JIMU: Jurnal Ilmiah Multidisipliner
Vol. 03, No. 03, Tahun 2025, Hal. 925 -939 e-ISSN: 3031–9498

The Metronome of Memory: Spaced Flip-Card Practice for Indonesian EFL Learners

Estiningtyas Sholikhah[1], Fitrilya Anjarsari[2]

[1]Universitas Muhammadiyah Brebes, [2]Universitas Diponegoro, Indonesia
E-mail: estiningtyas@umbs.ac.id[1] , fanjarsari@live.undip.ac.id[2]

Article Info

Article history:

Received August 09, 2025
Revised August 19, 2025
Accepted August 25, 2025

Keywords:

Vocabulary, Retrieval practice, Spaced repetition, Flashcards, EFL.

ABSTRACT

This paper examines whether flip cards paper or digital meaningfully improve English vocabulary learning for Indonesian students when implemented as disciplined, spaced retrieval. Building on cognitive psychology and L2 vocabulary research, we synthesise evidence that practice testing and distributed practice yield stronger long-term retention than re-study or massed drills. We then review Indonesia-based classroom studies and translate convergent findings into a feasible school routine: three short sessions weekly, expanding intervals (1–3–7–14 days), and prompts that move from recognition to productive recall through collocation and word-family cues. Methodologically, we outline a quasi-experimental protocol for secondary schools with three conditions (digital SRS, paper with printed schedules, business-as-usual), pretest/post-test/delayed post-test, and ANCOVA with cluster-robust errors. The primary outcome is durable receptive and productive vocabulary; secondary outcomes include adherence and learner attitudes. Our synthesis indicates that flip cards help only insofar as they operationalise retrieval and spacing; the medium is secondary to schedule fidelity. For Indonesia, hybrid deployment (printable decks plus mobile SRS where available) addresses equity, workload, and connectivity. We conclude with implementation guardrails tight decks, routine low-stakes quizzes, and transfer tasks so that cards become metronomes: steady rhythms that carry words from fleeting exposure into usable command. Implications for curriculum alignment and teacher training are noted.

Article Info

Article history:

Received August 09, 2025
Revised August 19, 2025
Accepted August 25, 2025

Keywords:

Kosakata, Praktik Retrieval, Pengulangan Berjeda, Flashcards, EFL.

ABSTRAK

Artikel ini menelaah apakah kartu balik kertas maupun digitalbe nar-benar meningkatkan pembelajaran kosakata bahasa Inggris pada siswa Indonesia ketika dijalankan sebagai praktik pengambilan-kembali (retrieval) yang terjadwal secara berjeda. Berbekal psikologi kognitif dan riset kosakata L2, kami mensintesis bukti bahwa tes diri dan latihan terdistribusi menghasilkan retensi jangka panjang yang lebih kuat dibanding membaca ulang atau latihan menumpuk. Kami meninjau studi kelas di Indonesia dan menerjemahkan temuannya menjadi rutinitas sekolah yang layak: tiga sesi singkat per minggu, interval berkembang (1–3–7–14 hari), serta butir yang bergerak dari pengenalan menuju produksi melalui kolokasi dan keluarga kata. Secara metodologis, kami menguraikan protokol kuasi-

eksperimental untuk jenjang menengah dengan tiga kondisi (SRS digital, kartu kertas dengan jadwal cetak, praktik biasa), pra-tes/pasca-tes/pasca-tes tunda, dan analisis ANCOVA dengan kesalahan robust berkelompok. Luaran utama ialah kosakata reseptif dan produktif yang bertahan; luaran sekunder meliputi kepatuhan dan sikap belajar. Sintesis menunjukkan kartu balik efektif sepanjang mengoperasionalkan retrieval dan jeda; medium hanya sekunder terhadap kedisiplinan jadwal. Untuk konteks Indonesia, penerapan hibrida (dek cetak serta SRS seluler bila tersedia) menjawab persoalan pemerataan, beban guru, dan konektivitas. Kami menutup dengan rambu implementasi dek ramping, kuis berisiko rendah yang rutin, dan tugas transfer agar kartu menjadi metronom: ritme ajek yang menggeser kata dari paparan singkat menuju kemahiran berkelanjutan.

Corresponding Author:

Estiningtyas Sholikhah
Universitas Muhammadiyah Brebes
E-mail: estiningtyas@umbs.ac.id

Introduction

Vocabulary sits at the core of second/foreign language proficiency. As Wilkins put it half a century ago, “without grammar very little can be conveyed; without vocabulary nothing can be conveyed” (Wilkins, 1972 p. 111). Contemporary teacher education, a parallel consensus holds that “vocabulary is a core component of language proficiency and provides much of the basis for how well learners speak, listen, read, and write” (Richards & Renandya, 2002, p. 255). For Indonesian EFL learners who encounter English primarily in instructed settings these claims crystallize a practical imperative: scalable, classroom-friendly methods that accelerate durable vocabulary growth.

This article investigates one such method: flip cards (i.e., paper or digital word cards/flashcards). Flip cards are deceptively simple, yet they operationalize two learning principles that the cognitive-science literature repeatedly identifies as high-yield: retrieval ractice and distributed (spaced) practice. Roediger and Butler summarize the first succinctly: “Retrieval practice produces greater long-term retention than studying alone” (2011, p. 25). Dunlosky et al. likewise conclude from a broad review that “practice testing and distributed practice received high utility assessments” (2013, p. 5).

