the text and then I had already forgotten like one word or maybe two*", and the delayed post-test "*I was a little embarrassed that I still didn't get all of those right*". Interviewee 3, at the delayed post-test, experienced this more intensely: "*I felt shame, because I haven't like, I haven't learned them*". It may be that this emotion was so acutely experienced because of learners' advanced proficiency levels and their dominant *ideal L2 self/own* motivations. More explicit reassurance to participants, particularly at study end, feels warranted, especially where experimental tasks or low-frequency target vocabulary may challenge their positive self-images as 'good' language learners, with potential to affect immediate and ongoing self-perceptions.

A further thematic topic of interest, and of relevance particularly in relation to replication efforts and future ILH research, centred on the treatment and methods employed. Specifically, interviewees proposed small changes that may increase the perceived educational benefit of such research, without negatively impacting findings. This includes, for instance, sharing answers to the gap fill in the RP condition and VLT/other vocabulary test scores. Another important methodological point of note is Interviewee 1's reflection on the *dynamic* nature of her engagement throughout the task, a factor rarely considered or accounted for in ILH-related research designs:

> OK, so at first I wanted to, to really be engaged in this, into this text to fill in. But then I, I guess in the middle of fill in the gaps, I, I thought, oh so, so difficult. Like I, I, I do not understand most of the words and maybe I will do it like very, very fast, just put in and then say… But then I, I thought, no, I should, I should calm down and try to like, to catch the context of the of the given text or paragraph, and try to fill in those words.

The qualitative dataset lastly confirmed our key methodological decisions related to topic choice and difficulty level. Interviewees perceived the focus of the experimental text on fake news as "*a topic that's quite popular and I think important to me and a lot of other people*" (Int 2) and the difficulty level of the task to be pitched overall at an appropriate level for their proficiency; "*I probably didn't get everything right, but I understood it quite well. I'm not saying it was so easy, but it was like a good level for me.*" (Int 2).

5 Discussion

Replicating and extending Hulstijn and Laufer's (2001) study into the effects of task-induced learner involvement on incidental vocabulary learning, our project has examined this process across three specific task conditions: Reading (R), Reading Plus (RP), and Writing (W). As a conceptual replication, our design followed the original study as closely as possible, while simultaneously introducing several elements of extension: a meaning recognition test, motivation and task engagement questionnaires, and qualitative individual interviews. Our aims in doing so were to increase methodological rigour and gain new insights into the impact of task-induced involvement on vocabulary learning, while foregrounding also a focus on some learner-related factors such as motivation and task engagement.

RQ1 examined participants' learning of ten target items and indicated that R was the least effective task, with both RP and W outperforming it. Interestingly, results did not indicate statistical differences between RP and W, which differs from Hulstijn and Laufer (2001). In their study, significant differences between the respective task conditions did occur and were taken as confirmation of the ILH-based claim that tasks with a higher involvement load (e.g. W) led to better vocabulary retention than tasks with lower involvement (e.g. R). Our replication with EFL learners in the Polish context can corroborate this only partially. The lack of difference between the RP and W conditions may have been due to various reasons, with the outcomes of learning likely going beyond just task effects and involving a multiplicity of contributing factors. We argue that one such key factor 'missing' from much extant vocabulary research is the learner themselves, an argument indeed supported by the findings of this study.

Further, by including two measures of vocabulary learning rather than just one, findings demonstrated that RP and W outperformed R not only in terms of meaning recall but also meaning recognition. As predicted, with the recognition test being easier, learners' recognition scores were higher than their recall scores, which clearly shows that ILH-based predictions need to consider newly learned words not only in terms of raw numbers, but also aspects of knowledge tested, as well as how and when they are tested. A case in point from the current analysis was the interaction between task type and time of testing: this interaction was found to be significant only in the recall data, suggesting that recall and recognition mastery in specific task conditions might follow different paths. None of this nuance would have emerged had this replication relied on only one measure of vocabulary or participants being tested only once.

