Predicting Personalized Vocabulary Learning Burden from Intralexical and Interlexical Factors

Mykhailo Diachenko
University of Galway
Galway, Ireland
hi@nekear.me

John P. McCrae
Institute for Data and AI Research (IDAIR)
University of Galway, Galway, Ireland
john.mccrae@universityofgalway.ie

Abstract

Existing measures of word difficulty rarely capture how much effort a particular learner needs to learn a particular word. We introduce personalized learning burden (PLB) — the number of spaced, non-cramming reviews a learner needs to bring a word to 14-day memory stability, extracted from existing review logs without human annotation. In Anki notes and review logs from four multilingual learners of Japanese, non-personalized features explain little of PLB: the lexical item accounts for at most 14.84% of its variance. We therefore ask whether PLB on known words transfers to unseen words through semantic, morphological, orthographic, or phonological similarity. Neighbor-informed models improve over a within-learner median for two of the three learners we test, but the gain does not come from similarity: random words the learner completed within one month of each neighbor reproduce it, and the target's position on the learner's timeline alone predicts best (up to Spearman's $\rho = .557$ and a 16.2% error reduction). Because neighborhoods drawn from a learner's history encode timeline position by construction, studies of interlexical effects need temporally matched controls. We release a dataset of about 66,000 Anki notes and over 670,000 review logs from eight learners with language profiles.

1 Introduction

Reading with adequate text coverage in a second language (L2) requires knowing around 8,000 word families, which comprise about 34,660 word forms (Nation, 2006, as cited in Schmitt, 2008). Vocabulary on this scale cannot be acquired from exposure alone and requires deliberate, sustained study (Schmitt, 2008; Nation, 2022). Flashcards are a well-established tool for such study. Learners who used them regularly gained over 1,000 words in eight months, compared with fewer than 100 for a control group (McLean et al., 2013, as cited in Nation, 2022). Flashcards can be used for receptive (passive) recall practice—retrieving a word's meaning from its form—which was the test type that best predicted course grades in Laufer and Goldstein (2004). Moreover, spacing reviews over time yields medium-to-large retention gains over massed practice (Kim and Webb, 2022). Applications such as Anki (Ankitects Pty Ltd, 2026) (Section 3.1), whose benefits are supported by empirical evidence (Hanson and Brown, 2020), combine these elements with schedulers that adapt review intervals to each learner's performance.

Words, however, differ widely in how much study they require. The effort needed to master a word — its *learning burden* — is highly idiosyncratic: the more a word aligns with patterns a learner already knows, the lighter its burden (Nation, 2022, p. 49). Prior exposure is a major determinant of how difficult a given word is for a given learner (Hashimoto and Egbert, 2019), and relationships with known words act in both directions. Shared structure and cognate overlap make acquisition easier, whereas confusable forms and near-synonyms cause interference that learners may not even notice (Laufer, 1988, 1990, 2025). A word's learning burden therefore depends not only on the word itself, but also on the vocabulary a particular learner has already built.

Existing computational approaches capture complementary parts of this problem: lexical complexity prediction models word difficulty, knowledge tracing models learner histories, and spaced-repetition schedulers model item-level memory (Section 2). However, to our knowledge none jointly predicts the effort a particular learner will need to learn an unseen word from both the word's properties and its relationships to the learner's

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 140–159, November 19–20, 2026, University of Tartu Library

known vocabulary, nor validates such predictions on real multilingual review data.

We introduce *personalized learning burden* (PLB) — the number of spaced, non-cramming reviews a specific learner needs to bring a lexical item to a target memory stability. Stability is the number of days after which the item's recall probability would fall to 90%, and we set the target to 14 days. PLB thus measures memorization effort from review logs that learners already generate, without human annotation. We predict PLB from progressively richer information: the learner's overall and recent performance, features of the target word, and, in neighbour-informed models, the observed PLB of known words that are semantically, morphologically, orthographically, or phonologically similar to the target. Because our difficulty feature must score arbitrary flashcard content for learners with different first languages (L1s) and no receptive-difficulty dataset spans multiple languages, we fine-tune a difficulty prior on productive difficulty scores (British Council, 2026). Productive knowledge is harder to acquire than receptive knowledge (Nation, 2022; Laufer and Goldstein, 2004), so we first test whether the prior transfers to receptive difficulty on MultiLS-Japanese, a dataset of Japanese lexical complexity ratings by non-native readers (Nohejl et al., 2024). We evaluate all models on Anki review logs from multilingual learners of Japanese with diverse L1s.

We address the following research questions (RQs), answering those on PLB for a cohort of four learners of Japanese, three of whom have enough data to evaluate RQ2:

- RQ1: How much of PLB can population-level features explain?
  - RQ1a: Does the productive difficulty prior transfer to receptive population-level difficulty?
  - RQ1b: What share of PLB variance is attributable to the lexical item at all?
  - RQ1c: Do the prior and word frequency from TUBELEX, a corpus of YouTube subtitles (Nohejl et al., 2025), predict PLB?
- RQ2 (main): Does a learner's observed PLB on known words transfer to unseen words through interlexical similarity?

Our contributions are:

1. We introduce and operationalize personalized learning burden (PLB) directly from existing review logs, without additional human annotation.
2. We release a dataset of about 66,000 Anki notes and over 670,000 FSRS (Free Spaced Repetition Scheduler; Ye, 2026) review logs from eight multilingual Anki users with language profiles, pairing receptive review histories with cross-linguistic lexical content.[1]
3. (RQ1) We show that the productive difficulty prior transfers only weakly to receptive difficulty on MultiLS-Japanese ($R^2 = .112$, against .41 for TUBELEX log frequency; Nohejl et al., 2025). On PLB in our four-learner cohort, the lexical item accounts for at most 14.84% of the variance, a ceiling on intralexical features that TUBELEX frequency comes nowhere near ($\approx 1\%$ of total variance). Conditioning on the learner's L1, as the prior does, does not help either (Spearman's $\rho = -.027$).
4. (RQ2) In our cohort, we find no evidence of similarity-specific transfer: neighbour-informed models improve prediction for two of three learners, but temporally matched random neighbours reproduce the gain, and the target's position on the learner's timeline alone predicts best (ρ up to .557, error reduction up to 16.2%).
5. We identify timeline position as a confound for neighbour-based models, since neighbourhoods drawn from a learner's history encode it by construction, and propose a temporally matched random-neighbour control and a days-elapsed baseline for future studies of interlexical effects.

2 Background and Related Work

A word's difficulty may stem from *intralexical* factors — properties of the word itself — or from *interlexical* factors — its relationships to other words the learner knows (Laufer, 1990; Schmitt, 2008; Laufer, 2025). Because the literature draws this boundary inconsistently, we apply a single criterion: a feature is *intralexical* if it can be computed from the L2 word alone, and *interlexical*

[1] Dataset (CC BY 4.0): https://huggingface.co/datasets/nekear/anki-language-learning; code (MIT): https://github.com/nekear/plb.

if its computation requires reference to another word, whether an L2 word the learner knows or a word in their L1.

2.1 Intralexical Difficulty

Frequency is commonly assumed to be the dominant intralexical predictor of difficulty, but the evidence is mixed. In a study of 403 adult L2-English learners and 33 intralexical variables, Hashimoto and Egbert (2019) found frequency to be the strongest single correlate of item difficulty ($r = .499$), yet it explained only about 25% of the variance, and its effect depended on the subcorpus used. Range and dispersion, by contrast, predicted difficulty consistently. The authors interpret all three as facets of one construct: the likelihood of a learner's repeated exposure to a word. Laufer (2025) excludes frequency from intralexical factors altogether, since it reflects the learning context rather than the word itself, and instead lists lack of phonotactic regularity, irregular sound-to-spelling correspondence, deceptive morphological transparency, polysemy, and figurativeness. Because words usually combine several of these characteristics, she argues, difficulty is hard to isolate and determine from intralexical properties alone.

As reviewed by North et al. (2023), computational work builds on such properties, using them as features in models that map lexical items to difficulty scores, a task known as lexical complexity prediction (LCP). Because generic models missed how individual learners perceive complexity, later work personalized them, conditioning on learner attributes such as L1 and proficiency or on judgments elicited from the learner through active learning (Gooding and Tragut, 2022). Today, state-of-the-art predictors are transformer-based, replacing hand-crafted features with learned representations (North et al., 2023). Skidmore et al. (2025) fine-tuned XLM-RoBERTa to predict productive difficulty on an extended version of the Knowledge-based Vocabulary Lists (KVL) dataset (Spearman $\rho = .78$–$.79$), and their model became the baseline of the shared task at the 2026 Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026 Shared Task; British Council, 2026). The winning system of its open track (Nohejl et al., 2026) introduced a soft-target cross-entropy loss for continuous prediction (Pearson $r > .91$) and showed that the same loss makes masked language models competitive with decoder-based large language models (LLMs) of comparable size. All of these approaches, however, estimate the probability that a learner already *knows* a word rather than the *effort required to learn it*, and therefore do not model learning burden.

2.2 Interlexical Effects

Drawing on Meara (1980) and Hatch (1983), Laufer (1990) describes a single mental lexicon shared by all of a multilingual's languages, in which words are grouped first into semantic categories, then into phonological networks, where a word's nearest neighbour is the one it most resembles in sound, and only then into L1 and L2 sets. Within this lexicon, words interact on the basis of semantic, phonological, and morphological similarity, across languages as well as within them. A new word must therefore be integrated into this network, and how easily it is learned depends on its connections to what is already known. Material with no association to previously learned material is difficult to learn (Higa, 1972, as cited in Laufer, 1990). Conversely, *meaningfulness* — the mean frequency of continued associations in 60 seconds (Noble, 1952, as cited in Laufer, 1990) — speeds learning: in studies with nonsense syllables, more meaningful items were learned faster (Noble and McNealy, 1957, as cited in Laufer, 1990).