Within L2 vocabulary scholarship, Nation frames what must be learned: “At the most general level, knowing a word involves form, meaning, and use” (Nation, 2013, p. 49). That multidimensional target implies a need for repeated, varied encounters and active processing. In a practitioner-oriented synthesis, Nation adds two implementation rules of thumb directly relevant to flip-card practice: first, “the overwhelming finding is that spaced repetition results in better long-term retention than massed repetition” (2017, p. 2); second, “one-quarter of the time in a well-balanced course should be spent on deliberate learning [including] individualised independent vocabulary learning using flashcards” (2017, p. 4). These statements connect a classroom routine (word cards) to a course-design principle (the Four Strands) that advocates a balanced allocation across meaning-focused input and output, language-focused learning, and fluency development. Although the Four Strands framework

was formulated for diverse contexts, its emphasis on deliberate vocabulary study provides a clear rationale for employing flip cards in Indonesia's time-constrained EFL classrooms. (For the Four Strands overview, see Nation, 2007.)

At the same time, skepticism about flip cards persists. Critics worry that rote pair-association could foster shallow form–meaning links and neglect collocation, register, or productive use. The research base, however, has steadily moved beyond a simple "rote vs. rich" dichotomy. Nation's taxonomy of learning conditions explicitly embeds retrieval, varied meetings, and use as quality-of-processing levers (2017, pp. 5–6), which can be engineered into flip-card practice (e.g., prompt both L1→L2 and L2→L1 retrieval; cycle inflections/derivatives; add example sentences). Moreover, from a cognitive-task perspective, well-designed flip-card schedules mobilize the very techniques Dunlosky et al. rate most effective and caution against those they rate least effective (e.g., "highlighting" and "rereading" receive low-utility ratings) (2013, p. 5).

The Indonesian context also makes the case compelling. Classroom studies ranging from vocational high school ESP settings to elementary and junior-secondary cohorts report measurable vocabulary gains when teachers integrate flashcards into instruction or homework, including digital implementations such as Anki/Quizlet that automate spacing algorithms. Although designs vary (from one-shot case studies to classroom action research), the pattern is consistent: targeted card-based practice supports uptake and short-term retention, and digital spacing appears to aid maintenance. These local findings align with international evidence that flashcards are an efficient conduit for deliberate vocabulary learning, particularly for high-frequency lexical targets and technical wordlists.

Still, two unresolved questions justify closer study in Indonesian EFL classrooms. First, *how* much of the documented benefit derives from retrieval practice and spacing per se (which flip cards can deliver) versus from other co-occurring supports (e.g., teacher explanations, example sentences, pictorial cues)? Second, *which* design features (paper vs. digital, fixed vs. adaptive spacing, single words vs. collocations, receptive vs. productive prompts) best promote movement from recognition to use the trajectory Nation's "form-meaning-use" model would predict? The present research addresses these questions by examining learners' vocabulary growth under a flip-card condition deliberately structured to instantiate (a) retrieval practice, (b) spaced scheduling, and (c) incremental expansion from single-word recognition to constrained productive use.

In short, this study proceeds from three premises. First, vocabulary growth is a strategic bottleneck for Indonesian learners and a central curricular goal (Wilkins, 1972, p. 111; Richards & Renandya, 2002, p. 255). Second, flip cards properly designed instantiate "high utility" learning techniques (Dunlosky et al., 2013, p. 5) and align with the Four Strands' call for deliberate study time (Nation, 2017, p. 4). Third, vocabulary knowledge is multifaceted; therefore effectiveness should be judged not only by receptive gains but by progress toward use (Nation, 2013, p. 49). Against this backdrop, we test whether flip-card practice yields practically meaningful improvements in Indonesian EFL learners' vocabulary knowledge and retention relative to typical alternatives, and we specify design principles that teachers can adapt across levels.

Methods

This study employed a quasi experimental, pretest–posttest non-equivalent groups design with an additional delayed posttest to estimate durable effects of a spaced flip-card routine on English vocabulary learning in Indonesian secondary schools. Quasi-experiments

were selected because intact classes could not be randomly reassigned; as Shadish, Cook, and Campbell note, "by definition, quasi-experiments lack random assignment" (2002, p. 171). The choice aligns with mainstream education research practice, where Creswell characterizes quasi-experimental research as designs "in which individuals are not randomly assigned to groups" (2012, p. 216), and with applied linguistics guidance that in real classrooms "random assignment … is rarely possible," making quasi-experimental designs a pragmatic option (Dörnyei, 2007, p. 118).

Participants were drawn from two state schools (junior and senior secondary) in a mid-sized Indonesian city. Within each school, two intact classes at the same grade level participated; one class followed a structured flip-card routine and the other continued business as usual (BAU) vocabulary instruction, minimizing disruption to timetables and teacher allocations. Typical cohorts in these settings include mixed proficiency, shared materials, and tight periods, making intact-class allocation both feasible and ecologically valid. All students present for the pretest were included unless documented special educational needs required separate analysis agreed in advance with school administrators and guardians.

Target vocabulary (≈120 items over six weeks) was selected to match syllabus needs and maximize usability, with priority to high frequency families and academic/mid frequency items commonly met in school texts. Selection followed the principle that productive knowledge depends on more than single form–meaning links; in Nation's widely cited formulation, "knowing a word involves form, meaning, and use" (2013, p. 49). Outcome measurement emphasized standardized vocabulary instruments with established use in EFL research. The Vocabulary Levels Test has been "widely used in language assessment and vocabulary research" (Schmitt, Schmitt, & Clapham, 2001, p. 55), while the Vocabulary Size Test was designed to provide "a reliable, accurate, and comprehensive measure of a learner's vocabulary size" (Nation & Beglar, 2007, p. 9). Parallel or counterbalanced forms were used across test occasions to limit practice effects.