Such a complex picture with a multiplicity of factors involved, be it in relation to specific aspects of knowledge, test formats or time of testing, has not been uncommon in prior vocabulary studies. For instance, Yanagiswa and Webb's (2021) meta-analysis revealed that ILH results looked different depending on whether immediate or delayed measures were considered. Similarly, Liu and Reynolds' (2022) review also found that individual ILH studies relied on different measures of vocabulary, underlining the importance of methodological considerations in interpreting findings, particularly when results are compared across different tasks or contexts of learning, which replication research rightly encourages (McManus, 2024; Szudarski & Mikołajczak, 2023). By way of example, a recent study by Gao et al. (2024) examined the impact of task-induced involvement on L1-Chinese learners' incidental vocabulary and findings clearly pointed to testing and time-on-task effects. In controlled time-on-task conditions, the impact of task-induced involvement was evident on the delayed but not on the immediate post-test.

Thus, the implication of these findings is that future research incorporating the ILH, or using it as a starting point for explorations of other factors relevant to learning, must reflect the complex nature of this process and methodologically draw on such designs that accurately capture how IL affects degrees of learning, and also how IL task-related effects might interact with other relevant factors. While this may not be a novel insight, our replication has strengthened this point in relation to learner-related factors, underpinned in particular by the novel, if modest, qualitative dataset. Schmitt (in press) has recently also highlighted this issue more directly: positioning vocabulary development as being *complex, integrative,*

incremental and learned through a process. It is not controversial to argue that the discussions of operationalisation of these aspects have not yet fully reached maturity in existing ILH-inspired scholarship.

In our study, RQ2 foregrounded the importance of this aspect by exploring participants' *task engagement*. To answer this research question, a self-report measure was included and invited participants to share their experiences of engagement while completing experimental tasks. Participants rated their engagement similarly across all task conditions, with behavioural, emotional and cognitive engagement reported as significantly greater than their agentic and social engagement. Considering the tasks themselves and the conditions under which they were completed, the lower levels of agentic and social engagement should not be a surprise. It is therefore perhaps more interesting to note that, although involvement loads differed across each task, participants reported equal levels of behavioural, emotional and cognitive engagement. One explanation for this may relate to the overall high language proficiency and motivation reported by participants. The de facto 'test conditions' created by data collection in this type of study may explain the higher and comparable levels of behavioural and cognitive engagement across tasks. On paper, the Writing task may have offered some additional scope for personal expression, potentially thus inspiring greater emotional engagement. However, participants appeared not to have perceived this or to have felt able to engage with the task in this way, perceiving it instead to be equally 'fun', interesting and enjoyable as the Reading and Reading Plus tasks.

In light of these findings, it is also instructive to consider the ways in which involvement load and task engagement differ. The ILH considers three factors (Search, Evaluation and Need), operationalised collectively as varying involvement loads that represent processing depth of *target vocabulary* in a task. Adopting a broader lens, task engagement involves all "deliberate actions on the part of learners in the service of completing a pedagogic task" (Hiver & Wu, 2023, p. 75). Engagement, by its very nature, is an embodied construct, that is, "it is 'done' by a learner/agent" (Hiver & Wu, 2023, p. 81), it is thus not possible to separate the learner from the learning. Thinking ahead, we ponder over the potential for study designs and methodological approaches that might bridge both perspectives. Such an approach would broaden empirical focus on task-related features (exemplified in extant ILH research) to one, or a constellation of, measure(s) that may come closer to representing a fuller picture of learning as experienced by the learner.

In pursing RQ3, a modest complementary dataset of qualitative interview data was collected, bringing in the "perceptions of people who are at the heart of this research" (Coxhead, 2025: 128). Findings from this dataset have fully justified this extension, providing further grist underpinning the role of qualitative approaches in vocabulary research as a critical yet under-utilised methodological avenue. Analysis of participants' perceptions pointed to important methodological considerations for future work. Having highlighted some of these already, in what follows we wish to spotlight one aspect in particular: the role of *emotions*.

