AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2023 American Psychological Association
ISSN: 0278-7393

2024, Vol. 50, No. 3, 435–457
https://doi.org/10.1037/xlm0001233

Does Testing Potentiate New Learning Because It Enables Learners to Use Better Strategies?

Dahwi Ahn and Jason C. K. Chan
Department of Psychology, Iowa State University

OPEN DATA PREREGISTERED TC

Testing can potentiate new learning, which is often called the *forward testing effect*. One potential explanation for this benefit is that testing might enable participants to use more effective learning strategies subsequently. We investigated this possibility by asking participants to report their encoding strategies in a multilist foreign language learning paradigm with four preregistered experiments and one non-preregistered pilot experiment. In Experiments 1–3, participants learned three lists of Chinese–English pairs; one group took a test after every list (i.e., test condition) and the other group took a test only for the criterial List 3 (i.e., restudy condition). In addition, participants completed a transfer test and a study strategy survey. Although we found a forward testing effect in all experiments, participants in the test and restudy conditions did not report differences in strategies. In Experiments 4 and 5, we used a within-subject design so that we could correlate changes in strategy use with the magnitude of the forward testing effect on an individual level. Interestingly, individual differences in strategy change were moderately associated with the magnitude of the forward testing effect, but even here, strategy change did not mediate the effect of testing on performance. Overall, our data showed that, at least for foreign language learning of Chinese characters, interim testing did not enhance new learning by altering participants' subsequent encoding strategies. Moreover, our data showed that interim testing did not promote the transfer of Chinese language learning to novel characters.

Keywords: retrieval practice, forward testing effect, strategy change, foreign language learning, transfer of learning

Supplemental materials: https://doi.org/10.1037/xlm0001233.supp

Taking a test on previously learned material enhances subsequent new learning. For example, if an instructor splits their class time in half and introduces a quiz after the first half, this practice will help students better retain the tested first half and also promote learning during the second half of the class. The enhanced learning of tested items is often referred to as the *testing effect*, whereas the enhanced learning of new material is called the *forward testing effect*. To clearly distinguish the two findings, we refer to the former as the *retrieval practice effect*. Although both phenomena have received considerable attention (for reviews, see Chan, Meissner, & Davis, 2018; Roediger & Karpicke, 2006; Rowland, 2014; Yang et al., 2018), teachers and students are often unaware of the advantages of testing in learning (Halamish, 2018; McCabe, 2011, 2018), which could lead to a reluctance to implement testing in the classroom. Thus, understanding how testing improves learning is important because it will allow us to better explain to instructors why introducing quizzes during a class is beneficial.

The goal of the current study is to investigate the *strategy-change account* of the forward testing effect, which suggests that testing facilitates future learning because it enables learners to use better strategies for new learning. Specifically, testing has been known to enhance the use of relational processing strategy (Chan et al., 2020; Zaromb & Roediger, 2010), which might promote conceptual learning in which discovering relations among different concepts is critical. An additional, related goal of this study was to examine whether testing can facilitate the transfer of conceptual learning using Chinese characters as material. We review the relevant literature and explain the motivation of our study below.

Testing and Learning Strategies

Several accounts have been proposed to address the question of "how testing potentiates new learning." In particular, a recent meta-analysis (Chan, Meissner, & Davis, 2018) identified four classes of theories—resource theories, metacognitive theories, context theories, and integration theories, which are not mutually exclusive and might explain different aspects of the forward testing effect. In this paper, we will focus on the metacognitive account, which claims that testing alters how students approach subsequent learning activities.

Chan, Manley, et al. (2018) proposed that testing promotes new learning because it helps learners optimize their encoding or

This article was published Online First November 13, 2023.
Dahwi Ahn https://orcid.org/0000-0002-4300-6520
This work is supported by the National Science Foundation Science of Learning and Augmented Intelligence Grant 2017333.
All data files are available at https://osf.io/uh8wr/ (Ahn & Chan, 2023).
The experiment was preregistered at https://osf.io/uh8wr/ (Ahn & Chan, 2023).

Correspondence concerning this article should be addressed to Dahwi Ahn, Department of Psychology, Iowa State University, 1347 Lagomarcino Hall, Ames, IA 50011, United States. Email: dahn@iastate.edu

retrieval strategies, based on the finding that tested participants show greater clustering during target list recall compared to nontested participants. In their study, participants studied words from five categories (i.e., animals, weather, fruits, body parts, and building parts), which were spread evenly across four lists. Participants in the test condition were tested after every list, whereas those in the control condition were tested only after List 4. When recalling List 4 words, tested participants were more likely to cluster the words from the same category than control participants. The researchers provided two explanations for this finding.

Retrieval Strategy Change

First, testing might enable participants to use better *retrieval strategies*. It has been known that greater semantic clustering during recall is associated with better recall performance (Mandler, 1967; Rawson & Zamary, 2019; Shuell, 1969). Higher clustering shown by the tested participants indicates enhanced organizational processes during recall, which might be evidence that participants used recalled items as extra retrieval cues. For example, after recalling some fruits such as *apple* or *banana*, participants might realize the relations between these words and then use the fruit category as a cue to recall other studied words such as *melon*. Specifically, the prior recall experiences during the interim tests might help participants realize the advantage of this method, whereas the control participants had no way to practice because the final test was their first and only testing experience. Other studies have shown compelling evidence for the idea that interim testing leads to subsequent retrieval strategy updating (Chan et al., 2020; Dang et al., 2021; Jing et al., 2016).

Encoding Strategy Change

Second, Chan, Manley, et al. (2018) suggested that enhanced semantic clustering may also reflect a change in participants' *encoding strategy*. The idea is that testing might help participants realize the categorical nature of the learning material, which encourages them to attend to the relations among the study words during subsequent encoding trials. Note that this encoding-based explanation is *inferred* from recall data (i.e., high clustering during recall might indicate a change to encoding). Indeed, most studies that have investigated the strategy-change account have focused on how interim testing changes later retrieval processes during the final test, and little research has directly examined how testing affects the *encoding* of new information.

One way in which testing promotes subsequent encoding is by preventing learners from terminating their study prematurely, which suggests a quantitative (rather than qualitative) change to encoding strategies. Specifically, Yang et al. (2017) investigated how testing affects future study time allocation by having participants decide how long they want to study each stimulus across multiple lists. Participants either took an interim test after each study list or restudied the same list. In the absence of interim testing, participants decreased their study time across lists. However, in the presence of testing, participants maintained (Experiment 1) or increased their study time (Experiment 2) across lists. Similarly, Davis and Chan (2022) reported that participants spent considerably longer time studying STEM text material when they received interim tests relative to interim restudy. Further, they showed that self-regulated study time was positively correlated with test performance.

Other studies, however, have shown that interim testing had little to no influence on study time regulation. For instance, Ha and Lee (2019) reported no difference in study time between tested and control participants, but the former group still performed better on the final test. Other studies showed that testing boosted learning in an experimenter-paced procedure, in which participants in both the test and control conditions spent an equal amount of time studying (Chan et al., 2020; Szpunar et al., 2008; Wissman et al., 2011; Yang et al., 2019). One interpretation of these findings is that interim testing may lead learners to use their study time more efficiently (e.g., by adopting better learning strategies).

Yang et al. (2022) provided further preliminary support for the encoding strategy-change explanation. Critically, participants learned unrelated words in this study. Unlike categorized words, unrelated words do not naturally lend themselves to meaning-based retrieval strategies, so temporal clustering at retrieval (e.g., clustering based on input position) provides a window into encoding processes. The researchers found that interim testing increased the likelihood that participants would cluster recall based on input positions, and the level of temporal clustering was positively associated with recall performance. Although this finding shows that interim testing fostered the encoding of temporal information, it does not address *how* this advantage is realized. It is possible that tested participants switched from simply reading the list to using an imagery-based encoding technique similar to the method of loci, which should promote the encoding of temporal orders and would signal a qualitative change in encoding strategy. Alternatively, testing might simply encourage more rehearsal during the encoding phase, which should also increase temporal clustering but does not constitute a qualitative shift in strategy.

Measure of Encoding Strategies

To date, few researchers have tackled the question of how or if encoding strategies of new information are changed qualitatively and/or quantitatively after testing (cf., Cho et al., 2017; Finley & Benjamin, 2012[1]), because most studies have relied on the indirect measure of strategies such as recall clustering to make inferences about encoding processes (Chan et al., 2020; Dang et al., 2021; Yang et al., 2022). Although this approach is useful for hypothesis generation, it is insufficient for confirmation. One way to investigate encoding strategies is to ask participants to report them explicitly, which has been used in previous studies and regarded as a valid measure of encoding strategies. For example, Finley and Benjamin (2012) created a strategy questionnaire based on participants' responses to examine how different test expectancies led to adaptive and qualitative changes in encoding strategy. In another study that investigated the relationship between metacognitive control and aging (Hertzog & Dunlosky, 2004), the researchers had participants select one of several strategies that they used most often or report their idiosyncratic strategies.

To our knowledge, the study by Cho and Powers (2019) is the only one that has examined the effects of testing (vs. restudying)

[1] These studies were designed to examine the retrieval practice effect, not the forward testing effect.

on participants' reports of study strategies. Although the goal of their study was to examine whether testing improves conceptual learning in the context of the retrieval practice effect (rather than the *forward* testing effect), their finding is consistent with the idea that testing might influence learners' encoding strategy. Specifically, participants studied Chinese–English word pairs (e.g., 江—river). The Chinese characters consisted of two or more sub-characters called *radicals*, which can provide information about the meaning of the whole character. For example, the radical 氵 represents "water" and can be found in characters such as 江 (river) and 海 (sea). Participants were not told about the existence of radicals, but if they noticed that some characters shared similar symbols and that these symbols were associated with a common meaning, they might be able to guess the meaning of characters based on the radicals.

During the initial study phase, participants learned seven characters from each of the six radicals. Then, participants in the test condition attempted to recall the English meaning of the studied Chinese characters, whereas those in the restudy condition were shown the complete pairs again. Afterward, the entire procedure was repeated and then participants completed an encoding strategy questionnaire, in which they reported to what extent they used the seven encoding strategies adapted from Finley and Benjamin (2012)—Target focus, Cue focus, Inter-item association, Mental imagery, Rote rehearsal, Inter-item narrative, and Intra-item narrative.

Cho and Powers (2019) reported significant differences in two of the strategy items between the test and restudy conditions. Specifically, the tested group reported higher usage of inter-item association than the restudy group, which indicates that testing encouraged participants to associate different characters with one another (presumably by using the information provided by radicals) during the encoding phase. For example, testing might have facilitated participants' ability to notice the shared symbol across characters and extract a common meaning (e.g., after studying 汁—juice, 江—river, and 池—pool, one might infer that 氵 is associated with water). Indeed, the tested group's performance exceeded the restudy group in a *transfer test* in which they had to guess the meaning of novel Chinese characters featuring the studied radicals. Furthermore, the tested group reported less usage of rote rehearsal, a shallow and ineffective strategy, than the restudy group. Together, these results suggest that testing might cause learners to shift from shallower strategies such as rote rehearsal to deeper ones such as inter-item association. However, we believe that these results should be interpreted with caution for the following reasons. First, Cho and Powers (2019) conducted seven *t*-tests, one for each strategy between the test and restudy conditions, thus greatly increasing the risk of Type-1 error. Second, Cho and Powers (2019) did not examine whether different strategy use actually affected test performance. Although it is reasonable to assume that rote rehearsal was ineffective and that inter-item association was effective at promoting the learning of Chinese characters, no correlation or mediation analysis was conducted to ascertain these possibilities.

Current Study

To summarize, despite compelling evidence that testing potentiates new learning by altering participants' *retrieval* strategies (Chan et al., 2020; Dang et al., 2021), much less is known about how interim testing might affect *encoding* strategies. Note that we believe that the encoding and retrieval strategy-change explanations are complementary rather than exclusive. Testing affects the regulation of study time in subsequent learning (Davis & Chan, 2022; Yang et al., 2017), but this finding is equivocal (Ha & Lee, 2019). It has also been reported that testing influences the encoding of temporal order (Yang et al., 2022), which might indicate an encoding strategy change, but this claim was based on an inference during participants' retrieval performance. The study that most closely examined the possibility of encoding strategy change is Cho and Powers (2019), but the study's goal was to examine whether testing can enhance conceptual learning, and the aforementioned statistical concerns made it difficult to draw strong conclusions. Table 1 provides a summary of previous findings about how testing affects the regulation of different study strategies.

In the current study, we adapted Cho and Powers' (2019) design to study the forward testing effect and examined how interim testing might affect encoding strategies of new information. Below, we highlight some important changes relative to Cho and Powers' (2019) experiments. First, participants studied several different lists instead of a single list repeatedly because our research goal was to see how testing promotes the learning of new items rather than relearning of the tested items. Second, instead of conducting a between-subjects *t*-test for each strategy, we conducted a factor analysis on strategies to extract strategy factors and then compared strategy factor scores across conditions to reduce the likelihood of Type-1 errors. To better align the strategy questions with the purpose of our study, we created a new strategy survey containing 12 items based on pilot testing and existing studies (Cho & Powers, 2019; Finley & Benjamin, 2012; Hertzog & Dunlosky, 2004). Third, we asked some participants to report their learning strategies after each list rather than only after the target list. The objective was to examine whether the use of strategies evolved as learners were tested on more lists. If testing indeed affects encoding strategies, one might expect that a difference in strategies between the test and restudy conditions to emerge as participants go through multiple study lists (but not on the first list). Fourth, we sought to examine the cognitive correlates of the forward testing effect by using a within-subjects design in Experiments 4 and 5, which enabled us to measure the magnitude of the forward testing effect per individual and correlate it with strategy use. Lastly, to examine the strategy-change account, we conducted a mediation analysis to test whether strategy use mediates the relationship between interim activity (i.e., test vs. restudy) and target list correct recall performance.

Overview of Experiments

In our Experiments 1–3, interim activity was manipulated between-subjects, and participants studied three lists of Chinese–English pairs. There were two types of interim activities for Lists 1 and 2, with participants in the test condition taking a test after studying each list, whereas those in the restudy condition restudied the same list. For the target List 3, all participants took a test.