Association with known words, however, does not always help: similarity can hinder learning as well as facilitate it. Between L1 and L2, *cognates* and *loanwords* (words that usually share both spelling and meaning across the two languages) are the easiest to learn, followed by *one-to-many correspondences* (where one word in one language corresponds to several in the other), since only the form difference must be learned. Difficulty increases with partial overlap and metaphorical extension, and is greatest for *lexical voids* (L2 words whose concept does not exist in the L1), since the learner must form a new concept, and for *false friends* (words similar in form but different in meaning), since a familiar form must be dissociated from a familiar meaning (Laufer, 1990, 2025). Within a language, semantic relations to known words matter as well: near-synonyms add a new meaning link that interferes with existing knowledge, whereas hyponymy, antonymy, and converseness facilitate learning (Laufer, 1990).

Form alone can also interfere: Laufer (1988) identifies ten categories of *synforms* (words similar in form) spanning morphology, phonology, and orthography, among which pairs differing only in suffix (*comprehensive*/*comprehensible*) or vowel (*adopt*/*adapt*) are particularly confusing. Laufer (2025) classifies synformy as intralexical, but because it arises only relative to other known forms, it is interlexical under our criterion.

Taken together, these findings suggest that difficulty arises when the degree of *meaning* similarity (semantic similarity) diverges from the degree of *form* similarity (morphological, orthographic, and phonological).

2.3 Learner and Memory Models

Knowledge tracing (KT) estimates a learner's evolving mastery of skills from their interaction history, using Bayesian or factor-analysis models or deep sequence models extended with memory, attention, graph structure, question text, or forgetting (Abdelrahman et al., 2023). Applied to vocabulary, this framework has three limitations. First, items are linked to skills through a Q-matrix (a binary item-by-skill matrix specifying which skills each item requires), but vocabulary has no natural item–skill mapping, and models that instead learn relations between items derive them from interaction co-occurrence or generic question text rather than linguistic analysis. These relations are also of a single, positive type (similarity or prerequisite), so they cannot express interference such as that between synforms or false friends (§2.2). Second, KT predicts the correctness of the next response rather than the effort required to master an item, which could only be approximated by rolling a forgetting-aware model forward. Third, prediction for items absent from a learner's history remains unaddressed.

Whereas KT tracks mastery of skills, spaced-repetition schedulers model the memory of individual items and use it to decide when each should be reviewed. Following the forgetting curve (Ebbinghaus, 1913), they time each review shortly before an item would be forgotten, reaching a target recall probability with as few reviews as possible. Half-life regression (HLR; Settles and Meeder, 2016) estimates an item's memory half-life from interaction features. It originally also used lexeme-tag features, but these were eventually excluded because they added little predictive value, caused overfitting, and prevented use without a predefined vocabulary. SSP-MMC (Stochastic-Shortest-Path-Minimize-Memorization-Cost; Ye et al., 2022) combines a half-life memory model with a planner that minimizes the expected review cost of bringing an item to a target half-life, but it supports only binary outcomes, lets item difficulty rise after failures without ever recovering, and is fitted at the population level, so it cannot adapt to an individual learner. Its successor FSRS, the scheduler currently recommended for Anki (Ye, 2026), addresses all three limitations: it supports Anki's four-button grading (Section 3.1), updates difficulty in both directions, and can be refitted to an individual learner's history. FSRS represents an item's memory state through *difficulty* and *stability* (the interval after which recall probability falls to 90%), and predicts *retrievability* (the current probability of recall) from stability and the time elapsed since the last review. Review intervals are solved analytically for a configurable desired retention (90% by default), and the changes to difficulty and stability caused by cramming are limited. FSRS, however, remains content-agnostic. It schedules each item from its review history alone, so neither the item's intralexical properties nor its interlexical relations play any role.

3 Approach and Findings

We first introduce the Anki terminology used throughout the paper (Section 3.1) and then describe our dataset and how PLB is derived from it (Section 3.2). We then address RQ1 to establish whether individual personalization is needed at all (Section 3.3). We validate the productive difficulty prior on receptive difficulty, since our models use it as a target-word feature. We also estimate the ceiling on how much of PLB any intralexical feature can explain, test how close TUBELEX log frequency comes to that ceiling, and test whether conditioning on L1 through the prior helps. Finally, we address RQ2 by testing whether PLB on known words transfers to unseen words through interlexical similarity (Section 3.4).

3.1 Anki Terminology

Anki (Ankitects Pty Ltd, 2026) is a flashcard application in which learners create or import their own study material, organized into *decks* of *notes*. Each note stores one lexical item in named *fields*

(e.g., the word, its reading, and its meaning) and generates one or more *cards*, each testing the note in one way: a receptive card shows the L2 word and asks for its meaning, a productive card asks the reverse, and a cloze card asks the learner to fill a gap in a sentence. Cards are scheduled independently. Each time a card is shown counts as a *review*: the learner tries to recall the answer, reveals it, and rates their recall as *Again* (forgotten), *Hard*, *Good*, or *Easy* (ratings 1–4). The scheduler then sets the card's next *interval*, the number of days until it is due again. Forgetting a card that had already been learned is a *lapse*, and the card returns to short *relearning* steps. Anki logs every review with its timestamp, rating, interval, and *type*, such as *learning* (a new card's first steps), *review* (a scheduled review of a learned card), or *relearning* (after a lapse).

3.2 Dataset and Prediction Target

Predicting PLB requires data that combine the lexical content being studied, each learner's ordered and graded review history, and information about the learner's languages. To our knowledge, no public dataset offers all three: LCP datasets lack learning histories (North et al., 2023), knowledge-tracing datasets lack lexical content (Abdelrahman et al., 2023), Duolingo's HLR data (Settles and Meeder, 2016) and the logs from the MaiMemo vocabulary app behind SSP-MMC (Ye et al., 2022) record only binary outcomes, and FSRS-Anki-20k (Open Spaced Repetition, 2024) has ordered, graded Anki histories but neither lexical content nor learner attributes.

We therefore collected a dataset of approximately 66,000 Anki notes (96,000 cards) and over 670,000 associated review logs from eight learners: six studying Japanese, whom we refer to as L01–L06 (L01 also studies Dutch, and L04 English), one studying Italian, and one studying Spanish and Portuguese. Participants uploaded their decks through a web-based survey, in which they also reported their L1s, additional-language proficiencies, and areas of interest, and consented to the public release of their notes, cards, and review logs (Appendix F). Table 6 lists each learner's L1s, self-rated target-language proficiency, interests, and data size, and Table 7 shows an example record from each file of the released dataset.

For our experiments, we retained notes that (1) represent a single-word Japanese lexeme, (2) have a clear form–meaning mapping, so that the receptive practice type can be inferred, and (3) have at most two cards. When two cards are present, we treat them as receptive and productive variants and, because receptive knowledge is easier to acquire (Nation, 2022), keep as the receptive card the one with the easier review history (where *easier* means a lower ratio of review count to achieved FSRS stability). This yielded 24,433 unique lexeme forms (9,478 shared by two or more learners), 42,369 unique form–meaning pairs, and 503,826 review logs.

Prediction Target

For each learner–item pair, we predict PLB — the number of clean reviews the learner needed to bring the item to an FSRS stability of 14 days. Because cramming inflates review counts without producing durable learning, a review is considered clean only if it is a scheduled review of a graduated card (`type = review`; Section 3.1), which excludes learning and relearning steps, and comes at least 24 hours after the card's previous review, which excludes repeated study within a day. PLB is thus a time-to-event quantity measured on a review-count axis rather than a calendar axis: it counts reviews, not days. It still depends on the intervals, however, because FSRS credits a successful review with more stability the longer it has been since the previous one, and therefore on the scheduler settings that chose them (Sections 3.2 and 5). Algorithm 1 in Appendix A gives the extraction procedure and its right-censoring rule.

Figure 1 shows the review timelines of five lexemes shared by all six Japanese learners. Lapses (relearning events) cause FSRS to raise difficulty and lower stability, prompting a much earlier next review than a correct answer would.