Research into the role of emotions in language learning has blossomed into a thriving field of study (Resnik et al., 2025). Yet there has, however, remained a relative 'blind spot' concerning the interface between emotions and vocabulary development. Throughout interviews, participants foregrounded both positive emotions experienced, for example interest and curiosity, alongside negative emotions, notably shame and embarrassment. Considering the highly complex relationship between emotions and cognitive processing, a more central positioning of emotions in the role of vocabulary learning feels warranted. An unintended side-effect of the enduring popularity of the task-focused ILH, and the dominant focus on its cognitive elements as determining incidental learning, may have relegated into the shadows the concurrent need for attention to learners' emotions and motivation as also playing a determining role.

A second methodological implication from this dataset concerns the *dynamic* nature of participants' engagement with tasks. Clear in their reactions to perceived task difficulties and their feelings of needing to "*calm down*" and rethink, participants' engagement – and by extension we hypothesise, their depth of involvement with target items – fluctuated throughout the task. Another element of inherent dynamicity

evident in the qualitative dataset was participants' perceptions of the 'usefulness' of researcher-identified target items and their described impacts on behaviour. The importance of *choice* underpins many key conceptualisations of motivation, including in the context of vocabulary learning: Wang et al. (2015) found that compared to researcher-chosen target vocabulary, participants who identified personalised target words exhibited greater task engagement and motivation (for similar discussion, see also Coxhead, 2025).

Crucially in terms of methodology, we were able to reach these conclusions because of the replication design employed, as well as owing to the more central positioning of the learner. As the results demonstrate, with direct comparisons with Hulstijn & Laufer (2001) still possible, findings have also thrown new light on the mechanics of incidental learning in specific task conditions, adding a learner-related perspective and revealing the dynamic nature of their engagement with the experimental text, tasks and vocabulary tests. We see this as an important methodological contribution to ILH studies and wider vocabulary research.

Continuing this trajectory and thinking forward, we argue that greater interdisciplinary collaboration may be needed. Certainly, the challenge of conducting interdisciplinary research is not straightforward. However, with increasing bridges across disciplinary and methodological traditions and a growing focus on the interrelationship between learner and task in SLA (e.g. exemplified in *TESOL Quarterly*'s 2025 Special Issue, see Lambert & Aubrey, 2025) and vocabulary research specifically (e.g. the 2026 JESLA Special Issue on interdisciplinary dialogue between vocabulary studies and language learning psychology, see Szudarski & Muir, forthcoming), we are optimistic and encourage growing uptake of this challenge.

6 Limitations and Future Research

This study is not devoid of limitations. First, the sample size accrued was small, with participants recruited from one specific context. As such, the findings from this replication are preliminary. Future research might consider multisite replications that include multiple samples across contexts and consequently allow more generalizations (see Peters et al. (2023) for an example).

Second, following Hulstijn and Laufer (2001), our study treated all ten target items as equivalent to each other, with the analysis based on average scores. However, a closer look at the target vocabulary suggests that some items may have been more difficult to learn than others (e.g. multi-word unit 'not one whit' vs. individual word 'inflammatory'). Future research should make a distinction between the two and explore whether the ILH assumptions hold for both the learning of words and phrases. Relatedly, qualitative findings also indicated the need to consult participants' perceptions of items (e.g. of relative difficulty or usefulness) as affecting, for example, their conscious attentional allocations.

Third, learners' behavioural, cognitive and emotional engagement has been demonstrated to exert significant, positive effects on vocabulary development, with specific dimensions of engagement having different influences and exhibiting interdependent and fluctuating relationships (Zhang et al., 2023). Idiodynamic methods, an inherently mixed methods approach integrating quantitative and qualitative data, may offer particular methodological promise, with its ability to capture the *dynamic* nature of learners' engagement and the interrelated effects of discrete aspects of engagement (e.g. cognitive and emotional dimensions; Aubrey, 2022). Future research might also move beyond self-reports of learners' engagement. Indeed, observable measures of engagement, such as number of words produced, have been found to better predict learners' vocabulary development than self-report measures (Hiver & Dao, 2025).