The forward testing effect has been observed using foreign languages such as Swahili–English (Cho et al., 2017) or Euskara–English pairs (Yang et al., 2017), but this study tested Chinese characters for the first time in a foreign language learning setting. Some studies used Chinese as learning material (e.g., Dang et al., 2021;

Table 1
Summary of Previous Findings About Testing and Strategy Change

Authors	Testing effect	Major finding	Potential mechanisms
Chan, Manley, et al. (2018)	Forward	Testing increased organization in recall	Retrieval and/or encoding strategy change
Cho and Powers (2019)	Backward	Testing influenced encoding strategies when restudy opportunity was provided	Encoding strategy change
Davis and Chan (2022)	Forward	Testing increased subsequent study time	Study time regulation
Ha and Lee (2019)	Forward	Testing did not increase subsequent study time	More efficient use of study time
Yang et al. (2017)	Forward	Testing increased subsequent study time	Study time regulation
Yang et al. (2022)	Forward	Testing increased organization in recall	Retrieval and/or encoding strategy change

Yang et al., 2022), but the participants in these studies were native Chinese speakers, so they did not constitute foreign language learning. Chinese is different from Swahili or Euskara because it is a logographic, rather than an alphabet-based, language. Thus, participants who learn Chinese as a foreign language must encode the characters as abstract visuospatial representations (Flaherty, 2003; Shen, 2010). Given that most studies in the forward testing effect literature have used verbal materials such as word lists (Ahn & Chan, 2022; Szpunar et al., 2008), paired associates (Davis & Chan, 2015; Kornell et al., 2009), or prose and lecture materials (Jing et al., 2016; Wissman et al., 2011), investigating another type of representation such as visuospatial information is important for the generality of the phenomenon (cf., Kang, 2010), especially given that Chinese is the most spoken language in the world. We did not have any prior reason that using Chinese materials would yield a different effect compared to using verbal materials (partly because these materials have received little attention thus far), but as Hintzman (2011) suggested, examining the effect with various materials is important to form more comprehensive theories, and our study is to first to examine whether the forward testing effect extends to pictorial languages.

Figure 1 shows the design of our experiments. In Experiment 1, participants reported their strategies after every list to examine whether participants' strategies changed across Lists 1–3. In Experiment 2, participants reported their strategies only after studying List 3, because we were concerned that the requirement of reporting strategies after every list might trigger metacognitive reactivity (Double & Birney, 2019)—i.e., reporting on strategies might cause a change in how participants approach the learning task beyond the putative effects of testing on learning. Experiment 3 was a combination of Experiments 1 and 2, such that reporting frequency was manipulated by having some participants report their strategies after every list (Strategy Report 3×), and others reported only after the target list (Strategy Report 1×). In Experiments 4 and 5, we manipulated interim activity within-subjects. Experiment 4 was a smaller exploratory study, and Experiment 5 was a higher power replication of Experiment 4.

Given that testing encourages learners to put more effort into their learning (Endres & Renkl, 2015; Pyc & Rawson, 2012), we expected that tested participants would employ deep encoding strategies such as mental imagery and relational strategies such as inter-item associations (Chan, Manley, et al., 2018) and restudied participants would employ shallow strategies such as rote rehearsal (Cho & Powers, 2019).

In addition to examining the forward testing effect for foreign language learning and strategy change, we also examined whether interim testing would promote the transfer of learning *to new Chinese characters*. In all experiments, participants took a four-alternative forced choice *transfer test*, in which they had to guess the meanings of novel Chinese characters not presented previously. All these new characters included a studied radical, and the test, which was modeled after the one from Cho and Powers (2019), was designed to provide a measure of conceptual learning. Because Cho and Powers (2019) found that testing boosted the transfer of learning to novel Chinese characters, we also expected that the tested participants would outperform their restudy counterparts in this transfer test.

Data Availability

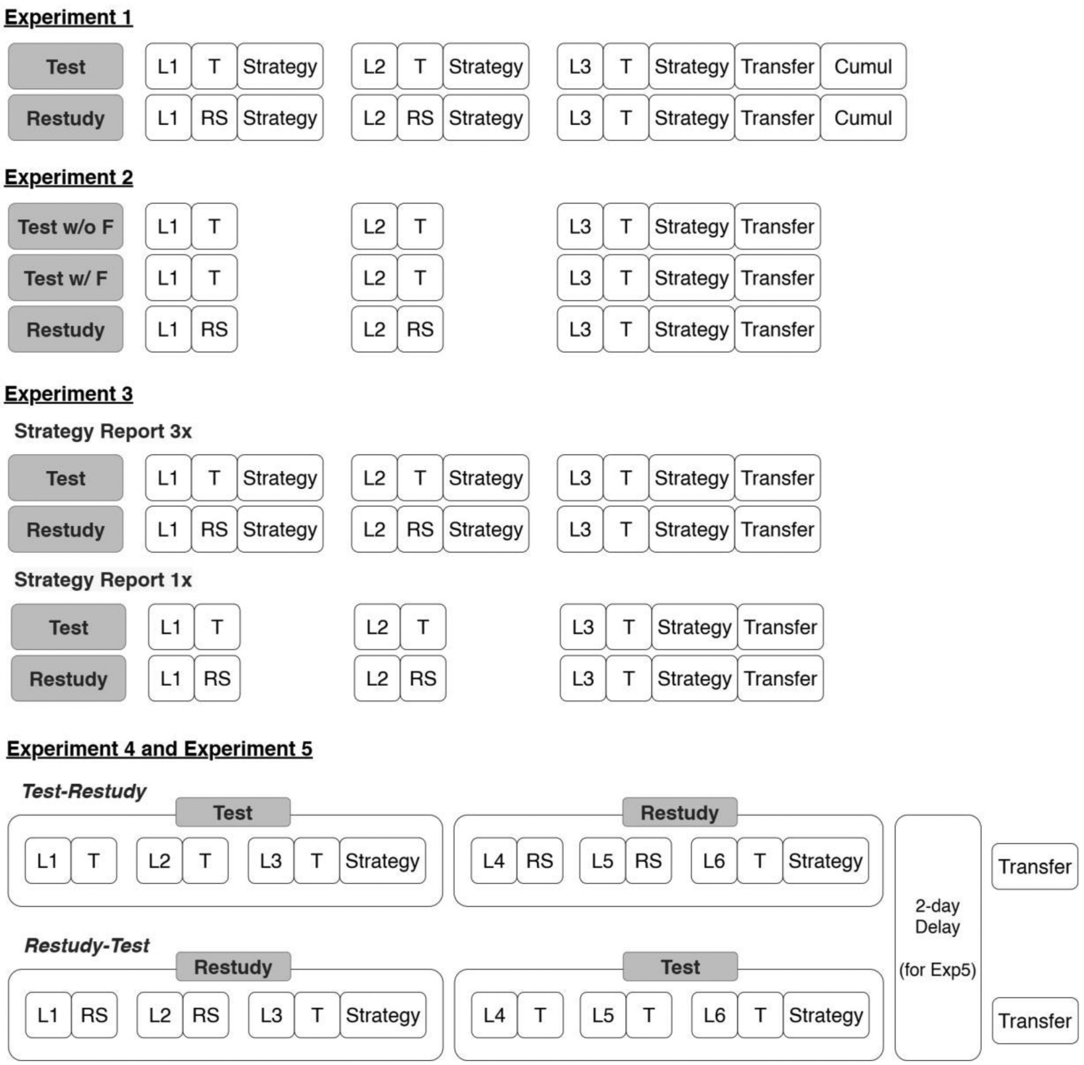
All experiments, except the exploratory Experiment 4, were preregistered on the Open Science Framework (OSF) at https://osf.io/uh8wr/. The preregistration included experimental protocols, and the coded data are available on the project page.

Experiment 1

Design and Participants

Interim activity (test vs. restudy) was manipulated between-subjects. We conducted a power analysis using G*Power (Faul et al., 2007) to determine the sample size. Our effect size was estimated based on the average effect size ($d = 0.63$, the effect of testing on transfer of learning) between Experiments 2 and 3 in Cho and Powers (2019). Although their study did not investigate the forward testing effect, we deemed it appropriate to use their effect size because they used Chinese characters as material, and the forward testing effect can be seen as one type of transfer in learning (Carpenter, 2012). To achieve 85% power to detect an effect size of .63 at a .05 α error probability, 37 participants per condition were necessary. We collected data from 95 undergraduate students at Iowa State University, who completed the experiment for course credits. Five data exclusion criteria were set in advance (see preregistration on OSF) and implemented as follows: Participants who failed an attention check ($N = 12$), those who knew Chinese ($N = 4$), those who reported not being alert at all during the study ($N = 4$), those who used a mobile device to complete the study ($N = 1$), and those who took notes during the study ($N = 1$). The final sample included data from 73 participants (37 women, 36 men; $M_{\text{age}} = 19.98$), with 36 in the test condition and 37 in the restudy condition.

Figure 1
Design for the Five Experiments


Experiment 1
Test | L1 T Strategy | L2 T Strategy | L3 T Strategy Transfer Cumul
Restudy | L1 RS Strategy | L2 RS Strategy | L3 T Strategy Transfer Cumul
Experiment 2
Test w/o F | L1 T | L2 T | L3 T Strategy Transfer
Test w/ F | L1 T | L2 T | L3 T Strategy Transfer
Restudy | L1 RS | L2 RS | L3 T Strategy Transfer
Experiment 3
Strategy Report 3x
Test | L1 T Strategy | L2 T Strategy | L3 T Strategy Transfer
Restudy | L1 RS Strategy | L2 RS Strategy | L3 T Strategy Transfer
Strategy Report 1x
Test | L1 T | L2 T | L3 T Strategy Transfer
Restudy | L1 RS | L2 RS | L3 T Strategy Transfer
Experiment 4 and Experiment 5
Test-Restudy
Test | L1 T | L2 T | L3 T Strategy
Restudy | L4 RS | L5 RS | L6 T Strategy
Restudy-Test
Restudy | L1 RS | L2 RS | L3 T Strategy
Test | L4 T | L5 T | L6 T Strategy
2-day Delay (for Exp5)
Transfer
Transfer

Note. L = Study list, T = Test, RS = Restudy, Strategy = Strategy reporting, Transfer = Transfer test, Cumul = Cumulative test. In Experiment 2, Test w/o F means *test without feedback* condition, and Test w/ F means *test with feedback* condition. Experiments 4 and 5 used the same procedure except that Experiment 5 employed a 2-day delay before the transfer test.

Materials

Chinese Characters

We adapted Cho and Powers' (2019) materials as a starting point but created our own Chinese–English pairs. We chose 12 characters from seven radicals each (i.e., eye, hand/arm, speech, water, woman, fire, and tree) and avoided very complicated characters (e.g., 護 or 熾) and those with very similar or ambiguous meanings. Then, a pilot test ($N = 31$) was conducted for the purpose of choosing which radical sets to use. In the pilot, participants studied four lists. Each list comprised three characters from four radicals for a total of 12 characters. Each Chinese–English pair was presented for 4 s, and participants studied each list twice in a different random order. After studying each list, participants took a cued recall test in which the Chinese character was presented and they had 8 s to type the corresponding English meaning.

Four radical sets (i.e., eye, hand/arm, speech, and water) with similar mean accuracy (i.e., 64%–68%) were chosen for the experiments. Among the chosen characters, the ones that we deemed too easy (more than 80% accuracy) or too difficult (less than 30% accuracy) were replaced by those with moderate difficulty (around 65% accuracy) after another pilot experiment ($N = 29$). The full material set is provided in Appendix A. The mean accuracy was similar across the four finalized radical sets, $F(3, 44) = 0.49$, $p = .693$, $B_{01} = 5.75$. We selected three items from each radical set to form four lists with a similar difficulty, $F(3, 44) = 0.04$, $p = .990$, $B_{01} = 8.73$. Three lists were presented during the learning phase, with the remaining list reserved for the transfer test. The assignment of items to lists was counterbalanced across participants, such that one participant might have studied 江—river in List 1, whereas another might have studied this pair in List 2 or 3. Alternatively, this item might not appear in the study lists but instead serve as an item on the transfer test.

Learning Strategies

The learning strategy questionnaire was constructed based on previous studies that examined learners' strategy use, but with questions specifically designed to assess strategies that were relevant to the learning of Chinese–English pairs (Cho & Powers, 2019; Finley & Benjamin, 2012; Hertzog & Dunlosky, 2004). To make questionnaire items more relevant to the current study, we asked participants in our pilot study to report their learning strategies with an open-ended question (i.e., Please describe the strategies that you used to remember the Chinese characters). Specifically, the first author coded participants' answers into several categories and modified or added several items to the strategy list. For example, Cue Focus Strategy (i.e., focused more on the left-hand Chinese Characters) in Cho and Powers (2019) was changed to Radical Focus Strategy (i.e., I focused more on the left side of each Chinese character). This change was made because some participants reported that they noticed the existence of radicals and found characters sharing the same radical were related to one another (e.g., *words that are related to the same subject had similar shapes of characters*), which was different from the general idea of focusing on the Chinese character as a whole (which always appeared on the left side, with the English meaning on the right side). Associative imagery strategy (i.e., I matched the Chinese characters with what their meaning would look like) was added to the strategy list because several participants indicated that they tried to link the shapes of characters to the words they matched (e.g., *I made the Chinese characters into pictures that symbolized the meaning*). The 12 learning strategies are presented in Table 2.

Procedure

All experiments were programmed using Qualtrics, and data collection was conducted online due to COVID-19. Participants were asked to eliminate any distractions in their environment and to complete the study in a single sitting.

Participants were informed that they would study lists of Chinese–English pairs. They were also told that after studying each list, they would solve some math problems and then either take a test on the just-presented list or restudy the list. Participants were told that whether or not they would be tested was determined randomly on a list-by-list basis. But in reality, participants in the test condition took a test after every list, whereas those in the restudy condition only took a test on List 3 and restudied Lists 1 and 2. Participants were also told that there would be a final cumulative test.

Before the presentation of each list, participants saw a prompt denoting the list number (e.g., This is Chinese–English word pair List 1). The list presentation started after participants clicked the arrow button on the screen. Each pair was presented for 4 s in a random order. The list was presented twice in a row, and the second presentation was ordered differently than the first. After the presentation of a list, participants solved 10 simple math problems at 6 s apiece to clear their short-term memory. Then participants in the test condition took a cued recall test on the list. Specifically, Chinese characters were presented with a question mark (e.g., 松–?) and participants had 8 s to recall the English meaning. Feedback was not provided. Instead of the recall test, participants in the restudy condition studied the same list for a third time in a different random order. To equate duration with the recall test, each pair was presented for 8 s. After the recall test or restudy presentation, participants completed the strategy survey by indicating the extent to which they used the 12 learning strategies while studying the just-presented list, using a 6-point Likert scale (see Figure S1 in the online supplemental material on OSF for the screenshot of the survey). Each choice was converted to a number during analyses (*Never* = 0, *Rarely* = 1, *Sometimes* = 2, *About half the time* = 3, *Most of the time* = 4, *Always* = 5).

After the presentation of List 3 and the math problems, all participants took a cued recall test and then completed the strategy survey for List 3. Additionally, they also completed another survey about strategy effectiveness separately. Specifically, participants were asked to indicate how effective they thought each of the 12 strategies was, using a 6-point Likert scale from Not effective at all to Extremely effective. Participants were told to guess the effectiveness if they had not used a certain strategy. The strategy effectiveness data were not analyzed because, in hindsight, we suspect that participants might have treated the strategy effectiveness and strategy usage questionnaires similarly.