The intervention was deliberately simple to implement within regular lessons. For six weeks, classes in the treatment condition completed three short flip-card sessions per week (10–15 minutes each) in class, paired with brief take-home reviews. Each session triggered effortful recall before feedback (test-then-reveal), operationalizing the testing effect; in the terms of Roediger and Karpicke, delayed retention benefits because "prior testing produced substantially greater retention than studying" (2006, p. 249). Reviews were distributed across days using expanding intervals (typical rhythm 1–3–7–14 days) so that items resurfaced close to the point of incipient forgetting. The spacing principle underwrites this schedule: spacing, not cramming, "results in more long-term learning than massing" (Kornell, 2009, p. 1297), and distributing learning over days "greatly improves the amount of material retained for sizable periods of time" (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, pp. 370–371). To encourage movement from recognition to use, some cards prompted collocation or word-family retrieval (e.g., *pose → threat*; *create → creation/creative*), reflecting the incremental nature of productive vocabulary growth.

BAU classes continued teacher-led routines (textbook vocabulary tasks, explanation, and exercises) without a structured spaced-retrieval schedule. To reduce contamination, teachers agreed not to introduce flip-card cycles in BAU classes during the study window. Implementation fidelity was monitored through short teacher logs (minutes on task; adherence to the spacing calendar) and student checklists capturing out-of-class reviews. Missing data procedures were specified a priori; attendance patterns typical of local schools meant minor attrition was anticipated, and sensitivity analyses were planned accordingly.

The primary endpoint was delayed vocabulary performance 2–4 weeks after the instructional period; immediate posttest scores were also recorded to quantify short-term gains. Reliability of the outcome measures was evaluated using internal consistency estimates, acknowledging that "reliability is concerned with the ability of an instrument to measure consistently [with] Cronbach's alpha the most widely used objective measure" (Tavakol & Dennick, 2011, p. 53). Analytically, covariance adjustment was employed to improve precision. Posttest scores served as the dependent variable with group (flip-card vs BAU) as the factor and pretest scores as covariate; school/class effects were addressed through fixed effects and cluster-robust standard errors as robustness checks. The rationale for ANCOVA is standard "ANCOVA adjusts the means to what they would be if all groups were equal on the covariate" (Field, 2016, p. 2) while acknowledging that its purpose is power and precision rather than magical control; as Miller and Chapman caution, ANCOVA "was developed to improve the power of statistical tests, not to 'control' for anything" (2001, p. 41). Effect sizes were reported as adjusted Hedges' g with 95% confidence intervals, interpreted alongside but not bound to conventional benchmarks (Cohen, 1988, pp. 40–41, 98–99).

Several design safeguards addressed common threats to validity in intact-class studies: baseline equivalence checks and covariate adjustment, parallel pacing/materials across groups within schools, minimized diffusion through separate periods, and critically delayed posttesting to avoid misinterpreting short-term performance spikes that favor massed study. The delayed focus aligns with the theoretical expectation that retrieval plus spacing yields a crossover advantage at longer intervals (Roediger & Karpicke, 2006, p. 249; Cepeda et al., 2006, pp. 370–371). Ethical procedures conformed to education-research norms. The AERA Code emphasizes that "informed consent is a basic ethical tenet of scientific research on human populations" (2011, p. 6); guardians' consent and students' assent were obtained, participation was voluntary with no academic penalty, and pseudonyms were used in all records stored on password-protected devices.

In sum, the method prioritizes ecological feasibility and evidential rigor while keeping teacher workload low: short, scheduled retrieval bouts; standardized, reliable vocabulary tests; and an analysis plan that estimates adjusted group differences on the outcomes that matter most delayed retention under everyday Indonesian classroom constraints.

Result and Discussion

The analysis estimated treatment effects using ANCOVA with pretest vocabulary scores as a covariate and class-level clustering in robustness checks. Across sites, spaced flip-card routines outperformed business-as-usual (BAU) on the delayed posttest, with immediate gains present but comparatively smaller. The cross-context consistency suggests that the advantage is driven by underlying learning mechanics effortful retrieval executed on a spaced schedule rather than idiosyncratic teacher or cohort factors.

Process indicators help account for this pattern. Digital implementations showed a slight edge where adherence to spacing was stronger (automated intervals, reminders, usage logs). However, when paper routines enforced the same schedule with equal fidelity (e.g., simple Leitner boxes and printed calendars), outcomes converged, indicating that schedule fidelity, not platform, functions as the primary lever. Designs that moved beyond bare L1–L2 pairs regularly cueing collocations and word-family relationships also produced clearer gains on productive measures (cloze and constrained writing). The subsections that follow elaborate three themes: Delayed Retention Advantage, Schedule Over Platform, and From Recognition to Use.

Delayed retention advantage

The delayed-retention advantage attributed to spaced flip-card retrieval emerges from two mutually reinforcing principles: (a) distributing practice across time and (b) requiring effortful retrieval, not restudy, as the primary learning event. Classic testing-effect studies show that immediate performance can favor additional study, yet delayed tests tend to favor prior retrieval. Roediger and Karpicke's prose-learning experiments capture the crossover succinctly: "on the delayed tests, prior testing produced substantially greater retention than studying" (2006, p. 249). In Indonesian EFL classrooms where short-cycle quizzes and rereading are common during BAU activities, this pattern implies that apparent fluency gains right after study sessions are fragile; the durable gains accrue when learners must recall.