Fourth, while this study has sought to underline the importance of learner-related factors in ILH research and expand focus by including measures of learners' motivation and task engagement, our design has not allowed empirical interrogation of revised proposals such as the ILH plus (Yanagisawa & Webb, 2022) or New TFA (Muir & Szudarski, 2025). In Nguyen's (2025) recent study comparing the ILH, TFA

and ILH Plus, the ILH Plus was found to better predict learners' vocabulary gains relative to task-induced involvement, with time-on-task also explaining vocabulary uptake and retention. Thus, it is clear that future research can involve comparisons based on these conceptualisations and alongside different factors. Equally, we propose it might be that making such comparisons will require different operationalisations of involvement load separately for studies foregrounding theoretical and pedagogical aims.

Finally, while we continue to underline the significance of the qualitative contributions to the study and the methodological and other implications this has suggested, we recognise the limited number and voluntary nature of our interviews. They similarly do not include participants from all experimental conditions, meaning the findings may present only a partial picture.

7 Conclusion

As a replication and extension of Hulstijn and Laufer (2001), this study has investigated English learners' incidental learning of vocabulary as dependent on task-induced involvement load across three task conditions: Reading (R), Reading Plus (RP), and Writing (W). While R was found to be less effective than both RP and W, the latter two resulted in participants' similar vocabulary scores, which does not fully replicate the findings of Laufer and Hulstijn (2001). Analyses have also shown that learners' task-related vocabulary performance was mediated by many factors, including aspects of lexical mastery measured, test formats used, and testing time. Finally, results have also pointed towards the dynamic nature of learners' emotions and engagement with these tasks, underlining the methodological need for a more established presence of learner-related variables within vocabulary learning research.

Collectively, this replication study deepens our understanding of the ILH in relation to the effectiveness of tasks and their impact on vocabulary gains. It further provides a range of theoretical and methodological insights regarding the construct of involvement load and how it is used to examine task-related vocabulary learning. We believe our findings demonstrate the value of replication research, while at the same time pointing to the benefits of interdisciplinarity and cross-pollination across different research traditions, in this case vocabulary studies and language learning psychology. There are some encouraging signs, e.g. Nation's (2024) addition of motivation as a key principle for delivering effective vocabulary learning and the 2026 JESLA Special Issue on interdisciplinarity across vocabulary studies and language learning psychology (Szudarski & Muir, forthcoming; see also Lambert & Aubrey, 2025 SI exploring the role of the learner in task performance). However, much more remains to be done in terms of facilitating robust interdisciplinary approaches to vocabulary development, and indeed other aspects of language learning. We hope this replication is a useful illustration of how such efforts can offer a useful stepping stone and contribute richer descriptions of the studied SLA phenomena.

Acknowledgements

We would like to thank our participants. We thank also Dr Maciej Laskowski and particularly Dr Emilia Wąsikiewicz-Firlej, without whose assistance this research would not have been possible. We share our grateful thanks also with Profs Jan Hulstijn and Batia Laufer for their kind correspondence, engagement with our questions, and useful feedback on earlier drafts of the manuscript. We also thank Prof Norbert Schmitt for acting as a generous sounding board and discussant. We lastly thank Dr Beatriz Gonzalez-Fernandez for acting as a critical friend and for her productive feedback on a final draft of this article.

Supplementary Materials

Check the Supplementary Materials on the journal website for all appendices.

References

Aubrey, S. (2022). The relationship between anxiety, enjoyment, and breakdown fluency during second language speaking tasks: An idiodynamic investigation. *Frontiers in Psychology*, *30*. https://doi.org/10.3389/fpsyg.2022.968946

Braun, V., Clarke, V., Hayfield, N., Davey, L., Jenkinson, E. (2022). Doing Reflexive Thematic Analysis. In S. Bager-Charleson, A. McBeath (Eds.) *Supporting Research in Counselling and Psychotherapy* (pp. 19–38). Springer.