Afterward, all participants took a transfer test for 12 new Chinese characters that did not appear during the study phase. Specifically, a Chinese character with a question mark (e.g., 眩–?) was presented, and participants were asked to guess its meaning (dizzy) among four options. To ensure that participants could not answer the

Table 2
List of Strategies

Strategy	Text in the questionnaire	Origin
Rote rehearsal	I repeated individual pairs over and over.	Cho and Powers (2019)
Mental imagery	I formed a mental image of the English meaning in my head.	Cho and Powers (2019)
Verbalization	I spoke words out loud or mouthed the words.	Finley and Benjamin (2012)
Personal significance	I related the words to something personally significant.	Finley and Benjamin (2012)
Looking	I focused on the word pairs by looking or staring.	Hertzog and Dunlosky (2004)
Associative imagery	I matched the Chinese characters with what their meaning would look like.	New
Avoid distraction	I told myself not to be distracted.	New
Intra-item narrative	I made a story for each Chinese–English pair.	Cho and Powers (2019)
Inter-item narrative	I combined multiple Chinese–English pairs into a story.	Cho and Powers (2019)
Radical focus	I focused more on the left side of each Chinese character.	Cho and Powers (2019)
Inter-item associations	I made associations amongst multiple Chinese characters that I've seen.	Cho and Powers (2019)
Self-testing	I tested myself during learning.	New

transfer questions correctly based on familiarity, each foil was associated with the meaning of a different *studied radical*. For example, the answer options were *read, lake, dizzy,* and *throw*, whose meanings were associated with the studied radicals of *word, water, eye,* and *hand/arm*, respectively. After the transfer test, participants took a cumulative cued recall test for Lists 1–3. Finally, participants completed a short survey on demographics, then were debriefed and thanked.

Results and Discussion

For the strategy questionnaire data, we initially planned to conduct a factor analysis in Experiment 1 but later realized that the sample size was too small to conduct an exploratory factor analysis. The recommended size is a minimum of 10 cases per measure (Russell, 2002). Given that we had 12 strategy questions, we needed 120 participants to satisfy this requirement. Thus, we collected more data in Experiments 2 and 3 using the same strategy questionnaire and then combined the data from across Experiments 1–3 for the factor analysis. Therefore, results for the strategy questionnaire were not presented for individual experiments. Instead, they are reported in the "Combined Analysis Across E1–E3" section. Below, we present results that pertain to the forward testing effect and the transfer of learning.

Proportions of interim test performance across all experiments are presented in Table 3. For all experiments, we first report target list correct recall to examine whether the forward testing effect emerged with Chinese character learning. Then, we report transfer recognition test performance. Lastly, the cumulative test result for Experiment 1 was reported on OSF in the online supplemental material. Because the cumulative test was not the focus of our research, we dropped it from all subsequent experiments.

For statistical inferences, two-tailed tests with $\alpha = .05$ were used. Bayes factors are reported for all analyses. When a result is significant in null hypothesis testing, we report B_{10}, for which a greater number indicates more support for the alternative hypothesis. When a result is not significant, we report B_{01}, for which a greater number indicates more support for the null hypothesis. This approach was taken for easier interpretation because a larger Bayes factor always provides more support for the effect (null or otherwise) under consideration. All Bayesian analyses were performed using the default priors in JASP (JASP Team, 2020).

List 3 Correct Recall

The left panel of Figure 2 shows the proportion of List 3 correct recall. An independent *t*-test was conducted with correct recall performance—i.e., correctly recalling the English meaning of a studied Chinese character—as the dependent variable and interim activity as the independent variable. The tested participants showed greater recall of the List 3 characters ($M = .51$) than the restudied participants ($M = .36$), $t(71) = 2.16$, $p = .034$, $d = 0.51$, $B_{10} = 1.75$, demonstrating a forward testing effect. This result shows that interim testing can enhance the learning of written Chinese as a foreign language, thus extending the forward benefit of testing from alphabet-based foreign languages to a pictorial one (Cho et al., 2017; Yang et al., 2017).

Table 3
Nontarget, Interim Test Performance Across Experiments 1–5

Experiment	List 1	List 2
Experiment 1	.42 (.26)	.59 (.23)
Experiment 2	.45 (.25)	.47 (.28)
Experiment 3	.45 (.25)	.52 (.27)
Experiment 4	.55 (.31)	.57 (.29)
Experiment 5	.55 (.29)	.58 (.29)

Note. *SD*s are in parentheses.

Transfer Test

The right panel of Figure 2 shows the proportion of correct responses on the transfer test. At first glance, one might wonder why the performance on the transfer test was similar to that of the List 3 memory test, given that transfer performance is typically lower than memory performance. We hasten to remind readers that the List 3 test was *recall* and the transfer test was *recognition*, so their performances are not directly comparable.

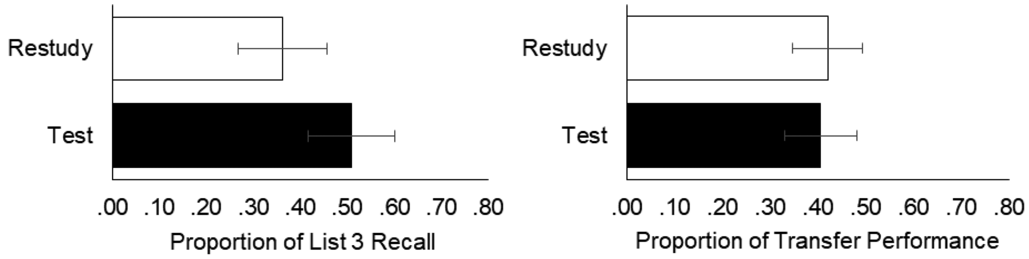
In striking contrast to the List 3 recall results, interim testing had virtually no effect on transfer performance ($M_{\text{test}} = .41$, $M_{\text{restudy}} = .42$), $t(71) = 0.22$, $p = .825$, $d = 0.05$, $B_{01} = 4.05$. This result differs from the one reported by Cho and Powers (2019), who found that retrieval practice promoted the transfer of learning for Chinese characters. One difference between our study and Cho and Powers' (2019) was that we did not provide feedback to participants, which might explain our absence of a testing benefit on the transfer of learning. However, because the tested participants were not informed of whether or not their answers were correct during the interim tests, connecting the radicals with similar meanings might have been too difficult, thereby negating the potential benefits of testing on transfer. This possibility is bolstered by data showing that providing feedback during tests can facilitate conceptual learning (Finn et al., 2018; Jacoby et al., 2010).

To address this possibility in Experiment 2, we added a condition in which participants received feedback during the interim tests. The logic is that testing, together with feedback, might enhance conceptual learning relative to restudy (Butler et al., 2013). Moreover, feedback is rarely manipulated in the forward testing effect literature. In fact, in a meta-analysis, Chan, Meissner, and Davis (2018) showed that providing feedback actually weakened the forward testing effect, although they cautioned against overinterpreting that finding (a) given how few studies included feedback and (b) the inclusion of feedback was confounded with other manipulations (e.g., interleaving of retrieval practice and new learning, Davis & Chan, 2015; Davis et al., 2017). Consequently, the provision of feedback in Experiment 2 offered a good opportunity to examine the influence of providing feedback on the forward testing effect.

Another possibility for the null effect of testing on transfer is that completing a strategy survey after every list might have helped the restudied participants discover effective strategies that they otherwise would not have considered, thereby minimizing the performance gap between participants in the test and restudy conditions (Double et al., 2018). For example, the statement "I made associations amongst multiple Chinese characters that I've seen" might have encouraged some participants to associate different characters with each other and to uncover the existence of radicals. Similarly, the statement "I focused more on the left side of each Chinese character" might have served as an instruction for participants to pay

Figure 2

Accuracy in Target List Recall Test and Transfer Multiple-Choice Test of Experiment 1



Note. Error bars display descriptive .95 confidence intervals.

closer attention to the radicals. We suspect that these strategy statements might not have been as helpful for the tested participants as they were for the restudied participants because taking the interim tests was hypothesized to promote more effective strategy use, so the statements might be redundant for participants in the test condition. In sum, answering the strategy survey might have informed the restudied participants about the strategies they could use during subsequent learning, which they otherwise would not have considered. Accordingly, in Experiment 2, we removed the strategy report after Lists 1 and 2, and participants reported their strategies only after List 3.

Experiment 2

In Experiment 2, two changes were implemented to examine why interim testing did not enhance performance on the transfer test in Experiment 1 relative to interim restudy. First, a test with feedback condition was added. If feedback during the interim test facilitates conceptual learning, participants in the test condition should perform better in the transfer test with feedback than without feedback (and restudy)—we termed this the *feedback hypothesis*. Second, participants reported their strategies only after List 3. If the restudied participants' transfer test performance was enhanced due to the exposure of strategies after Lists 1 and 2, then their performance should decrease when there was no strategy survey for Lists 1 and 2—which we named the *strategy introduction hypothesis*.

Design, Participants, and Procedure

Interim activity (test with feedback vs. test without feedback vs. restudy) was manipulated between-subjects. For the power analysis, the forward testing effect size from Experiment 1 ($d = 0.49$)[2] was used. A power analysis with the same criteria as Experiment 1 indicated that 61 participants per condition were necessary. Given that there were four counterbalances, we aimed to collect 64 participants per condition. In contrast to Experiment 1, which collected data using a college student sample, participants in Experiment 2 were 223 workers recruited via online experiment sampling platform Prolific. Participation was restricted to people between the age of 18–35, US nationality, and English-speaking monolinguals. Each participant was paid $3.50. Data from 26 participants who knew Chinese ($N = 20$) and took notes ($N = 6$) were excluded from analyses. The final sample included data from 197 participants (113 women, 83 men, and one non-binary; $M_{\text{age}} = 26.84$), with 65 in the test with feedback condition, 69 in the test without feedback condition, and 63 in the restudy condition. The procedure was identical to Experiment 1 except for three changes. First, the test with feedback condition was added. In this condition, the Chinese character with a question mark was presented for 6 s for participants to recall the English meaning, and then the corresponding English meaning was presented for 2 s. Second, all participants completed the strategy survey only after List 3. Lastly, the cumulative final test was omitted.

Results and Discussion

As mentioned earlier, the data for the strategy survey were analyzed after Experiment 3, when we combined the data collected from Experiments 1–3.

List 3 Correct Recall

The left panel of Figure 3 shows the proportion of List 3 correct recall. An analysis of variance (ANOVA) revealed that the main effect of interim activity was significant, $F(2, 194) = 14.58$, $p < .001$, $\eta_p^2 = .13$, $B_{10} = 12{,}246.04$. Specifically, the tested participants, with feedback ($M = .44$) or not ($M = .47$), recalled far more English meanings of the Chinese characters than the restudied participants ($M = .23$), ts > 4.47, ps $< .001$, ds > 0.78, B_{10}s $> 1{,}095.17$. Thus, the forward testing benefit for Chinese foreign language learning has been generalized to a non-college sample and a condition with feedback.

The issuance of feedback also allowed us to examine whether feedback affects the magnitude of the forward testing benefit—it did not, $t(132) = 0.49$, $p = .623$, $d = 0.09$, $B_{01} = 4.84$. This result indicated that the administration of feedback during initial tests did not enhance new learning further, nor did it cause the forward testing effect to diminish, as the moderator analysis from Chan, Meissner, & Davis (2018) showed.

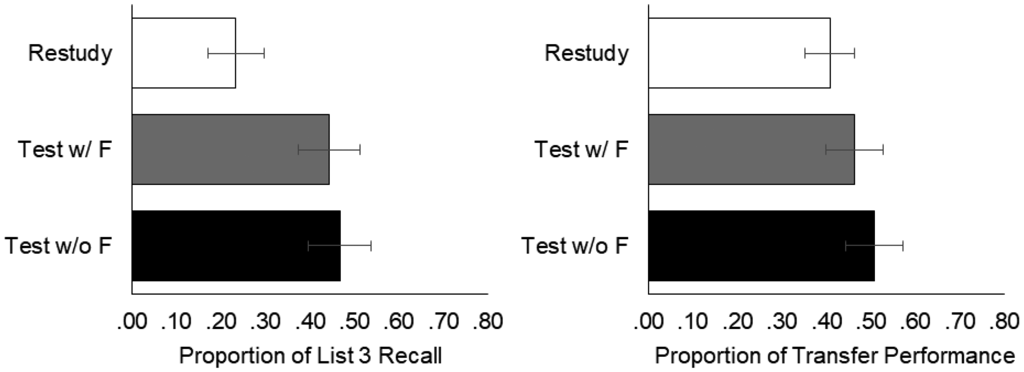
Transfer Test

The right panel of Figure 3 shows that the main effect of interim activity on transfer performance was not significant, $F(2, 194) = 2.70$, $p = .069$, $\eta_p^2 = .03$, $B_{10} = 1.80$. But somewhat unexpectedly, participants in the test without feedback condition ($M = .51$) selected the correct meaning of the novel Chinese items more

[2] The actual forward testing effect size in Experiment 1 was $d = 0.51$, but the power analysis was conducted when data collection of Experiment 1 was almost complete, and the effect size of the incomplete sample was $d = 0.49$.

Figure 3

Accuracy in Target List Recall Test and Transfer Multiple-Choice Test of Experiment 2



Note. w/ F = with feedback, w/o F = without feedback. Error bars display descriptive .95 confidence intervals.

frequently than those in the restudy condition ($M = .41$), $t(130) = 2.37$, $p = .019$, $d = 0.41$, $B_{10} = 2.31$, whereas the participants in the test with feedback condition did not ($M = .46$), $t(126) = 1.31$, $p = .193$, $d = 0.23$, $B_{01} = 2.43$ (note that both of the above Bayes factors provided only anecdotal support for their respective findings). These results are surprising because the feedback hypothesis proposed that the absence of feedback in the test condition in Experiment 1 might have hampered the transfer of learning for the tested participants compared to the restudied ones. Therefore, the prediction was that a transfer effect would occur when participants were given feedback. In sum, the data in Experiment 2 did not provide support for the feedback hypothesis.

The strategy introduction hypothesis proposed that the implementation of a learning strategy survey after every list might have informed restudied participants of learning strategies that they might not have considered. If this were the case, removing the interim strategy surveys should eliminate this possibility and participants in the test conditions, regardless of feedback, should show a superior *transfer of learning* compared to participants in the restudy condition—but our data showed that only participants in the test without feedback condition, but not in the test with feedback condition, showed a transfer effect.