Censoring

Not every item reaches the target stability by the end of the collection window. For these items, PLB is undefined, but the number of clean reviews already performed is known, so the observation is not missing - it is *right-censored*, and the true PLB is known to exceed the number of reviews observed so far. Censored items are usually the slowest to master, and censoring rates vary sharply across learners, from 4.6% to 44.5% (Table 4 in Appendix C), so removing them would affect learners unevenly. We separately exclude

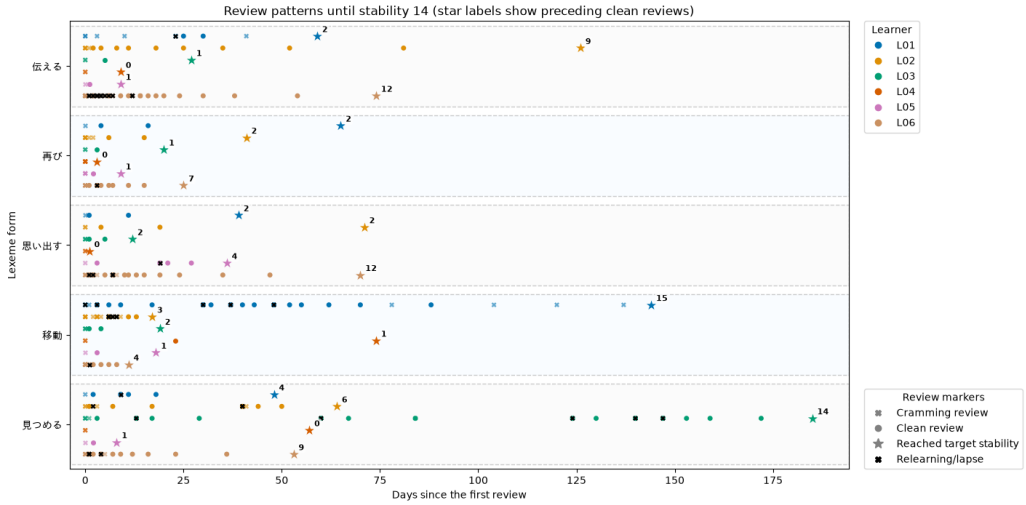


Figure 1: Per-lexeme review timelines for five shared lexemes across six learners. From top to bottom, the lexemes are *tsutaeru* 'to convey', *futatabi* 'again', *omoidasu* 'to recall', *idō* 'movement', and *mitsumeru* 'to gaze at'; within each lexeme, rows are ordered L01–L06. Stars mark PLB, semi-transparent crosses mark clean reviews, and black crosses mark lapses.

cards with zero reviews, since these represent study backlog rather than evidence of difficulty.

Review patterns are highly idiosyncratic (Figure 1): the number of clean reviews needed to reach the target FSRS stability (PLB) varies widely across learners, who require very different review counts to bring the same words to the same stability threshold. Kaplan–Meier burden distributions (Figure 3 in Appendix H) confirm this dispersion. Two learners, L04 and L05, reach the target in one or two reviews for most of their words, reflected in an interquartile range (IQR) of 0 (Table 4 in Appendix C). Based on their FSRS configurations, we attribute this to their FSRS weights and desired-retention settings rather than to genuine learning speed. Because an IQR of 0 means these learners exhibit essentially no within-learner variation in burden, they provide no difficulty signal for a ranking task, which motivates us to exclude them from subsequent processing. This leaves 15,553 unique lexeme forms (5,982 shared among two or more learners), 21,262 form–meaning pairs, and 382,697 review logs (190,619 clean).

3.3 RQ1: Limits of Population-Level Features

We test two population-level predictors of PLB, which describe a word's difficulty for learners in general or for learners who share an L1, without using the individual learner's review history: TUBELEX log frequency (Nohejl et al., 2025), an intralexical feature, and a productive difficulty prior that can score arbitrary flashcard content given the learner's L1.

Productive difficulty prior. We fine-tune mmBERT, a modern multilingual encoder (Marone et al., 2025), on the extended KVL dataset of the BEA 2026 Shared Task (Skidmore et al., 2025; British Council, 2026) (single-word productive difficulty scores from English-L2 learners with Spanish, German, and Chinese L1s), using the soft-target approach and prompt structure of Nohejl et al. (2026). The resulting model, which we call the *productive difficulty prior*, reaches a mean Pearson correlation of .801 and root mean squared error (RMSE) of 1.098 across the three L1s, close to the .833 and .989 reported by Nohejl et al.

RQ1a: Transfer to receptive difficulty. The prior was trained on *productive* difficulty in English, whereas Anki reviews test *receptive* knowledge of other L2s by learners with other L1s. We therefore evaluate it on MultiLS-Japanese (Nohejl et al., 2024), which contains 570 Japanese lexical units presented in context, each rated on a 1–5 scale by 10 annotators with L1s spanning English, Portuguese, French, Spanish, Swedish, and Basque. We use the `full_l2_deploy` prompt (Appendix D.6) because it mirrors an Anki review, showing only the Japanese word and its gloss in the learner's L1. Of six prompt variants (Table 5 in Appendix D), it ranks second by Pearson correlation. The best, `full_l2` (Appendix D.2), reaches $r = .375$ but also requires a sentence

context and a clue, which Anki notes do not reliably provide. On the 491 single words, the prior reaches $r = .334$ ($R^2 = .112$) against aggregated complexity, well below TUBELEX log frequency ($R^2 = .41$, consistent with Nohejl et al., 2024). Productive-to-receptive transfer is therefore limited. Frequency's advantage, however, depends on aggregation: against individual annotators' ratings, its fit falls to $R^2 = .13$ (Nohejl et al., 2024).

RQ1b: Item-level ceiling. PLB is highly idiosyncratic (Figure 1 and Figure 3 in Appendix H). Before testing features on it, we therefore estimate how much of its variance reflects stable differences between lexical items. For a learner i and lexical item j, we decompose PLB as follows:

$$\mathrm{PLB}_{ij} = \underbrace{\mu}_{\text{global baseline}} + \underbrace{\alpha_i}_{\text{learner effect}} + \underbrace{\beta_j}_{\text{item effect}} + \underbrace{\gamma_{ij}}_{\text{learner}\times\text{item interaction}} + \underbrace{\varepsilon_{ij}}_{\text{residual}}. \quad (1)$$

Only β_j represents a stable difference between lexical items across learners, so it is the only component that intralexical features can explain. Learners agree little on shared lexemes (mean pairwise Spearman $\rho = .161$, equivalent Kendall's $W = .371$; Figure 2 in Appendix B), and a mixed-effects model attributes only 14.84% of PLB variance to the lexical item (per-lexeme intraclass correlation). For our cohort, this is the ceiling for intralexical features, which are computed from the L2 word alone and so take the same value for every learner. However, the productive difficulty prior is not bound by it, because it also receives the learner's L1 and an L1 gloss. Therefore, it is interlexical under our criterion (Section 2) and can in principle also explain the part of γ_{ij} shared by learners with the same L1. However, because each learner–item pair has only one PLB observation, we cannot separately estimate a learner-by-item interaction and residual error.

RQ1c: Frequency and the prior. TUBELEX log frequency itself falls far short of this ceiling: it explains 2.6% of the lexeme-level variance, roughly 1% of total variance, and correlates with PLB at $\rho = .012$ within learners and $\rho = -.049$ across shared lexemes. Conditioning on L1 does not help either: the productive difficulty prior, given the first L1 each learner reported (Table 6), is equally uninformative ($\rho = -.027$).

In our cohort, intralexical features could thus explain at most 14.84% of PLB, frequency explains almost none of it, and L1-level conditioning adds no detectable signal. This motivates the individually personalized features examined in RQ2.

3.4 RQ2: Interlexical Transfer from Known Words

Neighbours. Our neighbour-informed models use a target word's neighbours: words the learner began studying earlier that are similar to the target along one or more of the similarity channels.

Therefore, to select neighbours, for a target word w_T and a neighbour w_N, we compute five similarity scores:

- *Semantic* (`sem_sim`): cosine similarity between the two words' definition embeddings from Multilingual E5 Base (Wang et al., 2024).
- *Morphological* (`morph_sim`): because Japanese kanji are morphographic, we segment each word into characters with Fugashi (McCann, 2020) and score character overlap.
- *Kanji orthographic* (`ortho_stroke_sim`): stroke-based edit distance computed with Niteru (Fleming, 2026). We additionally record a coverage indicator (`ortho_stroke_coverage`) marking how much of each pair the metric could score.
- *Kana orthographic* (`ortho_kana_confusion`): visual confusability derived from the Japanese letter-similarity data of Higuchi and Kobayashi (2023). Same as for kanji, we record a coverage indicator (`ortho_kana_coverage`) marking how much of each pair the metric could score.
- *Phonological* (`phon_sim`): both words are transcribed to the International Phonetic Alphabet (IPA) with Epitran (Mortensen et al., 2018) and compared using PanPhon's weighted feature edit distance (Mortensen et al., 2016), which weights phoneme differences by subjective distance, aligning with the confusion patterns Laufer documented for synforms (Laufer, 1988).

For each similarity channel, we estimate the 99.8th-percentile similarity threshold from 50,000

pairs randomly sampled from the cohort's 75.9 million eligible candidate pairs. Each candidate pairs a target with a word that the same learner began studying earlier, and learners are sampled in proportion to their numbers of candidate pairs. We then retain the target's top $K = 6$ neighbours above the threshold in each channel, yielding at most 30 neighbours across the five channels. We chose $K = 6$ heuristically to impose this limit.

Splits. The collected dataset was split into training, evaluation, and test sets (70/15/15) per learner, resulting in 11,337 training rows and 2,327 rows each for evaluation and testing. Learners L01, L02, and L03 were included in all splits, while L06 was included only in the training set because the number of uncensored words (480; Table 4 in Appendix C) was insufficient to calculate reliable confidence intervals (CIs).

Ablations. To test whether a learner's observed burden on known words provides predictive signal for unseen words through interlexical similarity, we construct a sequence of ablations that progressively introduces learner history, target-word features, and neighbour information. The models fall into three groups. In model names, `gbdt` and `st` denote the two model families introduced below, `w` stands for *with*, `nbrs` for *neighbours*, and `cen` for *censored*. Table 1 lists the inputs of each model. All learned models receive the same target-word features — the productive difficulty prior score with the prompt and TUBELEX log frequency — along with the within-learner median PLB, which is also available to the baseline.