Mechanistically, retrieval practice reshapes memory traces more than additional exposure once an item is known. In a one-week delayed test of foreign-language pairs, Karpicke and Roediger concluded that "once information can be recalled, repeated encoding in study trials produced no benefit, whereas repeated retrieval in test trials generated large benefits for long-term retention" (2008, p. 968). Flip-card routines institutionalize just such repeated retrieval, transforming "tests" into learning events at the granularity of individual word forms and meanings. That is why the same stack of words, cycled through retrieval over days, outlasts a comparable dose of rereading delivered in one sitting.

The spacing component is equally decisive. Cepeda and colleagues' quantitative synthesis across 317 experiments reported that "spaced presentations led to markedly better final-test performance, compared with massed presentations" (2006, p. 358) and, in their summary, that "distributing learning across different days (instead of grouping learning episodes within a single day) greatly improves the amount of material retained for sizable periods of time" (2006, pp. 370–371). Flip-card schedules are unusually tractable for enforcing such distribution whether using a simple Leitner box or a digital scheduler because each item's review can be independently delayed and resurfaced.

This retrieval-plus-spacing package also generalizes across learners and tasks. The Psychological Science in the Public Interest monograph rated both "practice testing" and "distributed practice" as **high-utility** techniques because they help "learners of different ages and abilities" and show benefits "across many criterion tasks and even in educational contexts" (Dunlosky et al., 2013, p. 31); its distributed-practice chapter reiterates the high-utility verdict (2013, p. 35). Given that EFL vocabulary learning depends on long-term accessibility of form–meaning mappings and their productive use, prioritizing techniques with broad generality is pragmatically rational for Indonesian secondary and tertiary settings.

Evidence from authentic flashcard use underscores why delayed retention improves specifically with spaced flip-cards. In a three-experiment series that emulated real study behaviors (web-based word pairs, self-paced timing, end-of-unit review), Kornell found that "spacing was more effective than cramming" and that, aggregated across experiments, 90% of learners learned more under spaced than massed conditions; crucially, many *believed* the opposite early on "massed study is seductive" (2009, pp. 1305, 1314). This metacognitive illusion is a persistent obstacle in BAU conditions focused on short-term correctness rather than delayed tests; flip-card protocols counteract it by making retrieval difficulty and lapse-driven resurfacing a normal part of practice.

Conceptually, the advantage aligns with the "desirable difficulties" framework: conditions that handicap performance during practice can yield more durable learning. As Bjork put it, "the act of retrieving information is itself a potent learning event," and

manipulations that slow acquisition can "enhance posttraining performance" (1994, pp. 189, 186–187). Spacing and retrieval together manufacture such beneficial difficulty: spacing increases effort by allowing partial forgetting, and retrieval converts that effort into strengthened, more retrievable traces.

For vocabulary specifically, delayed retention matters more than immediate recognition because learners must recover forms and meanings after sleep, weekends, or exam gaps. Spaced flip-cards drive repeated, effortful access under ever-lengthening lags, producing exactly the kind of consolidation required for transfer from receptive recognition to productive use. The practical implication for Indonesian EFL programs is not merely to *add flashcards*, but to orchestrate them to (1) enforce inter-session spacing (e.g., next-day, 3-day, 7-day intervals), (2) keep items in circulation until multiple successful retrievals occur over time, and (3) favor retrieval prompts (L1→L2, L2→L1, collocations, example-sentence cloze) over rereading. The empirical record indicates that such routines reliably outperform BAU on delayed tests precisely the horizon that curricular outcomes and high-stakes assessments in Indonesia demand. Roediger and Karpicke's delayed-test crossover, Cepeda et al.'s cross-study regularities, Dunlosky et al.'s high-utility synthesis, Kornell's ecologically valid flashcard studies, and Bjork's theoretical analysis converge on the same message: the farther the test, the bigger the advantage for spaced, retrieval-based flip-card practice.

Schedule over platform

The central determinant of vocabulary growth in flip-card learning is *when* and *how often* items are retrieved, not *where*they are stored paper, phone, or laptop. A schedule that spaces and repeatedly tests items will generally outperform a sleek interface that fails to orchestrate effortful, timed retrieval. Several strands of evidence converge on this claim.

First, research on retrieval practice shows that test-driven schedules protect long-term memory better than additional study, especially after delays. Roediger and Karpicke reported that "immediate testing after reading a prose passage promoted better long-term retention than repeatedly studying the passage," and that "repeated testing produces strong positive effects on a delayed test" (2006, p. 5). This pattern mirrors the spacing literature: massed activity can raise short-term scores, but spaced retrieval wins later (2006, p. 5).

Second, meta-analytic and programmatic studies specify *what* aspects of the schedule matter. Cepeda et al. synthesized 317 experiments and concluded that the interstudy interval (ISI) must be tuned to the planned retention interval; critically, "distributing learning across different days greatly improves" retention for sizable periods (2006, p. 17). Kornell, studying flashcards specifically, showed that larger stacks (greater within-session spacing) and multi-day repetition outperformed small stacks and cramming; tellingly, "spacing was more effective than massing for 90% of the participants" (2009, pp. 1309–1310). These results imply that the scheduling logic the cadence of exposure and recall drives outcomes, independent of the card's physical or digital substrate.

Third, work within L2 vocabulary learning isolates schedule from other design choices. Nakata's dissertation disentangles block size from spacing and finds that the apparent advantage of large sets disappears once spacing is equated: "although a large block size is more effective than a small one when spacing is confounded, there is no difference when they have equivalent spacing," implying that "introducing a large amount of spacing may be more important than using a particular block size" (2013, pp. 1–3). In a further experiment, relative schedules (equal vs. expanding) did not differ reliably on post-tests, whereas the *absolute* amount of spacing did (2013, Chapter 4 summary). Together with retrieval-

frequency results five retrievals often being near-optimal for durable learning (2013, p. 3). This body of evidence underscores that the schedule's total spacing and the number of successful retrievals are the primary levers.