Experiment 3

The data in Experiment 2 did not support either the feedback hypothesis or the strategy introduction hypothesis, which were both proposed to explain the null effect of interim testing on the transfer of learning. Moreover, we unexpectedly found a small transfer of learning effect in the Test without feedback condition in Experiment 2, which was not found in Experiment 1. To ascertain the replicability of our findings, we manipulated both interim task and strategy survey frequency in Experiment 3. We omitted the Test with feedback condition in Experiment 3 because it had little influence on both recall and transfer performance.

Design, Participants, Materials, and Procedure

Interim activity (test vs. restudy) and frequency of strategy surveys (report 1× vs. report 3×) were manipulated between-subjects. The effect size for the power analysis was based on the transfer test performance ($d = .41$) between the test without feedback and the restudy conditions in Experiment 2, and this analysis showed that 91 participants per condition were necessary. Participants were 416 undergraduate students from Iowa State University who completed the experiment online for course credits. Data from participants who failed an attention check ($N = 25$), those who knew Chinese ($N = 11$), those who used a mobile device to complete the study ($N = 6$), those who took notes ($N = 5$), and those who have participated in Experiment 1 ($N = 2$) were excluded from analyses. The final sample included 367 participants (230 women, 137 men; M_{age} 19.16), with 91 in the test-report 3× condition, 93 in the test-report 1× condition, 91 in the restudy-report 3× condition, and 92 in the restudy-report 1× condition. The materials and procedure were a combination of Experiment 1 (without the cumulative test) and Experiment 2 (without the test with feedback condition).

Results and Discussion

List 3 Correct Recall

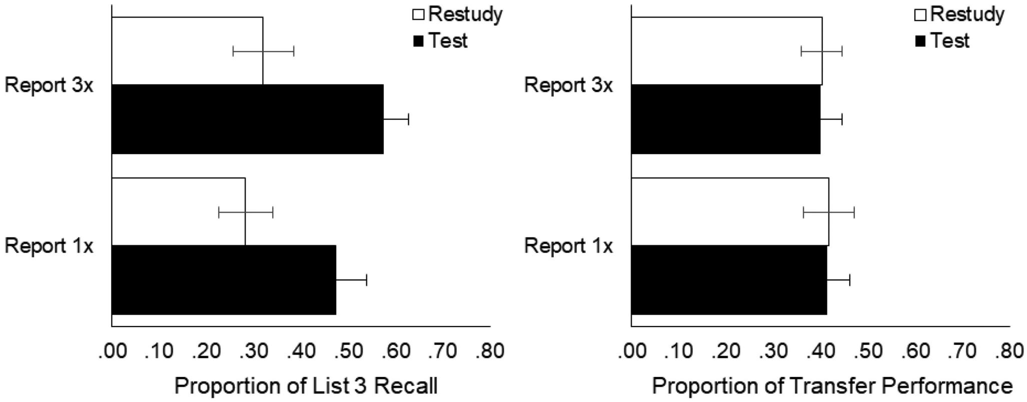
The left panel of Figure 4 shows the proportion of List 3 correct recall. A 2 (interim activity: *t*est vs. restudy) × 2 (report: 1× vs. 3×) ANOVA revealed a significant forward testing effect, with the tested participants ($M = .52$) recalling the meaning of far more Chinese characters than the restudied participants ($M = .30$), $F(1, 363) = 54.50$, $p < .001$, $d = 0.77$, $B_{10} = 4.427\text{e}+9$. Strategy reporting frequency also affected learning. Participants who completed the strategy survey after every list ($M = .45$) were slightly more likely to recall the English meanings than those who completed the survey after only the final list ($M = .38$), $F(1, 363) = 5.06$, $p = .025$, $d = 0.22$, $B_{10} = 0.93$ (but note the small effect size and the virtually neutral Bayes Factor). The interaction was not significant, $F(1, 363) = 1.04$, $p = .309$, $\eta_p^2 = .00$, $B_{01} = 3.95$.

Transfer Test

For the transfer test (see the right panel of Figure 4), neither interim activity, $F(1, 363) = 0.03$, $p = .864$, $d = .02$, $B_{01} = 8.55$, nor reporting frequency affected performance, $F(1, 363) = 0.40$, $p = .526$, $d = .07$, $B_{01} = 7.13$. Of greater interest, the interaction was also not significant, $F(1, 363) = 0.00$, $p = .984$, $\eta_p^2 = .00$,

Figure 4

Accuracy in Target List Recall Test and Transfer Multiple-Choice Test of Experiment 3



Note. Error bars display descriptive .95 confidence intervals.

$B_{01} = 6.82$. Similar to Experiment 1, transfer performance did not differ between the test-report 3× condition ($M = .40$) and the restudy-report 3× condition ($M = .40$), $t(180) = 0.14$, $p = .886$, $d = 0.02$, $B_{01} = 6.16$. In fact, performance in the two conditions were virtually identical. Further, we did not replicate the surprising finding of Experiment 2, in which the tested participants outperformed the restudied participants when the former reported strategies only for List 3. Instead, participants in the test-report 1× condition ($M = .41$) and those in the restudy-report 1× condition ($M = .42$) performed nearly identically, $t(183) = 0.10$, $p = .919$, $d = 0.02$, $B_{01} = 6.24$. Given that the sample size was larger in Experiment 3 ($N_{\text{test-report } 1\times} = 93$) than in Experiment 2 ($N_{\text{test w/o F}} = 69$), and the transfer effect was observed in only one comparison across three experiments, we suspect that the testing benefit on the transfer of learning in Experiment 2 was spurious. Given that the overall performance on the transfer test (~40%) in Experiments 1–3 seems quite low, one could think that participants were just guessing. However, although 40% performance might sound low, it far exceeds chance level performance in a four alternative forced choice test of 25%, $t(636) = 18.82$, $p < .001$, $d = 1.81$, $B_{10} = 3.445\text{e}+59$.

Overall, when considering the transfer of learning effect, our results clearly differed from Cho and Powers (2019), who observed that testing promoted the transfer of learning of Chinese characters. Although we have already addressed the difference of feedback, there was still an outstanding difference between our studies—namely, that participants in Cho and Powers' (2019) experiments studied the same list six times, whereas participants in our experiments studied three different lists twice. This discrepancy was unavoidable because we had disparate research goals. The different number of repetitions of the study list might play a role in the discrepancy between the results of our study and Cho and Powers' (2019). Specifically, it might be difficult for participants in our study to associate the characters with one another because they studied each list only twice. Furthermore, characters with the same radicals were distributed across three lists in our experiments, which could make relating characters more difficult. In contrast, Cho and Powers' (2019) participants studied the same list repeatedly, which might make it easier for participants to extract the meaning of the radicals. We will address these possibilities more fully in the General Discussion, but at the very least, our results show that testing does not always promote recognition of novel Chinese characters compared to restudying.

Combined Analysis Across E1–E3

In the following, we sought to address two questions based on the combined data of Experiments 1–3. First, to what extent did strategy use change across Lists 1–3 in the test condition compared to the restudy condition? Second, did List 3 strategy use mediate the relationship between interim activity (i.e., test vs. restudy) and List 3 recall? To accomplish these goals, we first conducted a factor analysis to consolidate results of the strategy questionnaire; we then used the extracted factors for our analyses. It is important to consider that we conducted parametric tests for the following analyses despite our measurements being ordinal. However, Robitzsch (2020) suggested that under many circumstances, ordinal data can be treated as scale data for analysis purposes.

Factor Analysis

In our pre-registration, we initially planned to compare deep versus shallow learning strategies. But during our data collection process, we decided to conduct factor analysis instead. This was because, in hindsight, we thought shallow versus deep categorization might not fully reflect the latent structures of the strategies. Thus, we conducted an exploratory factor analysis to identify possible strategy factors. We chose this approach to avoid conducting a t-test on each of the 12 strategy items, which either greatly inflates the Type-1 error rate without multiple comparison correction or greatly reduces power when implementing multiple comparison correction.

Responses on List 3 strategy were used for the factor analysis. We conducted principal axis factor extraction with varimax rotation in JASP. To decide on the number of factors, a parallel analysis was conducted, which is recommended over the traditional method of retaining only factors with eigenvalues greater than 1.0 (Hayton et al., 2004; Ledesma & Valero-Mora, 2007). A three-factor solution explaining 32% of the variance emerged. The first factor accounted for 14% of the variance, with each of the other two factors accounting for an additional 9% of the variance. Intercorrelation among factors was low (rs $< .15$).

Factor loading of .30 was used as a criterion (Costello & Osborne, 2005; Howard, 2016) to select items for each factor (see Table 4). The first factor was named "relational strategy" because most of the items meeting the criterion (i.e., Inter-item narrative, Intra-item narrative, Inter-item associations, Personal significance, and Self-testing) were associated with learners trying to find relations among different items (Dumas et al., 2013). The Self-Testing item was dropped from the factor because it is not clearly a relational strategy, given that any self-testing that occurs during an encoding phase is likely based on working memory retrieval (i.e., similar to rehearsal).[3] The second factor was named "rehearsal strategy" because all three items satisfying the criterion (i.e., Rote Rehearsal, Looking, and Verbalization) were associated with rehearsal. The third factor was named "imagery strategy" because both items passing the criterion (i.e., Mental Imagery, Associative Imagery) were related to imagery. The Radical Focus and Avoid Distraction strategies did not belong to any factors. Strategy scores were calculated by averaging the ratings of items within each factor. Of course, there are multiple ways to compute factor scores. However, we opted to use the mean rating system because we wanted all items in a factor to be weighted equally. Note that this approach is not without precedents and is commonly employed in the neuropsychology literature (e.g., Chan & McDermott, 2007; Glisky et al., 1995).

Strategy Use Across Lists

To validate our strategy measures, we calculated Spearman's rank correlation between strategies and List 3 recall, and all three strategy factors were positively associated with List 3 recall ($\rho_{\text{relational}}$ [635] = .16, $\rho_{\text{rehearsal}}$ [635] = .16, ρ_{imagery} [635] = .25, ps $< .001$, B_{10}s > 26.67). Consequently, employing the encoding strategies (especially imagery) was beneficial to learning in the present experiments.

If testing affects encoding strategies, one might expect a difference in strategy use to emerge across study lists, such that the tested participants might begin favoring relational strategies relative to the restudied participants. To investigate this possibility, a 3 (List number: 1 vs. 2 vs. 3) × 2 (Interim activity: Test vs. Restudy) ANOVA was conducted for each of the three strategy factors as a dependent variable. List number was a within-subjects variable, and interim activity was a between-subjects variable.

Table 4
Factor Loadings for List 3 Strategy Factors

Item	Relational	Rehearsal	Imagery
Inter-item narrative	**.80**	.06	.07
Intra-item narrative	**.71**	−.02	.18
Personal significance	**.41**	.21	.25
Inter-item associations	**.31**	.28	−.01
Rote rehearsal	.00	**.58**	.20
Looking	.02	**.46**	.11
Verbalization	.12	**.43**	.08
Mental imagery	.08	.18	**.85**
Associative imagery	.19	.13	**.39**
Self-testing	.30	.28	.16
Radical focus	.22	.26	−.15
Avoid distraction	.14	.17	.09

Note. Bolded numbers indicate that items were chosen for each factor.

As shown in Figure 5, each ANOVA resulted in a significant interaction, all of which were characterized by a similar pattern, relational—$F(2, 506) = 5.22$, $p = .006$, $\eta_p^2 = .00$, $B_{10} = 3.30$; rehearsal—$F(2, 506) = 5.71$, $p = .004$, $\eta_p^2 = .00$, $B_{10} = 5.52$; imagery—$F(2, 506) = 12.65$, $p < .001$, $\eta_p^2 = .01$, $B_{10} = 3{,}281.44$. Specifically, from List 1 to List 3, the tested participants reported increased use of all three strategies, Fs > 6.40, ps $< .003$, η_p^2s $> .05$, B_{10}s > 9.96. In contrast, for the restudied participants, strategy use remained at a similar level across lists for relational, $F(2, 254) = 0.20$, $p = .815$, $\eta_p^2 = .00$, $B_{01} = 29.09$, and rehearsal, $F(2, 254) = 2.21$, $p = .112$, $\eta_p^2 = .02$, $B_{01} = 4.62$, strategies and decreased for imagery strategies, $F(2, 254) = 9.41$, $p < .001$, $\eta_p^2 = .07$, $B_{10} = 142.38$. These results suggest that interim testing may have prompted participants to increase the use of each strategy. However, as can be seen in Figure 5, the restudied participants unexpectedly reported greater use of List 1 strategy than the tested participants in all three strategies, indicating that the baseline was different between the two conditions, which clouded the interpretation of this interaction, ts > 3.10, ps $< .003$, ds $> .39$, B_{10}s > 12.33.

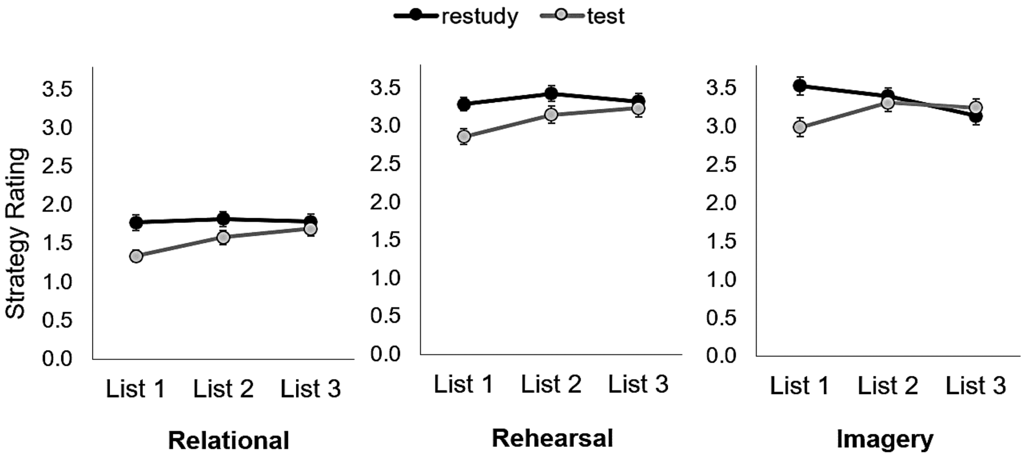
Taken together, the tested participants increased their use of strategies across Lists 1–3 for all factors, whereas the restudied participants maintained a similar or decreased strategy use. However, there was no qualitative shift in strategy use; instead, the tested participants simply reported increased use of all strategies. One might wonder why participants could increase their use of all strategies rather than shifting their use from one strategy to another across lists. Although strategy use being a zero-sum game is an intuitive concept, our survey did not allot participants with a maximum number of points and had them distribute those points across the 12 strategies. Instead, the strategies were not mutually exclusive, so participants could report having used more of every strategy after they were tested. For example, participants might report using more of all strategies because they put in more effort across the board and made use of strategies more often in general. Alternatively, as we have discussed in the paper, answering the strategy questionnaires might have made those strategies salient to participants, thereby increasing the reporting of multiple strategies. Perhaps most importantly, the tested participants did not use any of the List 3 strategy to a greater extent than the restudied participants, ts < 0.53, ps $> .513$, ds < 0.08, B_{01}s > 5.94, despite a substantial forward testing effect in List 3 recall. Although this dissociation in strategy use (i.e., null effect) and recall (i.e., forward testing effect) might be problematic for the strategy-change account, the null effect in List 3 strategy use might be attributable to the aforementioned baseline difference. If the baseline was similar across the two conditions, the tested participants might have reported greater use of strategies for List 3 than the restudied participants.