Constant (median-based). These establish how much can be predicted without lexical information: (1) `global_median`; (2) `within_learner_median`, which captures the learner's overall burden scale and serves as our primary baseline; and (3) `learner_recent_k_median` ($k = 20$), which predicts the median log PLB of up to twenty of the learner's most recently started training items whose PLB had already been observed by the target's first-review date. If no such items exist, it falls back to the learner's overall median. This tests whether recent performance provides predictive signal beyond the overall scale.

Gradient-boosted decision trees (GBDT; XGBoost-based, Chen and Guestrin, 2016). This family progressively introduces lexical information while representing neighbours through fixed, hand-designed aggregates. (4) `gbdt_no_nbrs` uses target-word features only, testing whether word-level properties of the target improve over the learner-level baseline; (5) `gbdt_w_nbrs` adds aggregates over uncensored, interlexically similar neighbours, testing whether observed burden on known words transfers to unseen words through interlexical similarity; (6) `gbdt_w_nbrs_random` computes the same aggregates after replacing each similar neighbour with a random word from the same learner completed around the same time (within one month); (7) `gbdt_w_nbrs_count_only` keeps only neighbourhood size, the number of distinct similar neighbours; (8) `gbdt_w_nbrs+cen` adds censored similar neighbours, testing whether incompletely observed items provide useful information beyond neighbours whose final PLB is already known; and (9) `gbdt_days` removes the neighbourhood altogether and instead gives the model the target's position on the learner's timeline — the number of days between the learner's earliest training item and the target's first review.

Ablations (6), (7), and (9) follow from the observation that PLB declines over time (Figure 4 in Appendix I). `gbdt_w_nbrs_random` asks whether neighbour similarity matters or whether any set of items completed by the learner in the same period serves equally well as a set of progress markers. `gbdt_w_nbrs_count_only` and `gbdt_days` ask whether even those markers are necessary, or whether neighbourhood size or the target's direct offset from the beginning of the learner's training history is sufficient. Neighbourhood size correlates with days elapsed since the learner's earliest training item at 0.764, 0.848, and 0.675 for L01–L03, respectively. Thus, `gbdt_days`, in this case, asks the same question as `gbdt_w_nbrs_count_only` without routing it through neighbour selection at all.

Set-based (SetTransformer, ST; Lee et al., 2019). This family mirrors the GBDT ablations while representing neighbours individually rather than through fixed aggregates. (10) `st_no_nbrs` serves as the architectural control for `gbdt_no_nbrs`; (11) `st_w_nbrs` introduces individual similar neighbours and tests whether learned attention-based weighting improves on GBDT's hand-designed aggregation; and (12) `st_w_nbrs+cen` tests the same hypothesis as

Model	Median	Target	Similar	Censored	Random	Count	Days	# Features
(1)–(3) Constant baselines	–	–	–	–	–	–	–	–
(4) gbdt_no_nbrs	✓	✓						3
(5) gbdt_w_nbrs	✓	✓	✓					23
(6) gbdt_w_nbrs_random	✓	✓			✓			23
(7) gbdt_w_nbrs_count_only	✓	✓				✓		4
(8) gbdt_w_nbrs+cen	✓	✓	✓	✓				38
(9) gbdt_days	✓	✓					✓	4
(10) st_no_nbrs	✓	✓						3
(11) st_w_nbrs	✓	✓	✓					3 + 13 per neighbour
(12) st_w_nbrs+cen	✓	✓	✓	✓				3 + 14 per neighbour

Table 1: Inputs of each RQ2 model. Median: within-learner median PLB; Target: target-word features (productive difficulty prior, TUBELEX log frequency); Similar / Censored: aggregates (GBDT) or per-neighbour features (ST) over uncensored / censored similar neighbours; Random: the same aggregates as (5), computed over temporally matched random neighbours; Count: number of distinct similar neighbours; Days: days from the learner's earliest training item to the target's first review. All neighbour features are computed at the target's first-review date. Individual features are defined in Table 8 (Appendix G).

gbdt_w_nbrs+cen, but under the set-based representation.

For all neighbour-based models, all neighbour scores (e.g., PLB, FSRS stability, FSRS difficulty) are calculated at the cutoff, defined as the target item's first-review date. Thus, only information available by the time the target is first reviewed is used, preventing leakage of future outcomes into earlier predictions.

Metrics. All metrics are computed within learner. We measure ranking with Spearman correlation and regression error with the mean absolute error (MAE) of log-transformed PLB, $\mathrm{MAE}(\log(1+X))$, where X denotes PLB; the log transform counts the gap between 1 and 3 reviews for more than that between 21 and 23. We report regression gains as the *relative error reduction* (RER) against the within-learner median, $\mathrm{RER} = 1 - \mathrm{MAE}_{\text{model}} / \mathrm{MAE}_{\text{baseline}}$, and compare two models by $\Delta\mathrm{RER}$, the difference between their RERs. We compute 95% CIs by paired bootstrapping over test items. A $\Delta\mathrm{RER}$ whose interval excludes zero indicates a statistically significant difference.

Results. Table 2 reports Spearman correlation and RER for all models, and Table 3 reports $\Delta\mathrm{RER}$ for the paired comparisons that isolate each model component.

Moving down the ablation ladder, the results first confirm that our target-word features provide little signal beyond the within-learner median, which every model receives as a feature. Neither gbdt_no_nbrs nor its SetTransformer counterpart improves reliably over the baseline for any learner — all $\Delta\mathrm{RER}$ CIs include zero (Table 3).

Adding uncensored similar neighbours improves both ranking and error for L01 and L02 (Table 2), and the gain over the no-neighbour models is significant for both learners under either architecture. L03 does not benefit, and under the SetTransformer, neighbours slightly but significantly reduce its accuracy.

However, we cannot attribute the gains to interlexical similarity. gbdt_w_nbrs_random, which preserves the neighbourhood's size, recency, and accrued PLB but replaces its words with random ones, performs almost identically to gbdt_w_nbrs for all three learners, with every $\Delta\mathrm{RER}$ CI including zero (bounds between -1.7 and $+1.9$ percentage points, pp). The signal in the neighbourhood features therefore appears to come from the target's position within the learner's history rather than from its similarity to previously studied words.

Isolating neighbourhood size supports this interpretation. gbdt_w_nbrs_count_only, which retains only the number of neighbours, matches the full similar neighbourhood for L01 and L02 and marginally outperforms it for L03. Because neighbourhood size correlates strongly with the number of days elapsed since the learner's earliest training item, we also evaluated gbdt_days, which uses elapsed days directly. It is the strongest model overall: it achieves the highest RER for all three learners (up to $+16.2\%$) and the highest Spearman correlation for L01 and L03, is on par

Model	L01		L02		L03	
	ρ	RER	ρ	RER	ρ	RER
Constant baselines						
Global median	–	0.0% (0.0, 0.0)	–	−9.6% (−12.9, −6.3)	–	−38.2% (−44.6, −31.8)
Recent-20 median	.138 (.065, .208)	−10.5% (−16.4, −4.8)	−.302 (−.365, −.229)	−30.0% (−35.1, −25.3)	.042 (−.029, .111)	−12.5% (−18.5, −6.7)
GBDT						
No neighbours	.170 (.099, .239)	+1.6% (−1.1, +4.1)	.203 (.134, .273)	+2.9% (−0.0, +5.8)	.101 (.035, .166)	+0.9% (−0.9, +2.6)
+ similar neighbours	.439 (.375, .495)	+9.9% (+6.7, +12.9)	.355 (.290, .412)	+8.4% (+5.0, +11.6)	.177 (.109, .241)	−0.5% (−2.2, +1.2)
+ random neighbours	.430 (.369, .482)	+9.6% (+6.4, +12.5)	.367 (.301, .426)	+8.8% (+5.3, +12.1)	.169 (.102, .237)	−0.0% (−1.5, +1.5)
+ neighbour count only	.410 (.346, .467)	+9.0% (+5.6, +12.0)	.319 (.251, .383)	+7.4% (+3.8, +10.8)	.171 (.105, .238)	+0.7% (−0.8, +2.4)
+ sim. and cen. neighbours	.425 (.362, .479)	+10.3% (+6.9, +13.4)	**.414** (.351, .479)	+11.1% (+7.0, +15.2)	.174 (.110, .233)	−1.0% (−2.7, +0.6)
+ days elapsed	**.557** (.498, .608)	**+16.2%** (+12.3, +19.5)	.412 (.351, .473)	**+12.2%** (+8.2, +16.2)	**.269** (.204, .333)	**+3.6%** (+1.6, +5.6)
SetTransformer						
No neighbours	.210 (.139, .279)	+2.3% (−0.0, +4.6)	.208 (.138, .278)	+2.8% (−0.1, +5.6)	.166 (.100, .233)	+0.5% (−0.2, +1.2)
+ similar neighbours	.437 (.373, .492)	+9.8% (+6.0, +13.2)	.347 (.283, .406)	+7.1% (+2.4, +11.1)	.186 (.121, .255)	−0.8% (−2.0, +0.3)
+ sim. and cen. neighbours	.460 (.396, .513)	+11.2% (+7.7, +14.2)	.408 (.347, .474)	+10.4% (+6.3, +14.6)	.158 (.085, .224)	−0.9% (−2.0, +0.2)

Table 2: Within-learner predictive performance by model ablation. GBDT denotes gradient-boosted decision trees. ρ denotes Spearman rank correlation and RER denotes relative error reduction in $\mathrm{MAE}(\log(1+X))$ against the within-learner median baseline, whose absolute $\mathrm{MAE}(\log(1+X))$ is 0.616, 0.517, and 0.373 for L01–L03 (4.62, 3.81, 1.48 reviews on the raw scale). Values in parentheses are paired-bootstrap 95% confidence intervals (CIs). Bold indicates the best result for each learner and metric.

with the best for L02, and significantly outperforms `gbdt_no_nbrs`, `gbdt_w_nbrs`, and `gbdt_w_nbrs_count_only` for every learner.