Coxhead, A. (2025). Understanding teachers' and learners' perceptions of incidental vocabulary learning: another piece of the puzzle. In M. F. Teng & B. L. Reynolds (Eds.), *Researching incidental vocabulary learning in a second language* (pp. 119-130). Routledge.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* Mahwah, NJ: Lawrence Erlbaum.

Folse, K. (2006). The effect of type of written exercises on L2 vocabulary retention. *TESOL Quarterly, 40*(2), 273-293. https://doi.org/10.2307/40264523

Fredericks, J. A., Blumenfield, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109. https://doi.org/10.3102/00346543074001059

Gao, M., Qian, J., & Rasool, U. (2024). Effects of task-induced involvement and time on task on incidental L2 vocabulary acquisition. *SAGE Open, 14*(2), 1-15. https://doi.org/10.1177/21582440241249340

Hazrat, M. (2020). The involvement load hypothesis and its impact on vocabulary learning. PhD thesis. University of Auckland.

Hiver, P. & Dao, P. (2025). From task motivation to L2 learning: Understanding links through learners' task engagement. *TESOL Quarterly*, *59*(2), 24–59. https://doi.org/10.1002/tesq.3410

Hiver, P. & Wu, J. (2023). Engagement in TBLT. In C. Lambert, S. Aubrey & G. Bui (Eds.) *The Role of the Learner in Task-Based Language Teaching* (pp. 74–90). Taylor & Francis.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *Modern Language Journal*, *96*(4), 544–557. http://dx.doi.org/10.1111/j.1540-4781.2012.01394.x

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539–558. http://dx.doi.org/10.1111/0023-8333.00164

Keating, G. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365–386. https://doi.org/10.1177/1362168808089922

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285-325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

Lambert, C. & Aubrey, S. (2025). The role of the learner in task performance and acquisition: Evidence from new and emerging perspectives. *TESOL Quarterly, 59*(2), 5–23. https://doi.org/10.1002/tesq.70047

Laufer, B. & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Laufer, B. (2020). Evaluating exercises for learning vocabulary. In S. Webb (ed.) *Routledge handbook of vocabulary studies* (pp. 351–368). London: Routledge.

Liu, S., & Reynolds, B.L. (2022). Empirical support for the Involvement Load Hypothesis (ILH): A systematic review. *Behavioral Sciences, 12(10),* 1-23. https://doi.org/10.3390/bs12100354

McManus, K. (2024). Replication studies in second language acquisition research: Definitions, issues, resources, and future direction. Introduction to the special issue. *Studies in Second Language Acquisition, 46(5)*, 1299-1319. https://doi.org/10.1017/S0272263124000652

McManus, K. (2026). Replications studies in TESOL. *TESOL Quarterly*, Early View, 1-18. https://doi.org/10.1002/tesq.70124

Muir, C., & Szudarski, P. (2025). Motivation, task-induced involvement load and incidental vocabulary learning. In M. F. Teng & B. L. Reynolds (Eds.), *Researching incidental vocabulary learning in a second language* (pp. 149-164). Routledge.

Nation, P. (2024). Re-thinking the principles of (vocabulary) learning and their applications. *Languages, 9*(160), 1-14. https://doi.org/10.3390/languages9050160

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston: Heinle, Cengage Learning.

Nguyen, C.-D. (2025). Best Model to Predict Vocabulary Learning Gains: ILH, ILH Plus, or TFA? *TESOL Journal, 16*(4), 1-14. https://doi.org/10.1002/tesj.70081

Papi, M., Bondarenko, A.V., Mansouri, S., Feng, L., & Jiang, C. (2019). Rethinking L2 motivation research. The 2x2 model of self-guides. *Studies in Second Language Acquisition*, *41,* 337–361. https://doi.org/10.1017/S0272263118000153

Papi, M., & Khajavy, G. H. (2021). Motivational Mechanisms Underlying Second Language Achievement: A Regulatory Focus Perspective. *Language Learning, 71*(2), 537-572. https://doi.org/10.1111/lang.12443

Paribakht, S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 174–200). Cambridge University Press.