It is not clear why the restudied participants reported greater use of all three strategies during List 1 than the tested participants, but one possibility could be a sampling difference. Although we deemed this possibility remote, we still wanted to address it. Thus, we used a within-subjects design in Experiments 4 and 5 to eliminate

[3] We conducted the analyses with including the self-testing item, but the overall result pattern remained the same.

Figure 5

Strategy Use Change Across Lists 1–3 for Each Strategy Factor



Note. The strategy rating of 1, 2, and 3 meant *rarely*, *sometimes,* and *about half the time*. Error bars display standard errors.

between-group differences. Furthermore, we expected participants would be better able to distinguish strategies used when they could experience both conditions.

Did List 3 Strategy Use Mediate the Relationship Between Interim Activity and Correct Recall?

We conducted a confirmatory factor analysis (CFA) first to test the fit of the measurement model before conducting structural equation modeling (SEM) to test the mediation of strategy-change account. In the mediation model, we considered interim activity and reporting frequency as independent variables, strategy factor as a mediating variable, and List 3 correct recall as a dependent variable (see Figure 6). For CFA, the three strategy factors were treated as latent variables, and each strategy item was treated as a manifest variable. CFA was conducted using Mplus 7.0. The measurement model showed a poor fit, $\chi^2 = 268.47$, $df = 49$, $p < .001$, with multiple indices suggesting that the model did not fit the data well, RMSEA = .08, CFI = .81, SRMR = .07. Moreover, none of the beta coefficients between interim activity and strategy factors were significant, βs < .061, $p > .255$. Thus, the SEM was not conducted because there would be no mediation effect given that the independent variable (i.e., interim activity) was not associated with the mediators (i.e., strategy use), although interim activity was significantly associated with List 3 correct recall (β = .051, $p < .001$).

Discussion

The results in the combined analysis showed an absence of a mediation effect of strategies; moreover, List 3 strategy use did not differ significantly between the tested and restudied participants. The latter finding contrasts with the one reported by Cho and Powers (2019), who showed that tested participants reported greater use of inter-item association and diminished use of rote rehearsal than restudied participants. One might wonder if this discrepancy can be attributed to the different analysis approaches, as we reduced our strategy questionnaire data from 12 questions into three factors, whereas Cho and Powers (2019) conducted one *t*-test per strategy question. To explore this possibility, we compared each List 3 strategy question between the test and restudy conditions. Out of 12 *t*-tests, only the Avoid Distraction strategy (i.e., I told myself not to be distracted) revealed a significant difference, with higher usage reported by the restudied participants ($M = 2.92$) than the tested participants ($M = 2.55$), $t(641) = -2.63$, $p = .009$, $d = 0.21$, $B_{10} = 2.57$. Verbalization strategy (i.e., I spoke words out loud or mouthed the words) showed a marginal difference in favor of the tested participants ($M_{\text{restudy}} = 2.55$ vs. $M_{\text{test}} = 2.85$), $t(641) = 1.96$, $p = .050$, $d = 0.16$, $B_{10} = 0.58$. Therefore, even when we used the same analysis approach as Cho and Powers (2019), we found little evidence that testing had encouraged participants to use deeper encoding strategies.

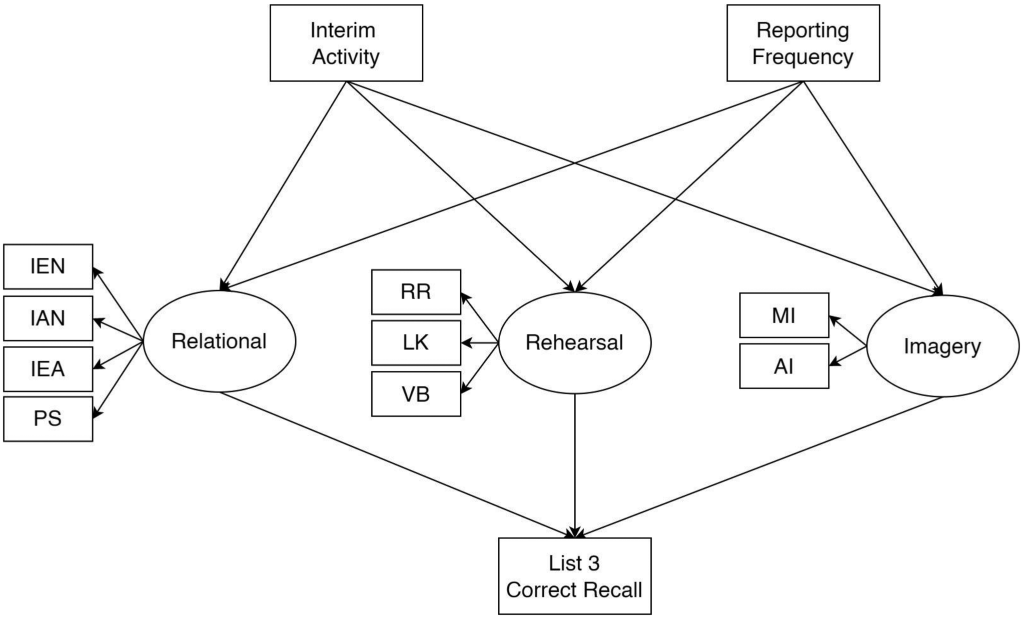
Another possibility that may have led to no difference in List 3 strategies between the test and restudy conditions is that reporting strategies after Lists 1 and 2 could have affected List 3 strategy use, because Cho and Powers (2019) had participants report strategy use only once. To examine this possibility, we conducted the same *t*-tests comparing List 3 strategies for only participants who reported their strategies once (i.e., Experiment 2 and the report 1× condition in Experiment 3). The results of this analysis showed a similar pattern, with the same two strategies showing a significant difference, Avoid Distraction—$t(381) = -2.76$, $p = .006$, $d = 0.29$, $B_{10} = 4.41$; Verbalization—$t(381) = 2.08$, $p = .038$, $d = 0.22$, $B_{10} = 0.91$).

Experiment 4

As a whole, our results suggest that testing did not influence participants' subsequent encoding strategies. But we wanted to investigate whether this finding still holds when we manipulate interim activity within-subjects. To this end, we conducted Experiment 4 as a non-preregistered pilot study. There were three reasons why we chose a within-subjects design. First, there was a baseline difference in strategy reporting between the test and restudy conditions, which suggests the possibility of a sampling issue. Second, participants might be able to better distinguish their strategy usage across conditions if they have experienced both, so the present experiment

Figure 6

The Strategy-Change Account Mediation Model



Note. Numbers are not included because we did not test the mediation model. We included this figure to depict the mediation model that we intended to test. IEN = Inter-item narrative, IAN = Intra-item narrative, IEA = Inter-item associations, PS = Personal significance, RR = Rote rehearsal, LK = Looking, VB = Verbalization, MI = Mental imagery, AI = Associative imagery.

might provide a more sensitive measure of strategy change. Lastly, we sought to investigate which encoding strategies might contribute to the forward testing effect at an individual level. Extant studies have not been able to associate the forward testing effect to a particular strategy use. By manipulating interim activity within-subjects, we could examine whether the magnitude of the forward testing effect (as demonstrated by each participant) is associated with increased usage of a particular encoding strategy between testing and restudying.

Design and Participants

Interim activity (test vs. restudy) was manipulated within-subjects. Because this study was conducted as a pilot, we did not conduct a power analysis in advance. Participants were 86 undergraduate students from Iowa State University who completed the experiment for course credits. Data of participants who knew Chinese ($N = 7$), failed an attention check ($N = 3$), did the same experiment previously ($N = 2$), and took notes ($N = 1$) were excluded from analyses. The final sample included 73 participants (56 women, 17 men; $M_{\text{age}} = 19.52$), with 36 completing the restudy condition first and 37 completing the test condition first.

Material and Procedure

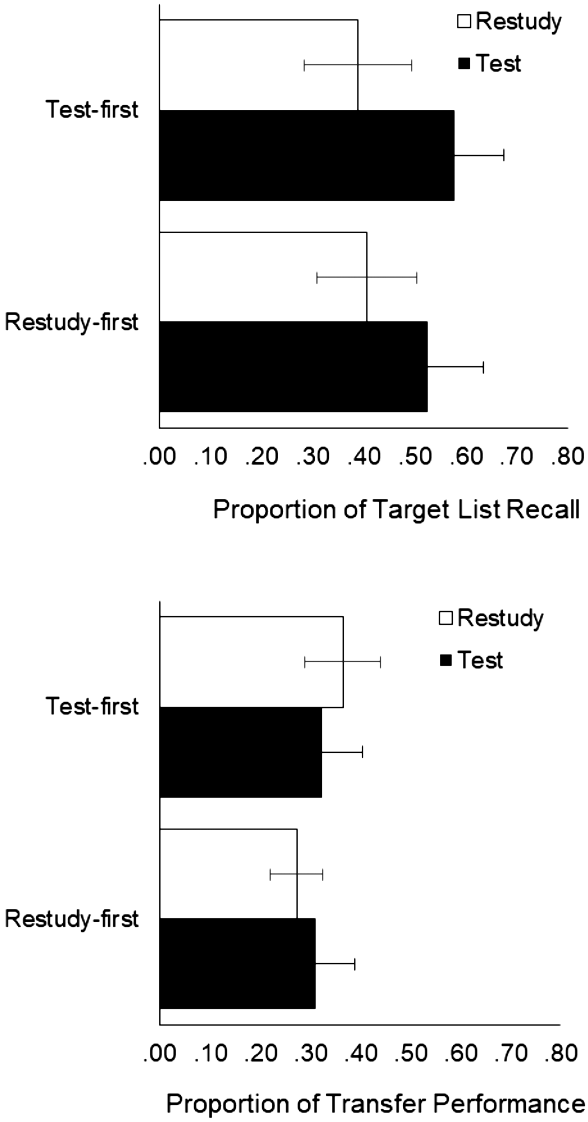
Two radicals (i.e., fire 火 and tree 木), each with 12 characters, were added to the four existing ones (i.e., eye, hand, water, speech) because more Chinese characters were necessary to conduct a within-subjects experiment. This resulted in 72 characters in total (see Appendices A and B). Two different material sets were constructed. Each set consisted of four lists of characters from three different radicals (i.e., Set A—eye, hand, fire; Set B—water, tree, speech). Each list contained three Chinese characters from each of the three radicals. Three lists from Set A and three lists from Set B were used for the list learning phase, with the fourth list from each set reserved for the transfer test. One set (i.e., A or B) was assigned to the restudy condition, and the other was assigned to the test condition. The order of the restudy or test conditions was counterbalanced across participants (i.e., restudy-first or test-first). This counterbalancing allowed us to examine whether the forward testing effect is affected by sequencing in a within-subjects design. For example, for participants who completed the test condition first, there might be some carryover effects of testing, so the forward testing effect could be smaller compared to participants who were assigned the opposite sequence. Immediately after the test of List 3 or 6, participants completed the strategy survey that was used in Experiments 1–3. Finally, at the end of the experiment, all participants took a transfer test on two novel lists (i.e., one from Set A and one from Set B).

Results and Discussion

Target List Correct Recall and Transfer Test

A 2 (interim activity: *t*est vs. restudy) × 2 (sequence: restudy-first vs. *t*est-first) ANOVA was conducted on target list recall (Figure 7 top panel). There was a main effect of interim activity, showing a within-subjects forward testing effect for Chinese foreign language

Figure 7
Accuracy in Target List Recall Test and Transfer Multiple-Choice Test of Experiment 4



Note. Error bars display descriptive .95 confidence intervals.

learning ($M_{\text{test}} = .55$, $M_{\text{restudy}} = .40$), $F(1,\ 71) = 15.69$, $p < .001$, $d = 0.47$, $B_{10} = 135.94$. Neither the main effect of sequence nor the interaction was significant: Sequence—$F(1,\ 71) = 0.08$, $p = .777$, $d = 0.07$, $B_{01} = 4.00$; Interaction—$F(1,\ 71) = 0.81$, $p = .372$, $\eta_p^2 = .00$, $B_{01} = 2.99$. There was no carryover effect of testing. Participants' restudy performance was similar regardless of sequence ($M_{\text{restudy-first}} = .41$ vs. $M_{\text{test-first}} = .40$), $t(71) = 0.24$, $p = .814$, $d = 0.06$, $B_{01} = 4.04$.

For transfer performance (Figure 7 bottom panel), neither the main effects nor the interaction was significant: Interim activity—$F(1,\ 71) = 0.00$, $p = .934$, $d = .01$, $B_{01} = 7.73$; Sequence—$F1(1,\ 71) = 1.46$, $p = .232$, $d = 0.28$, $B_{01} = 2.21$; Interaction—$F(1,\ 71) = 1.99$, $p = .162$, $\eta_p^2 = .00$, $B_{01} = 1.75$). Again, replicating the results of Experiments 1–3, transfer test performance was not affected by interim activity ($M_{\text{test}} = .32$ vs. $M_{\text{restudy}} = .32$).

Strategy Change and Forward Testing Effect

To examine whether there was a difference in strategy reports between the test and restudy conditions, we conducted a 2 (interim activity; test vs. restudy) × 3 (strategy factor; relational vs. rehearsal vs. imagery) ANOVA. We inherited the same strategy factors from Experiments 1–3 by averaging the ratings of items within each factor.

Table 5 shows strategy reporting by condition. There was a main effect of strategy factor, indicating that participants used each strategy factor to a different extent, $F(2,\ 144) = 45.47$, $p < .001$, $\eta_p^2 = .30$, $B_{10} = 2.021\text{e}{+}26$. Specifically, participants used imagery strategies most frequently ($M = 3.44$), then rehearsal strategies ($M = 2.99$), and lastly, relational strategies ($M = 1.98$). Neither the main effect of interim activity nor interaction was significant, $F\text{s} < 1.27$, $p\text{s} > .264$, $B_{01}\text{s} > 7.62$. Therefore, the data in Experiment 4 showed once again that interim testing did not change participants' encoding strategy compared to interim restudy, even when participants had experienced both interim tasks.

