Vision Research 227 (2025) 108543

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Brief memory reactivation may not improve visual perception

Jun-Ping Zhu, Jun-Yun Zhang *

School of Psychological and Cognitive Sciences, and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Perceptual learning
Brief reactivation
Reconsolidation
Test–retest effect
Plasticity

ABSTRACT

Visual perceptual learning often requires a substantial number of trials to observe significant learning effects. Previously Amar-Halpert et al. (2017) have shown that brief reactivation (5 trials/day) is sufficient to improve the performance of the texture discrimination task (TDT), yielding comparable improvements to those achieved through full practice (252 trials/day). The finding is important since it would refine our understanding of learning mechanisms and applications. In the current study, we attempted to replicate these experiments using a larger number of observers and an improved experimental design. Using between-group comparison, we did find significant improvements in the reactivation group and the full-practice group as Amar-Halpert et al. (2017) showed. However, these improvements were comparable to those of the no-reactivation group with no exposure to the TDT task over the same period. Importantly, our within-group comparison showed that both the reactivation and no-reactivation groups exhibited additional significant improvements after further practicing the TDT task for an additional three days, demonstrating that the full-practice effect was significantly superior to the effects of brief memory reactivation or simple test–retest. Besides, when refining the constant stimuli method with fewer stimulus levels and more trials per level, we still observed comparable improvements brought by the reactivation and no-reactivation groups. Therefore, our results suggested that brief memory reactivation may not significantly contribute to the improvement of perceptual learning, and traditional perceptual training could still be a necessary and effective approach for substantial improvements.

1. Introduction

Visual perceptual learning refers to performance improvement on visual tasks through training (Lu & Dosher, 2022; Sagi, 2011; Watanabe & Sasaki, 2015). It has shown powerful real-world applications in improving the sensory performance of healthy individuals and rehabilitating clinical populations with various types of vision loss, such as amblyopia (Levi & Polat, 1996; Zhang et al., 2014), macular degeneration (Plank et al., 2014), and cortical blindness (Das et al., 2014; Herpich et al., 2019). However, a significant limitation of perceptual learning's practical applications is that it usually requires a long period of extensive practice for adequate performance enhancement (Jeter et al., 2010; Li et al., 2008). For example, a healthy adult's performance usually reaches a plateau after practicing for 5–10 daily sessions in a texture discrimination task (Karni & Sagi, 1991; Wang et al., 2013). Additionally, patients with vision impairments like cortical blindness, a form of vision loss caused by primary visual cortex damage, require months of daily practice to restore normal performance on a motion integration task in the blind field, making the training difficult to attain and sustain (Das et al., 2014; Huxlin et al., 2009).

Amar-Halpert et al. (2017) previously reported that brief reactivation of encoded visual memories was sufficient to improve visual perception. This study is grounded in the reactivation-reconsolidation framework, which claims that memories remain dynamic even after initial consolidation. Reactivation of memory through exposure to salient training stimuli can induce destabilization, triggering a reconsolidation process during which memories become susceptible to modification and can be enhanced or impaired (Lee et al., 2017). In the study of Amar-Halpert et al. (2017), observers in the reactivation group performed a texture discrimination task with 252 trials on day 1 to encode and consolidate memory. Subsequently, memory reactivation was conducted with only 5 trials for three consecutive days. The results showed that brief reactivations were sufficient to improve memory, as evidenced by the significant learning outcomes observed in the post-test on day 5, which were comparable to those of the full-practice group that performed the task with 252 trials per day. Besides, the reactivation group outperformed a no-progress control condition measuring two-session learning without memory reactivations (day 1 to day 2 in the

* Corresponding author.
E-mail address: zhangjunyun@pku.edu.cn (J.-Y. Zhang).

https://doi.org/10.1016/j.visres.2025.108543
Received 1 November 2023; Received in revised form 7 January 2025; Accepted 7 January 2025

full-practice group). They also established far-threshold reactivation and no-reactivation groups with pre-test and post-test spaced nine days apart and consistently found that the former outperformed the latter.