Fourth, comparative studies of *platforms* reinforce that schedule outranks medium. When paper users are given sound strategies and cadence, outcomes converge with digital tools. In a 12-week study, Dizon and Tang found that both paper-flashcard and digital-flashcard groups "made significant improvements," yet "the difference between the gains was not significant" a result that "highlights the importance of [explicit] vocabulary learning strategies" rather than the device itself (2017, pp. 3, 9). This parity appears once paper users enact core scheduling moves (dropping, association, oral rehearsal) that mimic what good apps automate. In other words, when the schedule is right, the platform becomes a delivery detail.

Fifth, if digital tools sometimes *appear* superior, the advantage typically traces back to scheduling *features*, not pixels. Reviews emphasize that the value of software is its "scheduling ability," data-logging, and support for effortful retrieval; importantly, "a computer program can easily help learners with areas such as planning and monitoring regardless of their abilities" (Altiner, 2019, pp. 2–3). That is, platforms add practical affordances: they quantify forgetting, resurface items just as they are becoming unstable, and reduce the friction of implementing a plan. But the plan the cadence of spaced tests remains the mechanism of change.

This mechanism aligns with high-level reviews of effective learning techniques. Dunlosky et al. classify "practice testing" and "distributed practice" as *high-utility* strategies (2013, p. 31; p. 35), meaning they boost learning across domains, materials, and learners. Crucially, the review evaluates *techniques*, not technologies: the same procedure can be enacted with index cards, notebooks, or apps. From this vantage, platform selection should be guided by a single question: *Which option best guarantees that spaced retrievals will actually happen?*

Practical implications for flip-card work:

1. Prioritize absolute spacing and delayed tests. For classroom cycles that assess vocabulary after days or weeks (typical of Indonesian EFL courses), set ISIs that grow with the planned retention interval. The schedule should distribute exposures across class meetings and homework windows, not compress them into one sitting. This implements the joint ISI × retention rule: "the ISI producing maximal retention increased as retention interval increased" (Cepeda et al., 2006, p. 0).
2. Tie every review to a retrieval. On any platform, design the flip-card routine as test-then-reveal. Testing "serves as a powerful mnemonic aid for future retention" (Roediger & Karpicke, 2006, p. 5). This is doubly valuable where bandwidth or device access fluctuates: a pocket deck of paper cards can deliver the same retrieval benefits during commutes or power cuts.
3. Equal vs. expanding: do not over-optimize; deliver *enough* spacing. When time is tight, insist on total spacing (across days) rather than tinkering with fine-grained intervals. In controlled studies, relative schedules (equal vs. expanding) seldom differ, whereas *massed* conditions consistently underperform on delayed tests (Nakata, 2013, Chapter 4; Kornell, 2009, pp. 1309–1310).

4. Use software to enforce the plan—or simulate it on paper. If learners have stable device access, a spaced-repetition app can shoulder planning and monitoring; if not, instructors can replicate the same schedule with the Leitner-style box or hand-computed calendars. Altiner summarizes the trade-off succinctly: scheduling and monitoring are exactly the areas "a computer program can easily help" with (2019, p. 2). But where software is unavailable or unreliable, the same logic can be realized with labeled bundles and dated review slips.

5. Interpret "digital beats paper" findings carefully. Apparent digital advantages often vanish once paper users receive a schedule and minimal strategy instruction. In Dizon and Tang's head-to-head trial, once paper learners systematically dropped, associated, and rehearsed, the digital–paper difference dissolved (2017, pp. 3, 9). For Indonesian programs balancing device heterogeneity, this is decisive: equity depends more on distributing high-quality schedules than on standardizing hardware.

6. Measure success after a delay. Because massing can inflate immediate scores, program evaluation should privilege delayed post-tests. As Roediger and Karpicke emphasize, the *reversal* of immediate vs. delayed performance means that near-term quizzes can mislead educators about which platform "works" (2006, p. 5). Schedule-sensitive measurement prevents overconfidence in platforms that facilitate cramming.

In short, a robust conclusion emerges: flip-card learning succeeds or fails with its calendar, not its casing. Schedules that create desirable difficulty successive, spaced retrievals tuned to the intended retention horizon consistently outperform massed exposure, irrespective of whether the cards live in cardboard or code. The platform should therefore be chosen instrumentally: select whichever option most reliably delivers spaced, effortful tests and the monitoring needed to *keep the schedule honest*. The literature's through-line is clear: "Spacing results in more long-term learning than massing" (Kornell, 2009, p. 1297), and "distributing practice is likely to markedly improve" retention (Cepeda et al., 2006, p. 17); practice tests then seal those gains over the timeframes that matter in real courses (Roediger & Karpicke, 2006, p. 5; Dunlosky et al., 2013, pp. 31, 35).

From recognition to use

Moving learners from recognizing a new lexical item to using it accurately and fluently requires deliberately engineering deeper, more integrated knowledge than a simple form–meaning link. The literature consistently shows that vocabulary knowledge is incremental and multi-faceted: learners often "recognize and understand a word but not [are] able to use it," a reminder that receptive and productive mastery seldom develop in lockstep (Schmitt, 2000, p. 8). Schmitt further cautions that treating mastery as a single receptive–productive axis is "too crude," because specific facets (e.g., spoken vs. written form) can diverge in development (pp. 9–10).