Although interim activity did not affect strategy reports at the condition level, the within-subjects design of this experiment allowed us to investigate the relationship between the forward testing effect and strategy use at an individual level. Specifically, we examined whether the size of the forward testing effect was associated with a change in strategy use between test and restudy *at an individual level.* For each participant, we calculated a *strategy difference* score (instead of the group level strategy scores displayed in Table 5) by subtracting the restudy strategy score from the test strategy score for each type of strategy. For example, if a participant answered the mental imagery question with a 3 following a restudy trial but a 5 (i.e., always) following a test trial, the difference score would be 2 (i.e., $5 - 3 = 2$). Using this method, we calculated the average difference scores for each of the three factors (i.e., relational, rehearsal, and imagery). The forward testing effect size was calculated for each participant by subtracting target list recall performance of the restudy condition from that of the test condition. We used Spearman's ρ for correlation.

Interestingly, as can be seen in Figure 8, there was a positive correlation between the forward testing effect and the imagery strategy difference score, $\rho(71) = .27$, $p = .024$, $B_{10} = 4.03$. The correlation was not significant for relational strategy, $\rho(71) = .15$, $p = .216$, $B_{10} = 1.51$, and rehearsal strategy, $\rho(71) = .07$, $p = .550$, $B_{01} = 5.30$. These results suggest that an increase in imagery strategy usage (from restudy to test) was associated with a larger forward

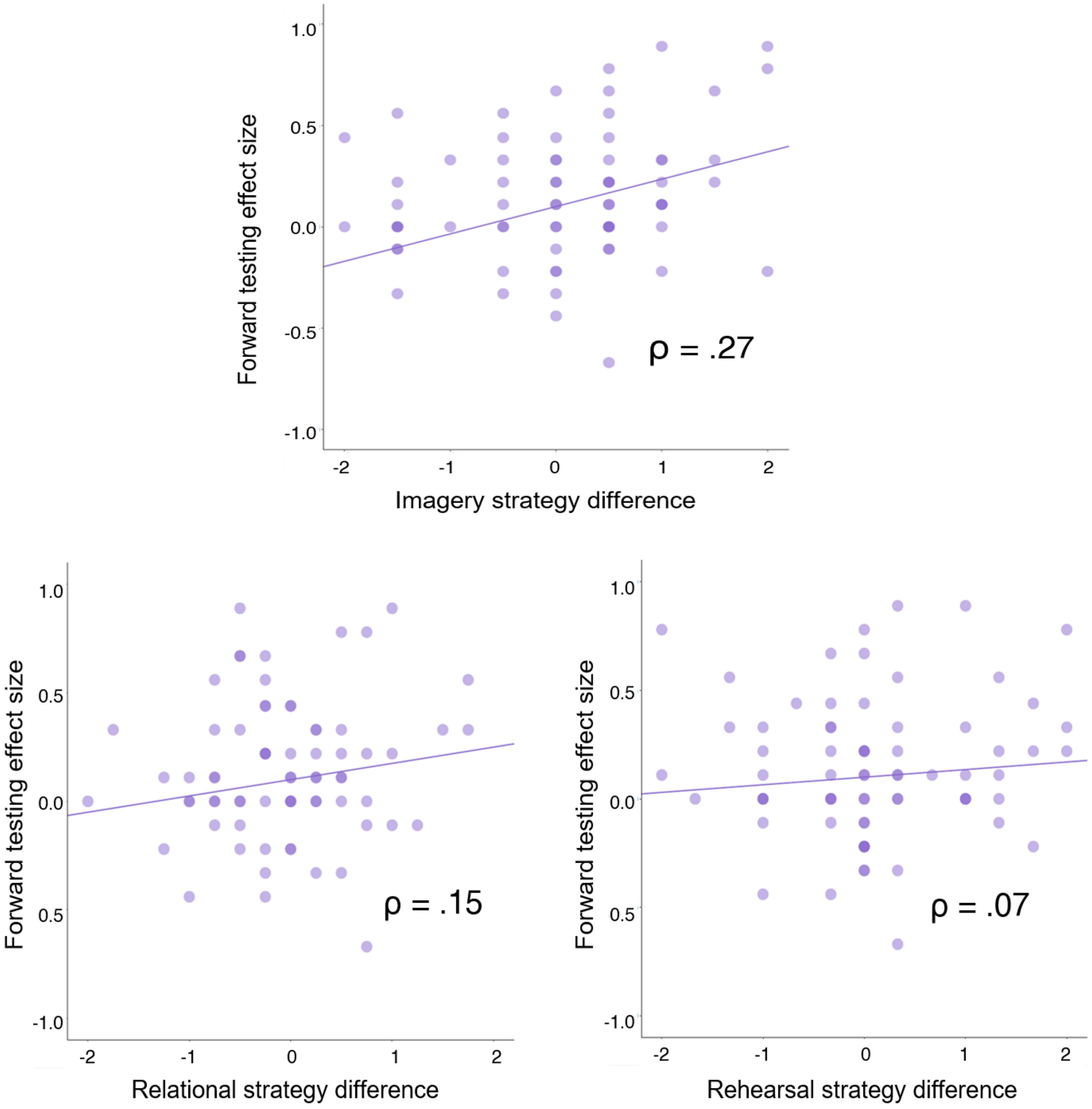
Table 5
Mean Strategy Reporting by Condition in Experiments 4–5

Experiment	Condition	Relational	Rehearsal	Imagery
Experiment 4				
	Test	1.99 (1.23)	3.04 (1.17)	3.50 (1.10)
	Restudy	1.97 (1.12)	2.95 (1.15)	3.37 (1.20)
Experiment 5				
	Test	1.89 (1.18)	2.88 (1.16)	3.25 (1.36)
	Restudy	1.85 (1.13)	2.86 (1.24)	3.13 (1.42)

Note. Standard deviations are in parentheses.

Figure 8

Scatterplots Showing the Association Between Strategy Difference Scores and Forward Testing Effect Size in Experiment 4



Note. Overlapping data points are shown as darker dots. See the online article for the color version of this figure.

testing effect. We discuss these findings after reporting the results from Experiment 5, which was preregistered and had a larger sample.

Experiment 5

In Experiment 5, we aimed to replicate the findings in Experiment 4 using a larger sample. In addition, to further examine whether testing can promote the transfer of learning to novel Chinese characters, we inserted a 2-day delay before the transfer test in Experiment 5. We opted for this procedure because Cho and Powers (2019) implemented their transfer test with a 2-day delay after the study phase.

Design, Participants, Material, and Procedure

Experiment 5 was identical to Experiment 4 except that participants took the transfer test after a 36–48 hr delay. Thirty-six hours after Session 1, participants received an email informing them to finish Session 2 (i.e., transfer test) within the next 12 hr. For exposition purposes, henceforth, we term this a 2-day delay.

For the power analysis, we used the correlation between the forward testing effect and the imagery strategy difference score (ρ = .30) in Experiment 4 as the effect size of interest. Given the novelty of this finding, we aimed to be conservative. Thus, we used 75% of the original correlation (ρ = .225), and a power analysis showed that 137 participants were necessary. Participants were 230 undergraduates from Iowa State University who completed the experiment for course credits. Data of participants who knew Chinese ($N = 17$), failed an attention check ($N = 14$), did the experiment previously ($N = 8$), took notes ($N = 4$), and used a mobile device ($N = 4$) were excluded from analyses. Out of the remaining 183 participants (105 women, 78 men; M_{age} 19.65), 72 did not complete the Session 2 transfer test. However, because our main interest was in the association between strategy use and the forward testing effect, which occurred in Session 1, we decided to include the participants who did not complete Session 2 in the analyses (and it would be exceptionally wasteful to drop data from 72 participants). Furthermore, a missing data analysis showed that there was no systematic differences in target list correct recall between those who completed both sessions and those who completed only Session 1, either in the test condition ($M_{\text{complete}} = .53$ vs. $M_{\text{incomplete}} = .52$), $t(181) = 0.08$, $p = .931$, $d = 0.01$, $B_{01} = 6.08$, or in the restudy condition ($M_{\text{complete}} = .32$ vs. $M_{\text{incomplete}} = .33$), $t(181) = 0.24$, $p = .814$, $d = 0.04$, $B_{01} = 5.95$.

Results

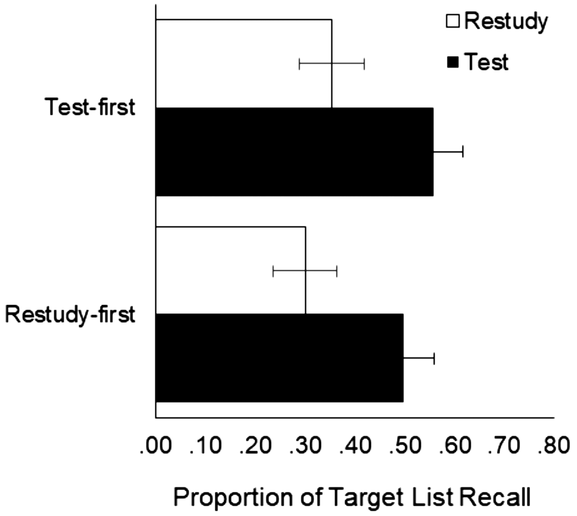
Target List Correct Recall and Transfer Test

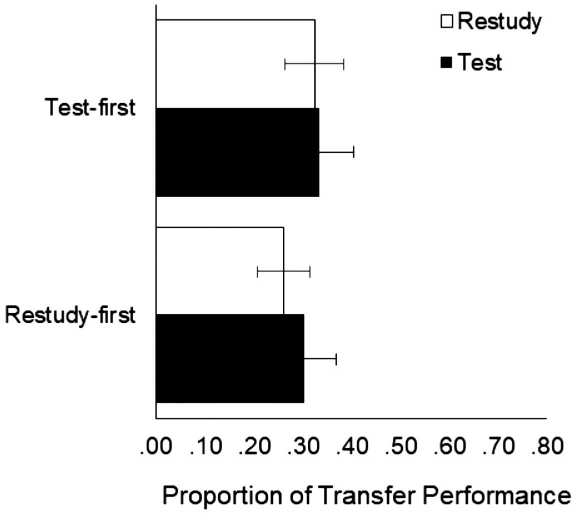
Figure 9 shows the proportion of target list correct recall (top panel) and accuracy on the transfer test (bottom panel). Once again, a forward testing effect was observed, $F(1, 181) = 98.25$, $p < .001$, $d = 0.74$, $B_{10} = 6.772\text{e+}15$, with participants recalling more English meanings of Chinese characters in the test condition ($M = .53$) than in the restudy condition ($M = .33$). Neither the main effect of counterbalancing nor the interaction was significant: Sequence—$F(1, 181) = 2.13$, $p = .146$, $d = 0.22$, $B_{01} = 2.32$; Interaction—$F(1, 181) = 0.00$, $p = .869$, $\eta_p^2 = .00$, $B_{01} = 6.55$. Similar to Experiment 4, there is little evidence for the carryover effect of testing, such that participants performed similarly in the restudy condition regardless of whether it was preceded by the test condition ($M_{\text{test-first}} = .35$) or not ($M_{\text{restudy-first}} = .30$), $t(181) = 1.20$, $p = .234$, $d = 0.18$, $B_{01} = 3.20$.

Extending the results from Experiments 1–4, interim activity did not affect transfer test performance even after a 2-day delay ($M_{\text{test}} = .32$ vs. $M_{\text{restudy}} = .30$), $F(1, 109) = 1.10$, $p = .296$, $d = 0.10$, $B_{01} = 5.70$. Neither the main effect of sequence nor its interaction with interim task was significant, $F\text{s} < 1.50$, $p\text{s} > .224$, $d\text{s} < 0.23$, $B_{01}\text{s} > 2.54$.

Figure 9

Accuracy in Target List Recall Test and Transfer Multiple-Choice Test of Experiment 5





Note. Error bars display descriptive .95 confidence intervals.

Forward Testing Effect Size and Strategies

A 2 (interim activity) × 3 (strategy factor) repeated-measures ANOVA showed that participants used each type of strategy to a different extent, $F(2, 364) = 101.52$, $p < .001$, $\eta_p^2 = .29$, $B_{10} = 2.236\text{e+}38$, with the same rank ordering as in Experiment 4 ($M_{\text{imagery}} = 3.19$, $M_{\text{rehearsal}} = 2.88$, $M_{\text{relational}} = 1.87$, see Table 5). Neither the main effect of interim activity nor interaction was significant, $F\text{s} < 2.05$, $p\text{s} > .153$, $B_{01}\text{s} > 9.87$.

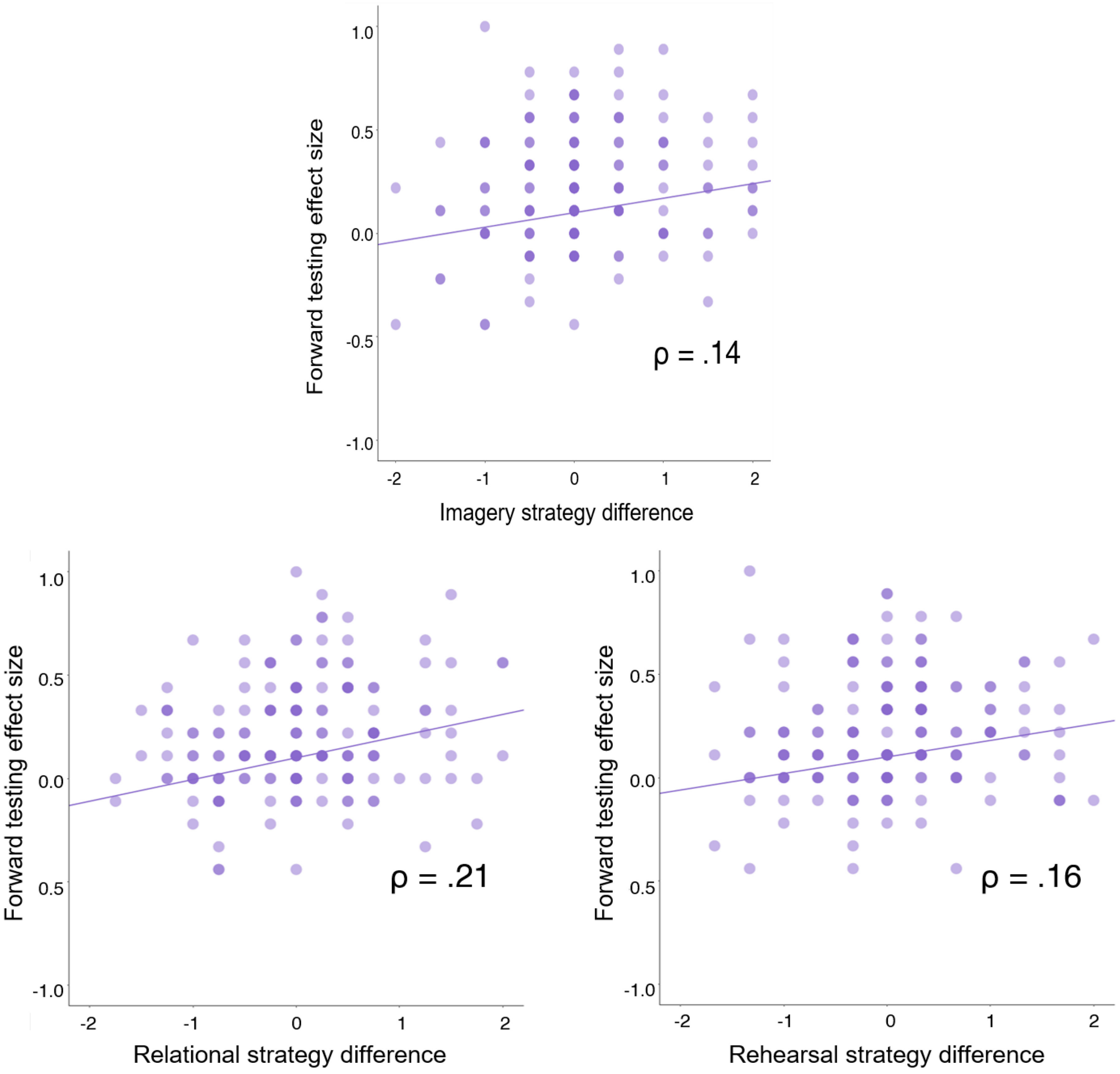
The correlations between the strategy difference scores and the size of the forward testing effect were calculated using the same method as in Experiment 4 (see Figure 10 for scatterplots). A small but positive correlation was observed between each of the strategy difference scores and the magnitude of the forward testing effect: Imagery—$\rho(181) = .14$, $p = .060$, $B_{10} = 0.73$; Relational—$\rho(181) = .21$, $p = .005$, $B_{10} = 6.96$; Rehearsal—$\rho(181) = .16$, $p = .029$, $B_{10} = 0.53$.