That said, censored neighbours provide a modest, learner-dependent additional benefit. They help only L02, where the gain is significant under both architectures but borderline, with lower CI bounds of $+0.1$ and $+0.3$ pp; for L01 and L03, the differences are not significant.

Finally, the choice of architecture has little effect: `st_w_nbrs` and `gbdt_w_nbrs` show no significant difference for any learner.

L03 remains an exception to the neighbour effect: none of the neighbour-based variants improves meaningfully over the within-learner median for it, and only timeline position yields a small but significant gain.

4 Discussion

Performance of population-level features. In our main experiments, we used two population-level features of the target word, but neither TUBELEX log frequency nor the productive difficulty prior was informative about PLB, and adding both to the within-learner median gave no significant improvement, although frequency predicts receptive population-level difficulty well ($R^2 = .41$). We see two reasons for this. First, both features describe an average learner, or at best an average learner with a given L1, whereas PLB describes an individual one, and averaging by itself removes the idiosyncrasy (Nohejl et al., 2024) that PLB retains. Even on MultiLS-Japanese, frequency falls from $R^2 = .41$ on aggregated judgments to $R^2 = .13$ at the individual-annotator level (Nohejl et al., 2024), and for PLB the lexical item itself accounts for at most 14.84% of the variance, which bounds what any non-personalized feature can explain. Second, both features estimate whether a learner is likely to already *know* a word, not the *effort* required to learn it through deliberate review — the resources behind the prior share this focus, as MultiLS-Japanese was built for lexical simplification rather than language learn-

Question	Comparison	L01	L02	L03
Target-word signal	GBDT no neighbours − learner median	+1.6 (−1.1, +4.1)	+2.9 (−0.0, +5.8)	+0.9 (−0.9, +2.6)
	ST no neighbours − learner median	+2.3 (−0.0, +4.6)	+2.8 (−0.1, +5.6)	+0.5 (−0.2, +1.2)
Neighbour signal	GBDT similar − no neighbours	**+8.4** (**+5.5**, **+10.9**)	**+5.5** (**+3.1**, **+7.8**)	−1.3 (−2.7, +0.1)
	ST similar − no neighbours	**+7.5** (**+3.5**, **+10.9**)	**+4.3** (**+0.9**, **+7.6**)	**−1.3** (**−2.4**, **−0.3**)
Similarity effect	GBDT similar − random	+0.4 (−1.3, +1.9)	−0.4 (−1.7, +0.9)	−0.4 (−1.6, +0.7)
Neighbourhood size	GBDT count only − no neighbours	**+7.4** (**+4.9**, **+10.0**)	**+4.5** (**+2.5**, **+6.5**)	−0.1 (−0.8, +0.6)
	GBDT count only − similar	−0.9 (−2.4, +0.7)	−1.0 (−2.6, +0.6)	+1.2 (+0.0, +2.4)
Timeline position	GBDT days − no neighbours	**+14.6** (**+11.1**, **+17.7**)	**+9.3** (**+5.9**, **+12.6**)	**+2.7** (**+1.2**, **+4.3**)
	GBDT days − count only	**+7.2** (**+4.4**, **+9.8**)	**+4.8** (**+1.7**, **+8.0**)	**+2.8** (**+1.4**, **+4.3**)
	GBDT days − similar	**+6.2** (**+3.4**, **+8.9**)	**+3.8** (**+0.8**, **+6.7**)	**+4.0** (**+2.5**, **+5.7**)
Censoring effect	GBDT similar+censored − similar	+0.3 (−1.6, +2.1)	**+2.7** (**+0.1**, **+5.5**)	−0.6 (−1.7, +0.5)
	ST similar+censored − similar	+1.4 (−0.9, +3.9)	**+3.4** (**+0.3**, **+6.8**)	−0.0 (−0.9, +0.9)
Aggregation effect	ST similar − GBDT similar	−0.1 (−2.8, +2.2)	−1.3 (−3.7, +1.1)	−0.4 (−1.8, +1.1)

Table 3: Paired comparisons between RQ2 ablations (GBDT: gradient-boosted decision trees; ST: SetTransformer). Each cell reports the difference in relative error reduction (ΔRER) in percentage points, with paired-bootstrap 95% confidence intervals (CIs) in parentheses. Positive values favour the first model in each comparison; negative values favour the second. Bold indicates that the 95% CI excludes zero.

ing (Alfter, 2025). Moreover, spaced repetition further weakens the usual link between frequency and ease: FSRS schedules every card toward the same target retention regardless of how often the word occurs outside Anki, and learners add cards for words they have not yet acquired, so the frequent words that become cards are disproportionately those that resisted incidental learning.

Performance of personalized features. Our main question was whether a learner's observed burden on *known words* transfers to *unseen words* through interlexical similarity. Our neighbour-informed models improved prediction for L01 and L02, but neighbours selected by semantic, morphological, orthographic, or phonological similarity performed no better than random words the learner completed in the same period. Each subsequent ablation removes more neighbour-specific information, yet performance holds or improves. Neighbourhood size alone performs as well as the full similar neighbourhood, and replacing the neighbourhood with the number of days elapsed since the learner's earliest training item, which correlates with neighbourhood size at .68–.85 across learners, yields the best model overall: `gbdt_days` achieves the highest RER for all three learners and Spearman correlations of .557, .412, and .269 for L01–L03. We therefore interpret the neighbourhood as a proxy for how far the learner has progressed rather than as a description of the target's lexical environment. The longer a learner has studied, the more words can pass the similarity threshold, so neighbourhood size grows with time, and random neighbours drawn from the same period carry the same temporal information.

Why timeline position predicts PLB. All GBDT and SetTransformer models receive the learner's median PLB on the training set, and because they are trained on data pooled across four learners, this feature effectively identifies the learner. Combined with the target's position on the timeline, it lets a model learn each learner's PLB trajectory rather than a single PLB level. Consistent with this, the within-learner median accounts

for 70.9% of the total gain in `gbdt_days`.

This matters in our case because PLB declines markedly over time for L01 and L02 but remains nearly flat for L03 (Figure 4), so for the first two learners, the date a word enters study already carries a large share of the rankable signal. This pattern is also mirrored in where neighbours helped: they improved prediction for L01 and L02 but not for L03, where they even slightly reduced accuracy for the SetTransformer. L03 also has the narrowest burden spread of the evaluated learners (interquartile range, Q_1–Q_3: 2–5 reviews), which leaves less variance to explain. Days elapsed nevertheless improves prediction for L03 slightly but significantly (+2.7 pp over `gbdt_no_nbrs`), whereas neighbourhood size does not, suggesting that direct timeline position also captures period effects, such as changes in study intensity or scheduler settings, that a noisier neighbourhood size cannot.

Why similarity carried no signal. The null similarity effect in our cohort does not show that interlexical effects are absent. We hypothesize that our features did not represent them in the form in which they affect learning. The literature reviewed in Section 2.2 predicts both facilitation (cognates, shared morphemes, meaningful associations) and interference (synforms, false friends, near-synonyms), and locates difficulty where form similarity and meaning similarity diverge. Our GBDT features aggregate each channel independently, so a form-similar neighbour that facilitates learning because its meaning is related and one that interferes because its meaning is unrelated enter the aggregates in the same way, and their effects may cancel out. The SetTransformer sees per-neighbour channel percentiles and could in principle represent such mismatches, but did not learn to do so at the present data scale. Moreover, a neighbour's PLB records how hard that word was for the learner, not whether it was confused with the target, and review logs do not record which word a learner mistook for which. We therefore hypothesize that similarity channels alone do not expose confusion patterns to the model.

How censoring and architecture affect performance. Including censored neighbours significantly improved prediction only for L02, and only marginally (+2.7 pp for GBDT and +3.4 pp for the SetTransformer, with lower CI bounds near zero). Because censored items tend to be the slowest to master, this suggests that they carry some information about the difficult tail that uncensored neighbours miss, though not much at this sample size. As for the architecture, the SetTransformer did not reliably outperform GBDT despite representing neighbours individually. We hypothesize that attention-based weighting has little to make use of when the useful signal is the neighbourhood as a whole, as a progress marker, rather than lying in individual neighbours, and that 11,337 training rows are too few for its additional flexibility to pay off.

Why absolute prediction remains hard. Even our best model ranks items only moderately well ($\rho \leq .557$) and reduces log-scale error by at most 16.2% relative to the within-learner median. Part of the reason lies in the PLB itself, which is an intrinsically noisy target: unlike population-level difficulty, it depends on the learner's behaviour, memory state, and scheduler configuration, each stochastic lapse adds reviews before the target is reached, and PLB shifts within a learner over time (Section 5).

5 Limitations and Future Work

Data and scope. Our evaluation covers three learners, one target language, and single-word lexemes only. The learners differ in L1 and self-rated proficiency (Table 6), but our cohort is too small to estimate the effect of either. The only L1-conditioned feature we test, the productive difficulty prior, is uninformative (Section 3.3). Multiword expressions, which appear both in learners' decks and in MultiLS-Japanese, are excluded. We also predict PLB only for unseen words of known learners: with three evaluable learners, a leave-one-learner-out evaluation of how the models generalize to unseen learners is not feasible. Our released dataset also covers learners of Italian, Spanish, and Portuguese, but both extensions require a larger cohort.