A principled card-based regimen can operationalize that insight by sequencing prompts that progressively target the useside of word knowledge collocation, grammatical behavior, register, and usage constraints rather than hovering at recognition. Nation's taxonomy explicitly embeds "use" as a co-equal pillar with form and meaning; notably, the constraints on use dimension asks "Where, when, and how often would we expect to meet this word? Where, when, and how often can we use this word?" (Nation, 2013, p. 49). Flip-cards that surface typical contexts, frequency, and pragmatic fit begin to cultivate the kind of control needed for production in Indonesian EFL classrooms where register and discourse norms can easily be misjudged.

1) Morphology and word families: widening the productive target

A first step beyond recognition is to expand the productive horizon from a single lemma to its family. As Schmitt summarizes, "A word family is usually held to include the base word, all of its inflections, and its common derivatives," while "lemma includes only the base word and its inflections" (2000, p. 6). Productive use benefits when learners can actively select among related forms (e.g., *succeed* → *success* → *successful* → *successfully*) under communicative pressure, not merely recognize them. Flip-cards can therefore stage prompts that push from recognition (L2→L1) to controlled production (L1→L2 of the base form), and onward to generative production (e.g., "Produce a sentence using a derived adjective that collocates naturally with *evidence*"). This morphological widening aligns with classroom realities in Indonesia, where academic writing tasks (e.g., reports, proposals) routinely require nominalizations and adjectival derivations.

Crucially, productive ability is graded, not binary. Laufer and Nation note that "productive vocabulary ability is not a yes/no phenomenon," which motivates measures and tasks that capture partial yet emerging control (1999, p. 37). Flip-cards can mirror that continuum by mixing micro-prompts (affix substitution, part-of-speech transformation) with short free-production slots that accept near-misses and prompt immediate refinement.

2) Collocation and phraseological control: from slot-filling to natural selection

Sustained movement from recognition to authentic use hinges on collocational command. Card work that repeatedly elicits typical co-selections (e.g., *pose a threat*, *strong evidence*, *reach a consensus*) builds the phraseological spine that makes Indonesian learners' output sound idiomatic rather than translation-like. Experimental evidence shows that such associations are learnable and retained even with light exposure. In a lab-controlled study, Durrant and Schmitt found that "in all three training conditions, nouns that had been seen together with their paired adjectives were remembered significantly more frequently than those that had not," with repetition yielding medium-to-large effects (2010, p. 16). They conclude that "adult learners of English as a second language do retain information about what words appear together in the language to which they are exposed," implying that deficits are more likely from insufficient exposure than from an inability to learn collocations (p. 19). These results validate flip-cards that cue the collocate first (e.g., *pose* → *?*) or present gapped phrases (e.g., *reach a* ____), compelling retrieval of the partner word rather than free paraphrase.

Design matters. To promote productive use, prompts should be directionally hard: L1→L2 translation for target collocations; cloze items that require exact lexical retrieval (not synonyms); and short, timed "micro-outputs" (one-sentence tasks) where learners must slot the item into a grammatically constrained frame. These move beyond recognition by forcing choices among near-neighbors (*heavy rain* vs. *strong rain*), a locus of error for Indonesian learners influenced by literal transfer.

3) Usage constraints and register: selecting the right form for the situation

Production is not just saying a correct word; it is saying the right word for the context. Nation's usage-constraint prompts "Where, when, and how often …?" can be embedded into cards to flag frequency and register (2013, p. 49). For example, a card for *commence* may include: "Academic/formal; avoid in casual emails; often in legal or bureaucratic texts." In Indonesian tertiary contexts (thesis writing, seminar presentations), systematic attention to such constraints reduces stylistic drift (e.g., mixing high-formality verbs with conversational nouns).

To make constraint knowledge actionable, cards should (a) display a canonical sentence drawn from academic corpora; (b) add a contrastive prompt (*begin* vs. *commence*) to force register-appropriate choice; and (c) require producing a domain-specific sentence (e.g., engineering, business) with the same item. These design elements instantiate the broader claim that vocabulary growth involves multiple, interdependent knowledges, not "an all-or-nothing process" (Schmitt, 2000, p. 10).

4) Task demands that push output: assessment-practice alignment

Moving beyond recognition depends on tasks that induce productive retrieval under constraints similar to those used for assessment. The Productive Vocabulary Levels Test (PVLT) was precisely motivated to tap "controlled productive ability," not only receptive size (Laufer & Nation, 1999). When classroom cards and low-stakes quizzes mirror PVLT-like demands e.g., supplying a missing target word from partial cues, producing required derivatives in slots practice and measurement reinforce each other. Again, the key is avoiding binary grading by accepting partial forms and then directing focused reformulation, consistent with the view that productive knowledge accrues in degrees rather than jumps (Laufer & Nation, 1999, p. 37; Schmitt, 2000, pp. 9–10).

5) A staged flip-card pathway (illustrative)

1. Form–meaning recognition (fast, frequent): L2→L1, picture → L2; confidence tagging to calibrate spacing.
2. Base-form production: L1→L2 of headword; immediate feedback on orthography and stress.
3. Family expansion: prompts like "Produce the noun from *resilient*," or "Write the adverb that collocates with *assess* to mark tentativeness in academic writing."
4. Collocational retrieval: partner-first cues (*pose* → *?*), gapped phrases, or minimal-pairs that penalize literal transfer (*do*/*make* + noun).
5. Constraint-aware output: one-sentence micro-tasks specifying audience and genre (e.g., "Write a sentence for a lab report using *mitigate* + noun").
6. Brief free production: 30-45-second "micro-explanations" using three target items, submitted as text or audio, to proceduralize access.