To further examine the overall influence of the three strategy difference scores on the forward testing effect, we conducted a multiple regression analysis with the three strategy difference scores as independent variables and the forward testing effect as the dependent variable using the combined data from Experiments 4 and 5.

The overall model explained a moderate amount of variance of the forward testing effect, $r = .29$, $F(3, 255) = 7.44$, $p < .001$, $B_{01} = 148.21$, with $\beta_{\text{imagery}\Delta} = .04$ ($p = .007$), $\beta_{\text{relational}\Delta} = .06$ ($p = .005$), and $\beta_{\text{rehearsal}\Delta} = .03$ ($p = .141$). This result indicates that people who showed a greater forward testing effect reported more strategy change following interim testing (compared with interim restudying). One might interpret this result as providing

Figure 10

Scatterplots Showing the Association Between Strategy Difference Scores and Forward Testing Effect Size in Experiment 5



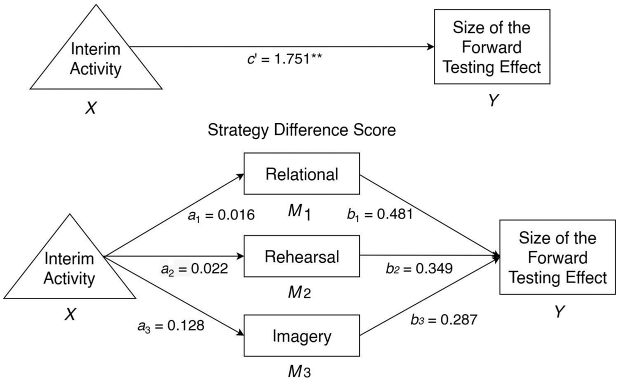
Note. Overlapping data points are shown as darker dots. See the online article for the color version of this figure.

partial support for the strategy-change account. Due to its correlational nature and the lack of overall difference in strategy reports across conditions, caution is necessary when interpreting these data. We will further consider the implications of this finding in the "General Discussion".

To further examine whether strategy change mediates the relationship between interim activity and target list recall, we conducted a mediation analysis. Before the mediation analysis, we checked the validity of strategy measure by correlating each of strategy factors (not the strategy difference scores) with target list recall, and all of them showed positive correlations in both restudy ($\rho_{\text{relational}} = .22$, $\rho_{\text{rehearsal}} = .25$, $\rho_{\text{imagery}} = .30$, $ps < .001$, B_{01}s > 43.31) and test ($\rho_{\text{relational}} = .16$, $\rho_{\text{rehearsal}} = .23$, $\rho_{\text{imagery}} = .21$, $ps < .001$, B_{01}s $>$ 2.39) conditions. We used Montoya and Hayes' (2017) approach for the within-subjects mediation analysis. In mediation analysis, there are three causal paths. In Figure 11, path *c* is the effect of predictor (X) on the outcome variable (Y), path *a* is the effect of predictor (X) on mediator (M), and path *b* is the effect of mediator (M) on outcome variable (Y). The total effect of X on Y is demonstrated by two components, a direct effect c', and an indirect-mediation effect *ab*, which is the product of *a* and *b*. Each of the three strategy factors served as a mediator. As can be seen in Figure 11, despite a powerful direct relationship between interim activity in Target list correct recall ($c' = 1.751$, CI [1.397, 2.105]), it was not mediated by any of the three strategy factors: Relational ($ab = 0.008$, CI [−0.046, 0.072]), Rehearsal ($ab = 0.008$, CI [−0.050, 0.063]), and Imagery ($ab = 0.037$, CI [−0.010, 0.118]). Total indirect effect was 0.052, CI [−0.048, 0.163].

Figure 11

A Mediation Model With Strategy Factors as Mediators in Path Analysis Form



Note. Size of the Forward Testing Effect was the difference in target list recall between the restudy and test conditions.
$^{**}p < .001$.

General Discussion

The main goal of the present study was to examine the *strategy-change* account of the forward testing effect. Specifically, we examined whether interim testing changes the strategies learners use to *encode* new information. Across five experiments, interim testing (relative to restudying) did not affect encoding strategies, and this null effect was found (a) regardless of whether participants reported strategy use after every study list or after only the criterial list, (b) whether the interim task was manipulated between or within-subjects, and (c) strategy use did not mediate the relationship between interim activity and recall. Most importantly, these null effects occurred in spite of a robust forward testing effect being observed in every experiment. Lastly, we found no evidence that interim testing promoted the recognition of novel Chinese materials compared to restudy in the transfer test.

Strategy-Change Account

If we take at face value the null effect of testing on participants' reported encoding strategies, a natural interpretation is that interim testing did not change encoding strategies for new materials. This finding thus places an important constraint on the strategy-change account, such that interim testing might affect learners' subsequent retrieval strategies (Chan et al., 2020; Chan, Manley, et al., 2018; Dang et al., 2021), but leaving the encoding strategies unchanged. Consistent with this idea, a recent study by Ahn and Chan (2022) suggested that strategy change may occur during retrieval rather than encoding. Their participants studied several category word lists arranged with the same category words blocked together (i.e., blocked) or with a mixture of different category words per list (i.e., intermixed). The researchers hypothesized that if testing facilitates new learning because it promotes relational processing during the encoding stage, then the forward testing benefit should be diminished in the blocked condition compared to the intermixed condition. Their rationale was that presenting the same category words consecutively would encourage relational encoding even in the absence of testing, which should in turn reduce the performance gap between test and restudy. However, Ahn and Chan (2022) found that the forward testing effect was unaffected by category arrangements, which suggests that testing may not affect how participants encode subsequent materials.

Although we did not find a strategy difference at the condition level, strategy change was moderately correlated with the forward testing effect at an individual level. There are several possibilities for these seemingly discrepant findings, although we believe that the contradiction is more apparent than real. First, despite the fact that we found a forward testing effect (e.g., $d = 0.74$ in Experiment 5), not all participants experienced the benefit of forward testing. In Experiment 5, 69% of the participants showed a forward testing effect, but 31% experienced either no benefit or a negative effect. The positive correlation between the forward testing effect and strategy change was driven by a small increase in strategy use for people who experienced a positive forward testing effect (+0.13 across the three strategies) and a small decrease in strategy use for people who did not show a forward testing effect (−0.09). Consequently, there was a correlation but no overall change in strategies based on testing at the group level.

Based on these results, we suspect that the individual differences in strategy use and their association with the forward testing effect were driven by a third variable. For example, students who tend to report more change in strategy use (based on testing) might be students who were naturally motivated to learn the material in the experiment, and these individuals might be particularly likely to demonstrate a forward testing effect (for similar demonstrations of individual differences in the retrieval practice effect literature, see Fellman et al., 2020; Minear et al., 2018). Moreover, as stated earlier, testing might instigate a shift in retrieval strategy, which leads to a forward testing effect, and changes in retrieval strategies might in turn cause *some participants* to change their encoding strategies, and this second-order change is then registered as the weak positive correlation between the encoding strategy difference scores and the forward testing effect observed in Experiment 5 (ρs between .14 and .21).

Before moving on, two existing findings must be considered in the context of our finding that "interim testing does not change participants' encoding strategy." First, interim testing has been shown to alter participants' expectations about *whether* they will be tested in the imminent future. Specifically, Weinstein et al. (2014) and Chan et al. (2020) reported that participants who had completed more interim tests were more likely to expect another test in the future, even if they were told that whether they would be tested on each trial was determined randomly. Therefore, participants' estimated test likelihood should remain at 50% across trials, but in reality, participants who are in the interim test condition increased their test expectancy across trials, whereas those in the interim restudy condition reported the opposite. Second, Finley and Benjamin (2012) showed that learners shifted their encoding strategies when they perceived a change in ongoing task demands. In Finley and Benjamin's (2012) experiments, participants were led to believe that they would later complete a recall or a recognition test, and participants' strategies changed based on the expected test format.

Taken together, if altering learners' expectations about the *type* of test they will receive changes their encoding strategies, and if interim testing alters learners' expectations about the *likelihood* that they would be tested, then a logical extension is that interim testing should affect learners' encoding strategies. The results of our five experiments, if proven generalizable beyond the present materials, suggest that learners' expectancy about *test likelihood* and *test type* are fundamentally different. Increasing test likelihood might not change people's encoding strategies because there are no qualitative changes to the task demand, whereas changing the type of test (between recall and recognition) does, and a change in perceived task demand might be necessary to elicit a qualitative shift in encoding strategies. Indeed, these results, together with the finding that interim testing can increase participants' self-regulated study time (Davis & Chan, 2022; Yang et al., 2017), suggest that interim retrieval might affect encoding strategies quantitatively (e.g., they might effort greater effort in general) but not qualitatively (e.g., they might not shift to a new encoding strategy), although this conclusion is clearly preliminary and awaits more research to be verified.

Limitations, Broader Implications, and Future Directions

We would like to address why the baseline of strategy use was different between the test and restudy conditions in the combined analysis across E1–E3. We suspect that it might be related to the timing of strategy reporting. Participants reported their strategies *after* they restudied the previous list or took an interim test. This was because we were concerned about participants in the restudy condition seeing the strategies *before* receiving the restudy opportunity, which could then lead them to implement those strategies during the restudy session. Specifically, the following question was provided during the strategy survey, "How much have you used the following strategies while you studied List 1?" The phrase "while you studied List 1" meant only the initial study trial, not the interim restudy or test trial. However, it is possible that when the restudied participants were doing the survey, they had considered the restudy trial to answer the questionnaire, whereas the tested participants would be more likely to consider only the initial study trial. Thus, in future studies, it might be necessary to vary when participants are asked to report their strategies. For example, instead of using a retroactive strategy report, Dunlosky and Hertzog (2001) asked participants what strategy was used for each item among the five strategies. Similarly, a follow-up study could use an item-based strategy measure instead of a list-based measure.

Finally, one could argue that we found no difference in encoding strategies because participants might not be able to recognize changes to their strategies, rather than because testing did not influence their encoding strategies. We caution against claims of this nature, because they make the strategy-change account difficult to falsify. An alternative, more plausible argument might be that our paired-associate materials may not be suitable to examine the strategy-change account, given that studies that led to the proposal of that account were based on associative word lists (Chan et al., 2020; Chan, Manley, et al., 2018; Dang et al., 2021). In those studies, participants benefited from associating different items with one another and that each recalled item could serve as a retrieval cue for other items. However, this type of relational reasoning might not be suitable for learning Chinese–English paired associates. Because participants were always provided with the Chinese character as a cue, it was sufficient to learn the association *within each* pair, without making associations across pairs, especially because participants were not told about the existence of radicals.

In fact, the imagery strategy might have been more useful than the relational strategy, given the characteristics of our material. Chinese is a logographic writing system because the shapes of many characters represent the actual meaning of the words. This feature is especially distinctive among characters created based on radicals. For example, 火 means fire, and its origin was based on the shape of fire around pieces of logs. Although participants were not told about the representational characteristics of the Chinese characters, some might have noticed the correspondence between the shape of the Chinese characters and their meanings. Indeed, participants reported using imagery strategies most frequently (see Figure 5 for data from Experiments 1–3 and Table 5 for data from Experiments 4–5). Moreover, when we asked participants about their learning strategies with an open-ended question during the pilot study, some participants reported associating the shapes of characters with their English meanings. Despite the above arguments, we had considered the present materials conducive to encoding strategies that use relational, and inter-item associations before the start of the experiments, so any reasoning to the contrary was done post hoc.

Future studies should examine whether our results would generalize beyond paired-associate learning. At the very least, our data highlight the importance of broadening the materials used in the forward testing effect literature, which is essential to developing well-rounded theories (Hintzman, 2011). Although some readers might find the conclusion that "interim testing does not cause a qualitative change to subsequent encoding strategy for Chinese–English paired associates learning" overly constrained, it is important to consider that this finding occurred while we repeatedly observed a robust forward testing effect. Consequently, our results have the potential to rule out encoding strategy change as an explanation for the forward testing effect, but it is also important to remember the context in which this finding emerged. Jenkins (1979; see also Roediger, 2008) proposed a tetrahedral model of human memory experiments, in which he argued that all memory phenomena must be considered in the context of their discovery, including subject variables such as age and cognitive abilities, orienting task variables such as instructions, criterial task variables such as test types (e.g., recall, recognition), and material variables such as the sensory mode (e.g., verbal, visual). In our opinion, it is the material variable that placed the greatest constraints on the generality of our critical finding.