Scalability. Computing PLB requires no annotation and a single pass over each card's review log (Algorithm 1), so its cost grows linearly with the number of reviews: processing our 503,826 review logs sequentially takes 1 min 34 s on an Intel Core i5-13600K CPU. Embedding the definitions of all 42,369 form–meaning pairs with Multilingual E5 Base takes 1 min 37 s on an NVIDIA GeForce RTX 4070 Ti (batch size 64). The dominant cost is neighbour selection: each target is compared

with every word the same learner began studying earlier, so the number of candidate pairs grows quadratically with a learner's deck size but only linearly with the number of learners (75.9 million pairs for our cohort, 30 min with 20 parallel workers on the same CPU). Training a GBDT or Set-Transformer model on our 11,337 training rows takes about a minute. Adding learners is therefore limited by data collection rather than computation.

Temporal confounds. PLB declines over time for L01 and L02 (Figure 4). L01's per-word PLB, for instance, reached 88 reviews early in the collection window but fell below 10 by its end. We did not test what drives this decline. It may reflect growing proficiency (e.g., a larger stock of known kanji), a change in the kinds of words added, or right-censoring, since an item first reviewed late in the window can reach the target before the window closes only if it needs few reviews. It may also reflect the scheduler: Anki lets learners change FSRS settings and re-optimize its parameters at any time, but our data do not record when such changes occurred, so we cannot tell how they shaped PLB. More generally, PLB depends on the scheduler's configuration, as the exclusion of L04 and L05 shows, and on the accuracy of FSRS's stability estimates. Future work could reduce this dependence, for example by counting only reviews taken close to their scheduled interval (e.g., within 0.8–$1.25\times$), so that overdue reviews after a break do not inflate stability gains. Overall, separating these explanations is necessary before the timeline effect can be read as learning progress.

Features and representation. Our target-word representation is limited to two features, one of which is a prior trained on productive rather than receptive difficulty, for a different L2 and different L1s than ours. Other factors reviewed in Section 2.1, such as polysemy or figurativeness, remain untested. Neighbour selection relies on heuristic choices, the 99.8th-percentile threshold and $K = 6$, whose sensitivity we did not analyse. Finally, similarity percentiles and neighbour PLB may be too coarse to capture confusion: they record how similar a neighbour is and how hard it was to learn, but not whether the learner confused it with the target. Three extensions could close this gap. First, data that record confusions directly, such as the wrong answers learners give, which Anki review logs do not store, would show which known words interfere with a given target. Second, more granular representations of each target–neighbour pair, such as part-of-speech tags and the specific kanji, radicals, or readings the two words share or differ in, would let models relate form similarity to meaning similarity and separate neighbours that facilitate learning from those that interfere. Third, a larger dataset would provide more target–neighbour examples and make these finer-grained interactions easier for models to learn.

6 Conclusion

We asked whether a learner's personalized learning burden (PLB) on unseen words can be predicted from the burden they showed on interlexically similar words they already know. Across semantic, morphological, orthographic, and phonological channels and three evaluable learners, we find no evidence that it can.

On our data, the population-level features we tested could not substitute for personalization: the lexical item accounts for at most 14.84% of PLB variance, log frequency explains about 1%, and the L1-conditioned prior, which transferred only weakly to receptive difficulty, is uninformative. Neighbour-informed models improved over the within-learner median for L01 and L02, but temporally matched random neighbours reproduced the gain and neighbourhood size alone matched it. The strongest model used no neighbours at all: days since the learner's earliest training item achieved the highest RER for every learner (up to $+16.2\%$; ρ up to $.557$). We therefore attribute the neighbour gain to the target's position on the learner's timeline rather than to what the neighbourhood contains.

Two implications follow. First, because a neighbourhood drawn from a learner's history encodes timeline position by construction, future tests of interlexical transfer should control for it, as our random-neighbour baseline does. Second, similarity alone may not expose confusion: modelling how form and meaning similarity interact for a particular learner is the more promising next step. To support such work, we release a dataset of approximately 66,000 Anki notes and over 670,000 review logs from eight learners with multi-year histories, each with a profile of their languages and areas of interest.

AI Disclosure

Following the taxonomy of Kamocki and Witt (2026), we used AI tools for *assistive use*, *editing*, *generation of minor elements*, and *compression/expansion*. AI tools were not used to design the study or interpret the results. Our use of AI coding assistants is documented in the repository's README.

Assistive use. Claude Sonnet and Claude Opus (versions 4.6–5.5; Anthropic) were used for topic discovery and literature search (the authors read and verified all cited works). Before submission, GPT-5.6 Sol (OpenAI) reviewed the manuscript to flag remaining inconsistencies. During revision, Claude Opus 5.5 helped address reviewer comments by identifying undefined abbreviations and inconsistencies and proposing table layouts.

Editing. The same Claude models were used for grammar and spelling correction and for improving flow, style, and tone. GPT-5.6 Terra was used to format LaTeX.

Generation of minor elements. During revision, Claude Opus 5.5 drafted short passages, such as the Anki terminology section and table captions, which the authors revised.

Compression/expansion. Claude was used to shorten some author-written sections to meet the page limit.

The authors reviewed all AI-assisted text and take full responsibility for the content of the paper.

Acknowledgements

John P. McCrae is supported by Research Ireland as part of the Rinn AI Centre and Grant Number 13/RC/2106_P2, ADAPT Research Centre.

References

Gehad Abdelrahman, Qian Wang, and Miguel A. Nunes. 2023. Knowledge Tracing: A Survey. *ACM Computing Surveys*, 55(11):1–37.

David Alfter. 2025. The need for truly graded lexical complexity prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 326–333, Vienna, Austria. Association for Computational Linguistics.

Ankitects Pty Ltd. 2026. Anki: Powerful, Intelligent Flashcards. https://apps.ankiweb.net/. Accessed: 2026-06-21.

British Council. 2026. BEA 2026 Shared Task: Vocabulary Difficulty Prediction for English Learners. https://www.britishcouncil.org/data-science-and-insights/bea2026st. Accessed: 2026-06-21.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. ACM.

Hermann Ebbinghaus. 1913. *Memory: A Contribution to Experimental Psychology*. Teachers College, Columbia University, New York. Translated from the original 1885 German edition by Henry A. Ruger and Clara E. Bussenius.

Kyle T. Fleming. 2026. Niteru: Search for Japanese Kanji by Visual Similarity. https://github.com/ktfleming/niteru. Accessed: 2026-07-31.

Sian Gooding and Manuel Tragut. 2022. One Size Does Not Fit All: The Case for Personalised Word Complexity Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365.

Aroline E. Seibert Hanson and Christina M. Brown. 2020. Enhancing L2 learning through a mobile assisted spaced-repetition tool: an effective but bitter pill? *Computer Assisted Language Learning*, 33(1-2):133–155.

Brett J. Hashimoto and Jesse Egbert. 2019. More Than Frequency? Exploring Predictors of Word Difficulty for Second Language Learners. *Language Learning*, 69(4):839–872.

Hiroki Higuchi and Tessei Kobayashi. 2023. Letter visual similarity of Japanese hiragana and katakana based on reaction times. *Current Psychology*, 42(15):12835–12844.

Pawel Kamocki and Andreas Witt. 2026. Authorship attribution in the times of LLMs within the framework of the CRediT taxonomy. In *Proceedings of the Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (LEGAL2026 and CALD-pseudo 2026) @ LREC 2026*, pages 35–39, Palma, Mallorca (Spain). ELRA.

Su Kyung Kim and Stuart Webb. 2022. The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis. *Language Learning*, 72(1):269–319.

Batia Laufer. 1988. The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education*, 2(2):113–132.

Batia Laufer. 1990. Words you know: How they affect the words you learn. In Jacek Fisiak, editor, *Further Insights into Contrastive Linguistics*, pages 573–593. John Benjamins, Amsterdam.

Batia Laufer. 2025. Second-Language Word Difficulty. In *The Encyclopedia of Applied Linguistics*, pages 1–6. John Wiley & Sons, Ltd.

Batia Laufer and Zahava Goldstein. 2004. Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning*, 54(3):399–436.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. *arXiv preprint arXiv:2509.06888*.

Paul McCann. 2020. fugashi, a Tool for Tokenizing Japanese in Python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51, Online. Association for Computational Linguistics.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for Many Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL.

I.S.P. Nation. 2022. *Learning Vocabulary in Another Language*, 3 edition. Cambridge University Press, Cambridge, UK.

Adam Nohejl, Akio Hayakawa, Yusuke Ide, and Taro Watanabe. 2024. Difficult for Whom? A Study of Japanese Lexical Complexity. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 69–81, Miami, Florida, USA. Association for Computational Linguistics.

Adam Nohejl, Frederikus Hudi, Eunike Andriani Kardinata, Shintaro Ozaki, Maria Angelica Riera Machin, Hongyu Sun, Justin Vasselli, and Taro Watanabe. 2025. Beyond Film Subtitles: Is YouTube the Best Approximation of Spoken Vocabulary? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9566–9585, Abu Dhabi, UAE. Association for Computational Linguistics.

Adam Nohejl, Xuanxin Wu, Yusuke Ide, Maria Angelica Riera Machin, Yi-Ning Chang, and Hitomi Yanaka. 2026. Sakura at BEA 2026 Shared Task 1: What Makes Vocabulary Difficult? In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1161–1178, San Diego, California, USA. Association for Computational Linguistics.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.

Open Spaced Repetition. 2024. FSRS-Anki-20k (Revision d2e5351). Hugging Face.