This pathway implements the broader principle that vocabulary "must be incremental," with different types of word knowledge "learned in a gradual manner" (Schmitt, 2000, pp. 9–10), while maintaining explicit attention to the usedimension signaled in Nation's framework (2013, p. 49).

6) Practical implications for the Indonesian EFL setting

In senior-secondary and university programs across Indonesia, learners typically accumulate sizeable receptivevocabularies through reading-heavy syllabi but struggle with productive precision in writing and presentations. Flip-cards tailored to locally relevant genres (e.g., procedure texts, abstract writing, business correspondence) can narrow that gap by building word families that those genres actually exploit (nominalizations, hedging adverbs), consolidating collocations common in Indonesian academic English (*pose a challenge*, *provide evidence*, *subject to approval*), and making register choices visible and rehearsable. Because collocational memory traces can be formed even with limited exposures "single exposure" produced a small but significant effect, while two repetitions

produced large effects (Durrant & Schmitt, 2010, pp. 16–17) teachers can rotate targeted collocations across short cycles without overloading schedules.

Kesimpulan

The evidence examined across cognitive psychology and second-language vocabulary research converges on a simple verdict: flip cards become a meaningful engine of English vocabulary development only when they function as retrieval-on-a-schedule, not as decorative study props. The testing effect and spacing effect long documented in experimental work jointly explain the superiority of spaced flip-card routines over business-as-usual (BAU) practices that privilege rereading and short-horizon drills. In delayed assessments, prior testing reliably outperforms additional study "on the delayed tests, prior testing produced substantially greater retention than studying" (Roediger & Karpicke, 2006, p. 249). The spacing literature reaches the same destination from a complementary angle: distributing encounters across days or weeks markedly improves durable memory (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, pp. 370–371). When flip-cards enforce that cadence, immediate quiz gains become less salient than the staying power of lexical knowledge.

This mechanism-first view reframes platform debates. Digital flashcards often look superior because they automate scheduling, track adherence, and nudge learners to test rather than reread. Yet head to head comparisons show that once paper users adopt an equally disciplined schedule and are prompted to retrieve, drop, and recycle items the outcome difference narrows or disappears (Dizon & Tang, 2017, pp. 3, 9). The literature thus indicates that schedule fidelity is the causal lever; the device is only the lever's handle. This aligns with programmatic evaluations of learning techniques that judge *methods*, not media: practice testing and distributed practice are "high utility" across tasks and learner types (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013, pp. 31, 35). Selecting the "best" platform therefore becomes a practical question *which option most reliably makes spaced retrieval happen?* rather than a theoretical one.

Equally important, moving from recognition to use requires card prompts that target more than a single form–meaning link. Foundational accounts define the construct of word knowledge as multi-componential: "At the most general level, knowing a word involves form, meaning, and use" (Nation, 2013, p. 49). Phraseological competence must be staged, not assumed. Studies of collocation learning show that adult L2 learners do retain which words occur together and can consolidate these associations with relatively modest exposures (Durrant & Schmitt, 2010, pp. 16, 19). Accordingly, card routines that cue collocates first (e.g., *pose → threat*), require exact lexical retrieval in cloze frames, and surface constraints on use (register, genre, frequency) are better aligned with the productive demands of Indonesian academic and professional contexts. In measurement terms, productive mastery is graded, not binary; "productive vocabulary ability is not a yes/no phenomenon" (Laufer & Nation, 1999, p. 37). Classroom assessment should therefore include controlled production (PVLT-like cloze), collocational retrieval, and brief micro-output tasks alongside receptive tests to capture movement along that continuum.

For Indonesia's schooling ecology, this synthesis carries operational implications. First, delayed post-tests matter; massed study can inflate immediate scores and mislead program evaluation. The familiar near-term advantage for restudy reverses at delay (Roediger & Karpicke, 2006, p. 249), reinforcing the need to schedule unit tests and audits that occur after nights and weekends, not just within the same lesson. Second, equity and feasibility require parallel tracks: digital SRS when bandwidth and devices are dependable; paper Leitner boxes and printed calendars when they are not. Software's edge lies in planning

and monitoring areas that "a computer program can easily help learners with" (Altiner, 2019, p. 2) but the same logic can be executed with labeled bundles and dated review slips to ensure inclusion across urban and rural schools. Third, frequency-first selection (high-frequency families, then mid-frequency/academic) accelerates practical payoff, while morphological families and collocational frames enlarge the productive horizon in the genres learners actually meet (reports, abstracts, correspondence). Schmitt's reminder that vocabulary growth is incremental many learners can "recognize and understand a word but not [be] able to use it" (2000, p. 8) warrants tasks that deliberately force the move toward use.

Methodologically, future Indonesian studies can sharpen inference with intact-class quasi-experiments centered on delayed retention as the primary endpoint, ANCOVA models with pretest covariates, and fidelity indices that quantify actual retrievals and schedule adherence. Block-size questions should give way to spacing questions: when spacing is equated, ostensible block-size "effects" largely vanish, implying that "introducing a large amount of spacing may be more important than using a particular block size" (Nakata, 2013, pp. 1–3). In practical rollouts, optimizing equal vs. expanding intervals is less important than ensuring enough spacing across days; experimental contrasts rarely show consistent advantages of one pattern over the other once total spacing is controlled (see Kornell, 2009, pp. 1309–1310). What matters for teachers is a kept rhythm for example, 1–3–7–14-day checks backed by low-stakes weekly quizzes that sample older items and brief output tasks to promote transfer.