Students learn different types of material daily, and different materials have different demands on learning. Take for example the finding that interim testing can reduce the frequency of mind wandering when participants watched a statistic lecture (Szpunar et al., 2013); a similar study found that interim testing only altered the *nature* of mind wandering (e.g., task-related vs. task-unrelated), but not the *frequency* of mind wandering, when participants watched a public health lecture, which was presumably more engaging than the statistic lecture (Jing et al., 2016). Indeed, the literature is replete with examples that show the elimination or reversal of a seemingly robust and powerful effect based on a change in the study material (e.g., Huff et al., 2016; LaPaglia & Chan, 2013, 2019; Pereverseff et al., 2020). An important goal, in our opinion, is to discover *how* different materials can affect the processes involved in learning. In sum, further research with different materials is necessary to ascertain the generality of our conclusion on the encoding strategy-change account of the forward testing effect.

Transfer of Learning

Across five experiments, we examined whether interim testing could enhance the transfer of learning. We included this research question because testing has been known to promote relational processing, and finding relations (i.e., radicals) among different characters can promote participants' ability to infer the meaning of novel Chinese characters (Cho & Powers, 2019). However, we did not find a benefit of testing on the transfer test, even when we included feedback (Experiment 2), varied the frequency of strategy reporting (Experiment 3), and inserted a 2-day delay (Experiment 5). These null effects contrast with the results shown by Cho and Powers.

When designing our experiments, we discovered several issues with the materials in Cho and Powers (2019) that could cloud interpretations of their transfer test results (see Figure S2 in the online supplemental material on OSF for their material), which prompted us to create new materials. The first issue is that some of their Chinese characters had very similar meanings or that the English words had substantial orthographic overlap. For example, for characters that used the 女—woman radical (i.e., the first column in Figure S2 in the online supplemental material), the first item (好) is beautiful, and the second item (姁) is beauty. A similar concern applies to the 目—eyes radical, in which three items (i.e., gaze, see, stare) have extremely similar meanings. This issue was exacerbated for items with the 氵 —water radical, with two items (i.e., 氾 and 汛) being assigned the identical English translation of "flood," and another character (汎) given the translation of "float." The overlap in meaning and orthography of the English words is problematic because their transfer test might be assessing retention of the verbatim item rather than transfer. Specifically, if participants had studied the character for "beautiful" during the encoding phase and then received the character for "beauty" in the transfer test, they might be biased towards selecting "beauty" as the answer not because they correctly inferred the meaning of the radical, but because they mistakenly recognized "beauty" as a studied item (this issue would be exacerbated if "氾—flood" was studied and "汛—flood" appeared on the transfer test).

The second issue is that some of Cho and Powers' (2019) Chinese characters look extremely similar. For example, for the 氵 —water radical, the second (汎) and third (汛) characters were nearly indistinguishable. The same applies to the fourth (灯) and seventh (灴) characters with the 火—fire radical. Furthermore, for the 木—tree radical, the second to fifth items (i.e., 材, 林, 杖, 材) have extremely similar appearances. In fact, the second and fifth items were identical but were assigned different English meanings. Accordingly, instead of recognizing a radical and inferring the meaning of the character—thereby showing transfer, participants in Cho and Powers (2019) might have recognized the entire character because of their similarity to the studied ones—thereby showing recognition memory.

Together, the high perceptual similarities of some of the Chinese characters and conceptual/orthographic similarities of the English words might have turned what was intended to be a transfer test into a recognition test, at least for some of the trials. Thus, the higher performance in the transfer test of the test condition compared to the restudy condition in Cho and Powers (2019) could have reflected, at least in part, enhanced retention by testing instead of transfer of learning.

Finally, although some evidence has shown that testing can help students apply learned knowledge to new contexts (see Carpenter, 2012), the results are not universal and dependent on various factors (Pan & Rickard, 2018). Some studies reported an advantage of testing on transfer (Pan & Rickard, 2017; Rohrer et al., 2010), whereas others reported no benefits (Brunyé et al., 2020; Tran et al., 2015). Critically, a recent meta-analysis found that several moderator variables, such as test format and material type, can determine the extent to which the benefits of testing would transfer to new environments (Pan & Rickard, 2018). Regarding our study, the moderator of interest is the number of training phase repetitions. Specifically, Pan and Rickard (2018) indicated that studies with more encoding repetitions produced a larger benefit of testing on the transfer of learning. In our study, the tested participants studied three different lists twice (w/o feedback) or three times (w/feedback), whereas Cho and Powers' (2019) participants studied a single list six times. This difference in encoding repetitions might explain why testing failed to improve the transfer of learning in our experiments relative to restudying. Additionally, Pan and Rickard (2018) pointed out that the study phase repetition effect was particularly pronounced when the transfer test comprised inference questions that required learners to apply knowledge to new contexts, as our transfer test did. In sum, our data showed that testing-based transfer effects for Chinese character learning might not be expected in all situations.

Conclusion

In the present experiments, we tested the strategy-change account for the forward testing effect. Across five experiments, we consistently found no effect of testing on participants' encoding strategy reports when they attempted to learn written Chinese characters—despite observing a robust forward testing effect in every comparison. Our study also showed that interim testing did not promote the transfer of learning to novel Chinese characters relative to restudying.

Finally, this study extended the forward testing effect to foreign language learning of Chinese characters. Although the benefit of interim testing on foreign language learning has been known, it had only been observed with alphabetic languages (Cho et al., 2017; Yang et al., 2017). Chinese characters are qualitatively different because of their logographic and visuospatial nature. Overall, our data suggest that interim testing can be used to enhance the learning of written Chinese—the most widely used language in the world—but further work is necessary to ascertain how interim testing produced its benefits.

References

Ahn, D., & Chan, J. C. K. (2022). Does testing enhance new learning because it insulates against proactive interference? *Memory & Cognition*, *50*(8), 1664–1682. https://doi.org/10.3758/s13421-022-01273-7

Ahn, D., & Chan, J. C. K. (2023). *How does testing affect future learning strategies?* https://osf.io/uh8wr/

Brunyé, T. T., Smith, A. M., Hendel, D., Gardony, A. L., Martis, S. B., & Taylor, H. A. (2020). Retrieval practice enhances near but not far transfer of spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(1), 24–45. https://doi.org/10.1037/xlm0000710

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), 290–298. https://doi.org/10.1037/a0031026

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Chan, J. C. K., & McDermott, K. B. (2007). The effects of frontal lobe functioning and age on veridical and false recall. *Psychonomic Bulletin & Review*, *14*(4), 606–611. https://doi.org/10.3758/BF03196809

Chan, J. C. K., Manley, K. D., & Ahn, D. (2020). Does retrieval potentiate new learning when retrieval stops but new learning continues? *Journal of Memory and Language*, *115*, Article 104150. https://doi.org/10.1016/j.jml.2020.104150

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language*, *102*, 83–96. https://doi.org/10.1016/j.jml.2018.05.007

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology*, *70*(7), 1211–1235. https://doi.org/10.1080/17470218.2016.1175485

Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory and Cognition*, *8*(2), 166–177. https://doi.org/10.1016/j.jarmac.2019.01.003

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), Article 7. https://doi.org/10.7275/jyj1-4868

Dang, X., Yang, C., & Chen, Y. (2021). Age difference in the forward testing effect: The roles of strategy change and release from proactive interference. *Cognitive Development*, *59*, Article 101079. https://doi.org/10.1016/j.cogdev.2021.101079

Davis, S. D., Chan, J. C., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 434–441. https://doi.org/10.1016/j.jarmac.2017.07.002

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1741–1754. https://doi.org/10.1037/xlm0000126

Davis, S. D., & Chan, J. C. K. (2022). *Effortful tests and deep metacognitive reflection enhance future learning* [Manuscript in preparation]. Department of Psychology, University of North Florida.

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology*, *10*, Article 2755. https://doi.org/10.3389/fpsyg.2019.02755

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational reasoning and its manifestations in the educational context: A systematic review of the literature. *Educational Psychology Review*, *25*(3), 391–427. https://doi.org/10.1007/s10648-013-9224-4

Dunlosky, J., & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition*, *29*(2), 247–253. https://doi.org/10.3758/BF03194918

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*, *6*, Article 1054. https://doi.org/10.3389/fpsyg.2015.01054

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fellman, D., Lincke, A., & Jonsson, B. (2020). Do individual differences in cognition and personality predict retrieval practice activities on MOOCs. *Frontiers in Psychology*, *11*, Article 2076. https://doi.org/10.3389/fpsyg.2020.02076

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 632–652. https://doi.org/10.1037/a0026215

Finn, B., Thomas, R., & Rawson, K. A. (2018). Learning more from feedback: Elaborating feedback with examples enhances concept learning. *Learning and Instruction*, *54*, 104–113. https://doi.org/10.1016/j.learninstruc.2017.08.007

Flaherty, M. (2003). Sign language and Chinese characters on visual-spatial memory: A literature review. *Perceptual and Motor Skills*, *97*(3), 797–802. https://doi.org/10.2466/pms.2003.97.3.797

Glisky, E. L., Polster, M. R., & Routhieaux, B. C. (1995). Double dissociation between item and source memory. *Neuropsychology*, *9*(2), 229–235. https://doi.org/10.1037/0894-4105.9.2.229

Ha, H., & Lee, H. S. (2019). Effect of interim testing on learners' metacognitive judgments, study time, and learning performance in concept and category learning. *The Korean Journal of Educational Psychology*, *33*(2), 125–152. https://doi.org/10.17286/KJEP.2019.33.2.01

Halamish, V. (2018). Pre-service and in-service teachers' metacognitive knowledge of learning strategies. *Frontiers in Psychology*, *9*, Article 2152. https://doi.org/10.3389/fpsyg.2018.02152

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205. https://doi.org/10.1177/1094428104263675

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. *Psychology of Learning and Motivation—Advances in Research and Theory*, *45*, 215–251. https://doi.org/10.1016/S0079-7421(03)45006-8

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "Coordinates of Truth." *Perspectives on Psychological Science*, *6*(3), 253–271. https://doi.org/10.1177/1745691611406924

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

Huff, M. J., Weinsheimer, C. C., & Bodner, G. E. (2016). Reducing the misinformation effect through initial testing: Take two tests and recall me in the morning. *Applied Cognitive Psychology*, *30*(1), 61–69. https://doi.org/10.1002/acp.3167

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(6), 1441–1451. https://doi.org/10.1037/a0020636

JASP Team. (2020). *JASP* (Version 0.14) [Computer software].

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (Vol. 1, pp. 429–446). Erlbaum.

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interim testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*(3), 305–318. https://doi.org/10.1037/xap0000087

Kang, S. H. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition*, *38*(8), 1009–1017. https://doi.org/10.3758/MC.38.8.1009

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

LaPaglia, J. A., & Chan, J. C. (2013). Testing increases suggestibility for narrative-based misinformation but reduces suggestibility for question-based misinformation. *Behavioral Sciences & the Law*, *31*(5), 593–606. https://doi.org/10.1002/bsl.2090

LaPaglia, J. A., & Chan, J. C. K. (2019). Telling a good story: The effects of memory retrieval and context processing on eyewitness suggestibility. *PLoS One*, *14*(2), Article e0212592. https://doi.org/10.1371/journal.pone.0212592

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research and Evaluation*, *12*, Article 2. https://doi.org/10.7275/WJNC-NM63

Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 1, pp. 327–372). Academic Press.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. https://doi.org/10.3758/s13421-010-0035-2

McCabe, J. A. (2018). What learning strategies do academic support centers recommend to undergraduates? *Journal of Applied Research in Memory and Cognition*, *7*(1), 143–153. https://doi.org/10.1016/j.jarmac.2017.10.002

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*(1), 6–27. https://doi.org/10.1037/met0000086

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied*, *23*(3), 278–292. https://doi.org/10.1037/xap0000124

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pereverseff, R. S., Bodner, G. E., & Huff, M. J. (2020). Protective effects of testing across misinformation formats in the household scene paradigm. *Quarterly Journal of Experimental Psychology*, *73*(3), 425–441. https://doi.org/10.1177/1747021819881948

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 737–746. https://doi.org/10.1037/a0026166

Rawson, K. A., & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language*, *105*, 141–152. https://doi.org/10.1016/j.jml.2019.01.002

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5*, Article 589965. https://doi.org/10.3389/feduc.2020.589965

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*(1), 225–254. https://doi.org/10.1146/annurev.psych.57.102904.190139

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239. https://doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, *28*(12), 1629–1646. https://doi.org/10.1177/014616702237645

Shen, H. H. (2010). Imagery and verbal coding approaches in Chinese vocabulary instruction. *Language Teaching Research*, *14*(4), 485–499. https://doi.org/10.1177/1362168810375370

Shuell, T. J. (1969). Clustering and organization in free recall. *Psychological Bulletin*, *72*(5), 353–374. https://doi.org/10.1037/h0028141

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interim memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392–1399. https://doi.org/10.1037/a0013082

Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*, *22*(1), 135–140. https://doi.org/10.3758/s13423-014-0646-x

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1039–1048. https://doi.org/10.1037/a0036164

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140–1147. https://doi.org/10.3758/s13423-011-0140-7

Yang, C., Chew, S. J., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology*, *111*(5), 809–826. https://doi.org/10.1037/edu0000320

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 263–277. https://doi.org/10.1037/xap0000122

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ Science of Learning*, *3*(1), Article 8. https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2022). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(8), 1127–1143 https://doi.org/10.1037/xlm0001021

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995–1008. https://doi.org/10.3758/MC.38.8.995

(*Appendices follow*)

Appendix A

The Full Material Set Used in Experiments 1–3

Radical	Eye (目)	Hand/arm (扌)	Speech (言)	Water (氵)
Chinese–English	眠—sleep	抄—copy	說—speak	汁—juice
	瞪—glare	扛—carry	講—explain	污—dirty
	瞎—blind	扣—hook	詩—poetry	波—wave
	瞰—overlook	拉—drag	諭—proclaim	渴—thirsty
	盹—nap	捏—pinch	諮—consult	沉—sink
	眩—dizzy	扔—throw	讀—read	湖—lake
	睫—eyelash	指—finger	訓—instruct	海—ocean
	瞄—glance	接—receive	誹—slander	池—pool
	眼—eye	打—hit	記—record	沁—soak
	眨—wink	托—hold	話—word	江—river
	瞳—pupil	挑—lift	議—discuss	清—clear
	眯—squint	抛—toss	計—count	汤—soup

Appendix B

Added Chinese Characters in Experiments 4–5

Radical	Fire (火)	Tree (木)
Chinese–English	烙—bake	板—board
	炫—bright	椅—chair
	烤—barbeque	植—plant
	燭—candle	柏—cypress
	灼—burn	松—pine
	炸—frying	楓—maple
	熔—melt	林—forest
	炬—torch	棉—cotton
	烟—smoke	核—seed
	炆—stew	杆—stick
	炉—stove	杪—twig
	灯—lamp	柳—willow

Received March 13, 2022
Revision received January 5, 2023
Accepted January 8, 2023 ■