Norbert Schmitt. 2008. Instructed second language vocabulary learning. *Language Teaching Research*, 12(3):329–363.

Burr Settles and Brendan Meeder. 2016. A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer Architectures for Vocabulary Test Item Difficulty Prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Vienna, Austria. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

Jarrett Ye. 2026. The FSRS Spaced Repetition Algorithm. `https://github.com/open-spaced-repetition/awesome-fsrs/wiki/The-Algorithm`. Open Spaced Repetition FSRS wiki; accessed: 2026-06-21.

Junyao Ye, Jingyong Su, and Yilong Cao. 2022. A Stochastic Shortest Path Algorithm for Optimizing Spaced Repetition Scheduling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*, pages 4381–4390, Washington, DC. Association for Computing Machinery.

A Personalized Learning Burden Computation

Algorithm 1 Personalized learning burden of learner i on item j.

Require: review log $r_1, \ldots, r_n$ of the item's card, sorted by time t_k; each r_k has a type τ_k, a rating $g_k \in \{1, 2, 3, 4\}$ (*Again*–*Easy*; Section 3.1); FSRS parameters $\mathbf{w}_i$ of the learner's deck; target stability $S^* = 14$ days
Ensure: PLB_{ij}, or a right-censored lower bound on it

$C \leftarrow \langle\,\rangle$ ▷ clean reviews
for $k = 2, \ldots, n$ **do**
  $\Delta_k \leftarrow t_k - t_{k-1}$ in days
  if $\tau_k = \texttt{review} \wedge \Delta_k \geq 1$ **then**
    append r_k to C
  end if
end for
if $C = \langle\,\rangle$ **then**
  return excluded ▷ backlog: no evidence
end if
$\sigma \leftarrow$ new FSRS memory state
for $m = 1, \ldots, |C|$ **do** ▷ replay clean reviews only
  $\sigma \leftarrow \mathrm{FSRS}_{\mathbf{w}_i}\big(\sigma, g(c_m), t(c_m)\big)$ ▷ c_m: m-th element of C
  if $S(\sigma) \geq S^*$ **then**
    return $\mathrm{PLB}_{ij} = m$
  end if
end for
return censored: $\mathrm{PLB}_{ij} > |C|$

B Learner Agreement on Shared Lexemes

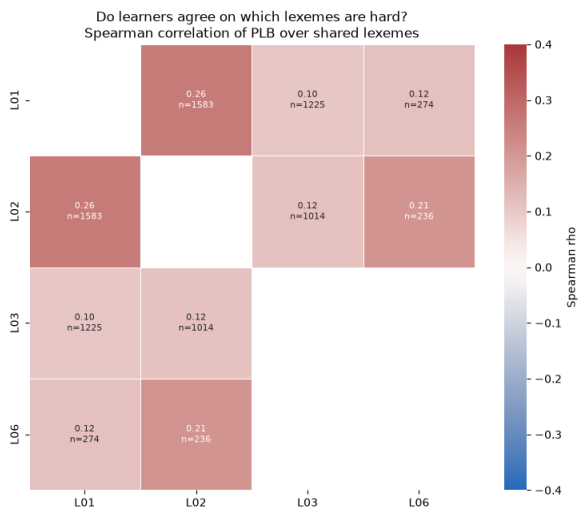


Figure 2: Pairwise Spearman agreement between learners on shared lexemes.

C Censoring and Review-Count Spread

Learner	Reached	Censored	Cens. %	Q_1–Q_3
L01	5,191	3,525	40.44	3–8
L02	4,594	485	9.55	3–10
L03	5,726	876	13.27	2–5
L04	10,612	508	4.57	1–1
L05	1,253	308	19.73	2–2
L06	480	385	44.51	3–8

Table 4: Items that reached the target stability, censored items, censored share, and interquartile range of clean review counts (Q_1–Q_3) by learner, over word instances with at least one clean review.

D Prompt Templates

Table 5 compares the six prompt variants listed below on the 491 single words of MultiLS-Japanese. For prompts using L1-translated context or a gloss, we queried GPT-5.6 Terra with the translation prompt provided in Appendix E.

Rank	Prompt	Pearson	RMSE
1	**`full_l2`**	**.375**	1.050
2	`full_l2_deploy`	.334	1.058
3	`full_l2_no_context`	.317	**1.006**
4	`full_l1`	.192	1.102
5	`ja_word_only`	.164	1.297
6	`l1_name_only`	.151	1.325

Table 5: Pearson correlation and RMSE of the productive difficulty prior against aggregated receptive complexity on MultiLS-Japanese, by prompt variant, ranked by Pearson correlation. Bold marks the best value in each column. `full_l2_deploy` is used in all experiments.

D.1 `full_l1`

```
Rate how easy it is for learners to guess the
↪  Japanese word based on the {l1_name} word,
↪  context and clue on a scale from 1 to 5 (1=very
↪  easy, 5=very difficult).
{l1_name} word: {l1_gloss}
{l1_name} context: {l1_context}
Clue: {clue}
Japanese word: {target}
Difficulty: 1
```

D.2 `full_l2`

```
Rate how easy it is for learners to guess the
↪  {l1_name} word based on the Japanese word,
↪  context and clue on a scale from 1 to 5 (1=very
↪  easy, 5=very difficult).
Japanese word: {target}
Japanese context: {context}
Clue: {l1_clue}
{l1_name} word: {l1_gloss}
Difficulty: 1
```

D.3 `ja_word_only`

```
Rate how difficult the Japanese word is for
↪  learners on a scale from 1 to 5 (1=very easy,
↪  5=very difficult).
Japanese word: {target}
Difficulty: 1
```

D.4 l1_name_only

```
Rate how difficult the Japanese word is for
↪  learners whose native language is {l1_name} on a
↪  scale from 1 to 5 (1=very easy, 5=very
↪  difficult).
Japanese word: {target}
Difficulty: 1
```

D.5 full_l2_no_context

```
Rate how easy it is for learners to guess the
↪  {l1_name} word based on the Japanese word and
↪  clue on a scale from 1 to 5 (1=very easy,
↪  5=very difficult).
Japanese word: {target}
Clue: {l1_clue}
{l1_name} word: {l1_gloss}
Difficulty: 1
```

D.6 full_l2_deploy

```
Rate how easy it is for learners to guess the
↪  {l1_name} word based on the Japanese word on a
↪  scale from 1 to 5 (1=very easy, 5=very
↪  difficult).
Japanese word: {target}
{l1_name} word: {l1_gloss}
Difficulty: 1
```

E L1 Cue-Generation Prompt

```
You create L1-side vocabulary cues for a controlled
↪  cross-lingual vocabulary-difficulty experiment.

The Japanese context contains exactly one span
↪  marked <TARGET>...</TARGET>. For every
↪  requested L1:

1. Translate the complete Japanese context
↪  naturally into that L1 as l1_context,
↪  preserving meaning and register.
2. l1_gloss must represent only the marked target
↪  expression. Never translate or paraphrase the
↪  surrounding sentence or clause into l1_gloss.
3. The literal text between <TARGET> and </TARGET>
↪  is the complete target boundary. Text outside
↪  the tags is context only, even when an adjacent
↪  particle, auxiliary, affix, or punctuation mark
↪  is orthographically attached. Never absorb
↪  outside material or its semantic contribution
↪  into l1_gloss.
4. If the L1 necessarily realizes the marked target
↪  jointly with adjacent outside material, keep
↪  l1_gloss target-only and use discontinuous,
↪  paraphrastic, or uncertain. Never label that
↪  joint realization exact or inflected, and
↪  explain the fusion in alignment_note.
5. Return exactly one best context-specific gloss.
↪  Do not return alternatives separated by
↪  semicolons, slashes, pipes, newlines,
↪  parentheses, or lists.
6. Prefer a common natural equivalent over an
↪  unnatural equivalent selected only to preserve
↪  Japanese part of speech. A single natural
↪  multiword equivalent is allowed.
7. If the marked target has a standalone
↪  context-appropriate L1 equivalent, set
↪  gloss_kind to translation.
8. If the marked target is a function word and no
↪  standalone L1 equivalent exists, set gloss_kind
↪  to metalinguistic. Give one brief L1-language
↪  description of the target's grammatical or
↪  discourse function, set alignment_kind to
↪  not_applicable, and set l1_target_surface to
↪  the empty string. Do not import the speech act
↪  or lexical meaning of the surrounding clause.
9. For translation glosses, use exact or inflected
↪  only when l1_target_surface is a contiguous
↪  realization of the marked target alone in
↪  l1_context, and copy that surface verbatim. Use
↪  discontinuous, paraphrastic, or uncertain when
↪  appropriate.
10. Do not estimate difficulty, simplify for
↪  learners, infer a score, or transliterate
↪  merely to preserve Japanese form.
11. Return exactly one cue for each requested
↪  language and no additional languages.
```

F Data Collection and Release

Participants uploaded their Anki decks (`.apkg` files; Section 3.1) through a web-based survey, in which they also reported their L1s, additional-language proficiencies, and areas of interest. Throughout, participants kept control over what they shared. Decks were parsed in the participant's browser, where they could preview the content and exclude individual notes, fields, and tags before anything was uploaded. Submission required explicit consent to the public release of the notes, cards, and review logs, while releasing the profile information required a separate, optional consent. In the released dataset, identifiers and deck names are replaced with synthetic ones, and each learner's timestamps are shifted by a random constant offset.

The privacy guidelines are available at `https://nekear.me/research/overview`, and the source code of the survey, including the in-browser deck parser, at `https://github.com/nekear/vocabulary-acquisition-survey` (MIT licence).