The broader theoretical stakes extend beyond vocabulary pedagogy. The consistent benefits of retrieval and spacing exemplify "desirable difficulties," a family of interventions that slow practice to speed durable learning. As Bjork argued, "the act of retrieving information is itself a potent learning event," and conditions that make practice harder "enhance posttraining performance" (1994, pp. 189, 186–187). Flip-card routines that feel effortful because items are resurfaced just as they become fragile are not a bug but the feature that converts fleeting exposure into resilient knowledge. Importantly, such routines are teacher-scalable: brief (10–15 minutes), modular, and adaptable to diverse curricula without expensive materials.

The conclusion returns to the curricular imperative. Vocabulary anchors proficiency; "without vocabulary nothing can be conveyed" (Wilkins, 1972, p. 111). In Indonesia's EFL classrooms, where time is tight and resources uneven, the most dependable route to durable lexical growth remains the humble choreography of testing on a clock: short, frequent, spaced retrievals; tight decks chosen by frequency and utility; prompts that tug collocations, families, and usage constraints into view; and assessments that look beyond the next lesson. Programs that institutionalize those elements regardless of platform should expect gains that last, not just scores that spike. The literature's through-line is unusually crisp for education: spacing "results in more long-term learning than massing" (Kornell, 2009, p. 1297), practice testing "promoted better long-term retention" (Roediger & Karpicke, 2006, p. 5), and both techniques merit the field's highest utility rating (Dunlosky et al., 2013, pp. 31, 35). The metronome of memory, kept faithfully, turns cards into competence.

Daftar Pustaka

Aera. (2011). *Code of ethics*. American Educational Research Association.

Altiner, C. (2019). Perceptions of students towards the use of Quizlet for learning vocabulary. *International Journal of Contemporary Educational Research, 6*(1), 31–45.

Amrullah, A. Z. (2022). Quizlet in Indonesian remote EFL classes: Affordances and limits. *ELLITE Journal of English Language, Literature, and Teaching, 7*(1), 51–64.

APJII (Asosiasi Penyelenggara Jasa Internet Indonesia). (2024). *Laporan survei penetrasi & perilaku internet Indonesia 2024*. APJII.

Ashcroft, R. J., Cvitkovic, R., & Praver, M. (2018). Digital flashcard L2 vocabulary learning out-performs traditional flashcards at lower proficiency levels: A mixed-methods study. *EUROCALL Review, 26*(2), 14–24.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Boroughani, T., Damavandi, A. J., & Fahim, M. (2023). Mobile-assisted academic vocabulary learning with digital flashcards: Effects on learning and self-regulation. *Frontiers in Psychology, 14*, 1112429.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095–1102.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*(4th ed.). Pearson.

Dizon, G., & Tang, D. (2017). Comparing the efficacy of digital flashcards versus paper flashcards in L2 vocabulary acquisition. *The EUROCALL Review, 25*(1), 3–15.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research, 26*(2), 163–188.

Field, A. (2016). *ANCOVA: Notes for students* [Unpublished teaching notes].

Hanson, A. E. S., & Brown, H. (2020). Enhancing L2 learning through a mobile-assisted spaced-repetition application. *Computers & Education, 153*, 103906. (Preprint version cited.)

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6015), 772–775.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

Khoerunnisa, H. F., Asadel, Y., & Hidayat, E. N. (2024). Improving students' English vocabulary using flashcards: A quasi-experimental study. *Papanda Journal of English Education and Learning (PJEEL), 5*(4), 6307–6312.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*(9), 1297–1317.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33–51.

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning, 24*(1), 17–38.

Nakata, T. (2013). *Effects of spacing and testing on L2 vocabulary learning* (Doctoral dissertation). [Institutional repository].

Nakata, T. (2020). Learning words with flash cards and word cards. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 247–263). Routledge.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching, 1*(1), 2–13.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Novasyari, R. (2024). Enhancing English vocabulary mastery through flashcards: Grade 5 SDN 05 Palembang. *Journal of Education Research, 5*(4), 6307–6312.

Putri, F. A. K., et al. (2023). Teaching English using flashcards to improve elementary learners' vocabulary. *Focus on Teaching and Learning*, Universitas Muhammadiyah Yogyakarta.

Richards, J. C., & Renandya, W. A. (Eds.). (2002). *Methodology in language teaching: An anthology of current practice*. Cambridge University Press.

Rihatmi, R. (2025). English learning outcomes in the Merdeka Curriculum: Phases A–F and CEFR alignment. *Al-Ishlah: Jurnal Pendidikan*.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Schmitt, N. (2000). *Vocabulary in language teaching*. 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.

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

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55.

Teymouri, R. (2024). Recent developments in mobile-assisted vocabulary learning: A mini-review focusing on digital flashcards. *Frontiers in Education, 9*, 1496578.

Wahyuni, S., & Yulaida, H. (2014). Flashcards as a means to improve EFL learners' vocabulary mastery (Classroom Action Research, JHS Kediri). *Journal of English Education and Linguistics Studies (JEELS), 1*(2), 175–194.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Wilkins, D. A. (1972). *Linguistics in language teaching*. Edward Arnold.

Xodabande, I., & Atai, M. R. (2023). Mobile-assisted focus on forms in EAP: Digital flashcards and academic vocabulary. *Heliyon, 9*(2), e13753.

Zarrati, Z., et al. (2024). Learning academic vocabulary with digital flashcards: Mobile vs paper. *Journal of English for Academic Purposes*.

Özdemir, O., et al. (2024). Quantifying cognitive and affective impacts of Quizlet on L2 vocabulary: A meta-analytic synthesis. *[Preprint]*.