Peters, E., Puimège, E., & Szudarski, P. (2023). Repetition and Incidental Learning of Multiword Units: A Conceptual Multisite Replication Study of Webb, Newton, and Chang (2013). *Language Learning, 73*(4), 1211-1251. https://doi.org/10.1111/lang.12621

Philp, J. & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics*, *36*, 50–72. https://doi.org/10.1017/S0267190515000094

Resnik, P., Dewaele, J-M., Li, C. & Botes, E. (2025). The role of positive and negative emotions in foreign language learning: A research agenda. *Language Teaching*, 1–26. https://doi.org/10.1017/S0261444825101031

Rott, S. (2012). The effect of task-induced involvement on L2 vocabulary acquisition: An approximate replication of Hulstijn and Laufer (2001). In G. Porte (ed.) *Replication research in applied linguistics* (pp. 228–267). Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. https://doi.org/10.1177/026553220101800010

Schmitt, N. (2010). *Researching vocabulary. A vocabulary research manual.* Palgrave Macmillan.

Schmitt, N. (in press). *A CIPP perspective of vocabulary knowledge and acquisition.* Multilingual Matters.

Sudsa-ard, S. (2023). The effects of technique feature analysis of retention of form recall in written production. PhD thesis, University of Leeds.

Szudarski, P. & Mikołajczak, S. (2023). Proficiency, language of assessment, and attention to meaning and form during L2 comprehension: Methodological considerations in L2 replication research. *Studies in Second Language Acquisition, 45*(1), 276-288. https://doi.org/10.1017/S0272263122000171

Szudarski, P. & Muir, C. (forthcoming). Vocabulary studies and the psychology of language learning: towards greater interdisciplinary collaboration. *Journal of the European Second Language Association*.

Teng, M. F. & Zhang, D. (2024). Task-induced involvement load, vocabulary learning in a foreign language, and their association with metacognition. *Language Teaching Research, 28*(2), 531-555. https://doi.org/10.1177/13621688211008798

Ushioda, E. (2009). A person-in-context relational view of emergent motivation, self and identity. In Z. Dörnyei & E. Ushioda (eds.) *Motivation, language identity and the L2 self* (pp. 215-228). Multilingual Matters.

Wang, H.-Ch., Huang, H.-T., & Hsu, Ch.-Ch. (2015). The impact of choice on EFL students' motivation and engagement with L2 vocabulary learning. *Taiwan Journal of TESOL, 12*(2), 1-40.

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning*, *71*(2), 487–536. https://doi.org/10.1111/lang.12444

Yanagisawa, A., & Webb, S. (2022). Involvement load hypothesis plus. Creating an improved predictive model of incidental vocabulary learning. *Studies in Second Language Acquisition*, *44*(5), 1279–1308. https://doi.org/10.1017/S0272263121000577

Zare, J., & Derakhshan, A. (2024). Task engagement in second language acquisition: a questionnaire development and validation study. *Journal of Multilingual and Multicultural Developmen*t, 1-12. https://doi.org/10.1080/01434632.2024.2306166

Zhang, R., Zou, D., & Cheng, G. (2023). Learner engagement in digital game-based vocabulary learning and its effects on EFL vocabulary development. *System, 119,* 1-19. https://doi.org/10.1016/j.system.2023.103173

Paweł Szudarski is Associate Professor of Applied Linguistics at the University of Nottingham, UK. He works in the areas of second language acquisition, corpus linguistics and TESOL, focusing in particular on the acquisition of vocabulary and phraseology by second/foreign language learners. His other interests include corpus-based analysis, replication research, online pedagogies, and teaching and learning of international students. ORCiD: https://orcid.org/0000-0002-8270-2932

Christine Muir is Assistant Professor in Second Language Acquisition at the University of Nottingham, UK. Working broadly in the area of the psychology of language learning and teaching, her research focuses in particular around student (long-term) language learning motivation and language teacher wellbeing. Her other interests include teaching English for Academic Purposes (EAP) in UK Higher Education and qualitative research methodologies. ORCiD: https://orcid.org/0000-0001-8491-7593