Table 6 summarizes the eight learners' language profiles and data sizes. The learners of Japanese, L01–L06, form the cohort of our experiments: L04 and L05 are excluded from modelling for lack of within-learner variation (Section 3.2), and L06 is used for training only (Section 3.4). The dataset consists of six files linked by identifiers: learner profiles, decks with their scheduler settings, note types with their field names, notes, cards, and reviews. Table 7 shows one record from each file.

ID	L1	Target (self-rated)	Interests	Notes	Cards	Reviews
L01	fr, en	ja (unsure), nl (unsure)	technology	15,832	30,253	200,496
L02	hi, en	ja (unsure)	travel, academic, business, legal, technology, literature, music, pop culture	6,720	7,113	143,227
L03	en	ja (B2)	travel, pop culture, other	9,658	9,658	77,362
L04	id	ja (B2), en (C1)	technology, pop culture, music	13,561	13,561	116,165
L05	en, fr	ja (B1)	technology	8,653	15,008	7,285
L06	en	ja (A1)	pop culture, technology, literature	2,312	2,312	27,595
L07	zh-Hant, en	it (A2)	technology	8,523	17,248	96,046
L08	en	es (B1), pt (A1)	literature, travel	965	1,106	10,206

Table 6: Learner profiles and data sizes in the released dataset. L1 and target languages are given as ISO 639-1 codes (zh-Hant: Chinese in Traditional script). Target-language proficiency is self-rated on the scale of the Common European Framework of Reference for Languages (CEFR, A1–C2) or reported as unsure. Interests are the domains learners selected in the survey.

File	Example record (learner `L02`)
`learners`	`native_languages: [hi, en]; target_languages: [{lang: ja, proficiency: Unsure}]; domain_interests: [travel, academic, ..., pop_culture]`
`decks`	`deck_id: 7, parent_id: 3, scheduler: FSRS, desired_retention: 0.9, fsrs_weights: [0.033, 0.196, 0.672, ...], target_language: ja`
`note_types`	`model_id: 2, field_names: [VocabKanji, VocabFurigana, VocabDef, SentEng, SentKanji, SentFurigana, ...], excluded_field_indexes: []`
`notes`	`note_id: 1677, model_id: 2, fields: [見 方, 見 方[み か た], view, "I began [...] abroad.", 留学してから、ものの<b>見方</b>が変わった。, ...], tags: [JLPT::N3, 単語N2]`
`cards`	`card_id: 2070, note_id: 1677, deck_id: 7, state: 2, interval: 4, reps_total: 8, lapses_total: 1, fsrs_stability: 4.07, fsrs_difficulty: 8.05`
`reviews`	`card_id: 2070, timestamp_ms: 1733935081425, rating: 3, interval_before: 0, interval_after: 1, time_taken_ms: 5366, review_type: 0` `...29 d → rating: 1 (lapse) → interval_after: −600 s (relearning) → ...`

Table 7: One record from each file of the released dataset, for learner L02: a note, one of its cards, and that card's reviews (...: values shortened). Card `state` and `review_type` follow Anki's codes (Section 3.1): state 2 is a review card and review type 0 a learning review. Positive intervals are in days and negative ones in seconds. `excluded_field_indexes` lists the fields the learner withheld at upload.

G Neighbour Features

Table 8 lists the input features of the learned RQ2 models. *Group* names the feature block and the models that use it, *Features* lists its features, and n gives their number. GBDT models summarize the neighbourhood with fixed-length aggregates, computed either per similarity channel (one feature for each of the five channels) or across channels (over the union of all selected neighbours). Censored neighbours have no observed PLB, so the censored block omits the log-PLB means and instead adds the share of selected neighbours that are censored. Model 7 uses only the distinct-neighbour count of the uncensored block. The SetTransformer receives one feature vector per neighbour, in which `s_missing` marks neighbours not reviewed before the target's first review. All neighbour features are computed at the target's first-review date.

Group	Features	n
Target and learner (all learned models)	Within-learner median PLB; productive difficulty prior score (`full_l2_deploy`); TUBELEX log frequency; days elapsed (model 9 only)	3 (+1)
GBDT, uncensored neighbours (models 5, 6, 8)	Per channel: neighbour count, mean similarity percentile, mean log PLB	15
	Across channels: distinct neighbours, mean best-channel percentile, mean log FSRS stability, mean FSRS difficulty, mean log PLB	5
GBDT, censored neighbours (model 8)	Per channel: neighbour count, mean similarity percentile	10
	Across channels: distinct neighbours, censored share, mean best-channel percentile, mean log FSRS stability, mean FSRS difficulty	5
SetTransformer, per neighbour (models 11, 12)	PLB, review count, log FSRS stability, FSRS difficulty, `s_missing`; similarity percentiles for the five channels, `phon_missing`; kanji and kana orthographic coverage indicators; `is_censored` (model 12 only)	13 (+1)

Table 8: Input features of the RQ2 models, by feature block.

H Kaplan–Meier Burden Distributions

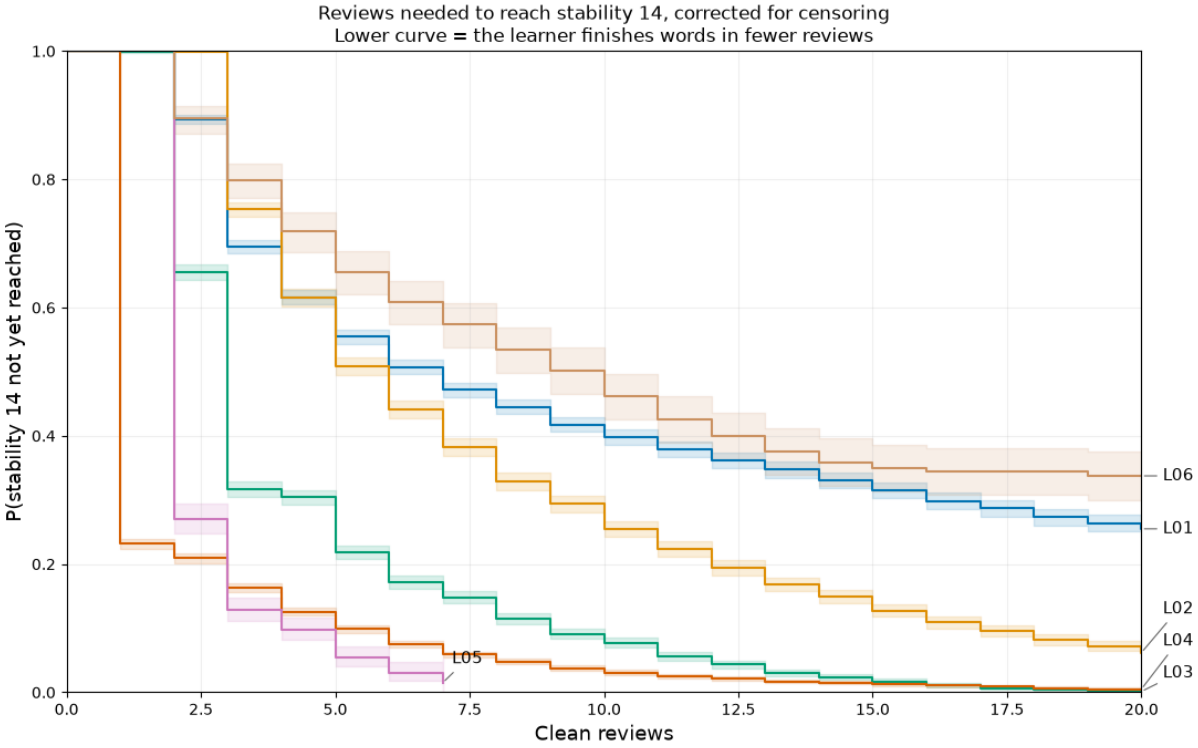


Figure 3: Kaplan-Meier estimates of reviews required to reach the target stability, by learner.

I Temporal Trend in Personalized Learning Burden

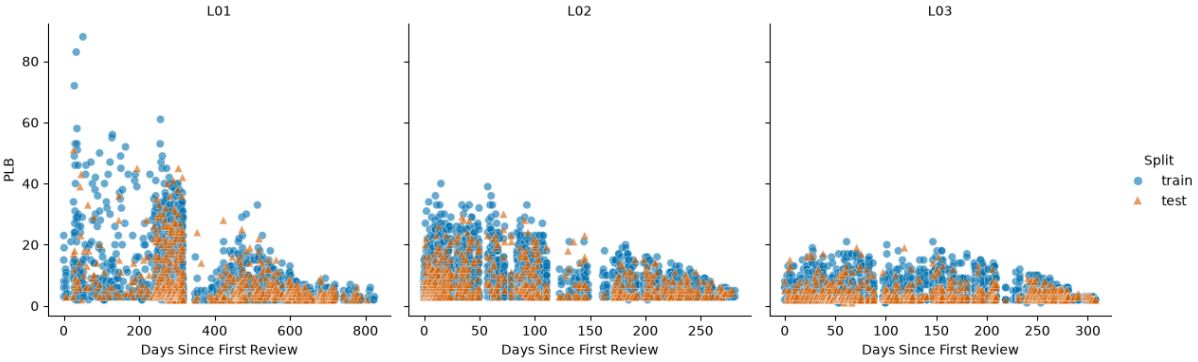


Figure 4: PLB as a function of days since first review for L01–L03, shown separately for training and test items. PLB generally declines over time, motivating the temporally matched random-neighbour control.