Several investigations using task interference have suggested the reactivation and reconsolidation process in perceptual learning (Bang et al., 2018; Dayan et al., 2016; Herszage & Censor, 2017; Huang et al., 2023; Shibata et al., 2017; Walker et al., 2003). For example, Bang et al. (2018) demonstrated that reconsolidation did occur after reactivation in visual perceptual learning. They asked observers to practice the detection of two orientations in a blockwise manner and found that the timing between the blocks (either short: 0 h or long: 3.5 h) led to either interference and performance decline, or no interference and performance improvement. These results suggested that reconsolidation occurred during the 3.5-hour interval following the reactivation of the trained orientation detection task. To the best of our knowledge, Amar-Halpert et al. (2017) is the first study reporting the reconsolidation phenomenon in the domain of visual perceptual learning. The finding of Amar-Halpert et al. (2017) is significant as it has challenged the fundamental principle of procedural learning theory, which states that practice makes perfect. Instead, their finding suggests a more efficient mechanism underlying improvement in visual perception, which has far-reaching clinical applications. The same research team has also generalized these results to other fields of procedural learning, including motor skill learning (Herszage et al., 2021), numeric domain (Schrift et al., 2022), and clinical populations like individuals with autism (Klorfeld-Auslender et al., 2022). This generalization across different memory fields and populations has great theoretical significance for the reconsolidation theory itself, given that the reconsolidation phenomenon has predominantly been based on Pavlovian fear-conditioning models in rodents since its initial discovery (Lee et al., 2017; Misanin et al., 1968; Nader et al., 2000; Schneider & Sherman, 1968).

More recently, Chen and de Beeck (2021) have investigated to what extent the similar effects of reactivation as Amar-Halpert et al. (2017) have shown in the texture discrimination task could be found in a paradigm that fo *rocedure.* A standard trial of the texture discrimination task in the current study was nearly identical to that of Amar-Halpert et al. (2017). Each trial began with a 400 ms presentation of a fixation cross, followed by a 500 ms blank screen display. Then a target frame was briefly presented for 13.3 ms, followed, at various stimulus onset asynchronies (SOAs, measured from the onset of the target to the onset of the mask), by a 100 ms patterned mask. After the mask, the screen went blank until the observer made a response. The observers were asked to make two responses: first to report the foveal letter (T or L), and then to report the orientation of the target configuration (horizontal or vertical). Immediate auditory feedback was provided only for the incorrect foveal letter identification. There was a 250 ms inter-trial interval. The average accuracy of the foveal letter identification task is approximately 95 %, indicating effective foveal fixation.

The study employed the constant stimuli method, in which several predetermined stimulus levels were used and each level consisted of a fixed number of trials. The stimulus levels were stimulus onset asynchronies (SOAs, measured from the onset of the target to the onset of the mask), which were multiples of 13.3 ms frame duration, ranging from 3 frames to 26 frames (40, 67, 80, 107, 120, 147, 160, 187, 200, 227, 240, 267, 307, 347 ms). In the main experiment, the 14 SOA levels were randomized across all trials, with 18 trials per SOA level. Thus, a standard block comprised a total of 252 trials. In the control experiment, the constant stimuli method was refined by reducing the 14 SOA levels to 6 or 7 levels, with each level containing 42 or 36 trials, maintaining the total number of trials at 252.

Before the formal experiment, a pretraining block consisting of 10 trials at a 347 ms SOA was administered repeatedly until observers reached a 90 % accuracy rate. Observers who completed ten pretraining blocks without attaining the required accuracy were excluded from the study. Those who successfully passed the pretraining phase then progressed to a familiarization block comprising 14 trials, each reactivation group and half of the observers in the new no-reactivation group experienced 6 SOA levels (40, 67, 107, 160, 240, 347 ms), with 42 trials per SOA level. The other half of observers in the new no-reactivation group underwent 7 SOA levels (40, 80, 120, 160, 200, 267, 347 ms), with 36 trials per SOA level. Reactivation trials were set individually at one of the 6 ~ 7 SOAs that was closest to each observer's pre-test threshold (See Table S1 for a specific value of each observer). Using the modified constant stimuli method, the new reactivation and no-reactivation groups only experienced Phase 1.

2.4. ata fitting and statistical analysis

To evaluate the impact of different data fitting methods on the results, we employed two fitting methods to fit psychometric curves for threshold estimates.

The first fitting method is consistent with the approach described by Amar-Halpert et al. (2017). The threshold was calculated for each standard block (252 trials) using the Weibull fit for the psychometric curve, with slope β and finger error (mistyping) parameter 1 – p, yielding the function:

$$P(t) = p\{1-\frac{1}{2}\exp[-\left(\frac{t}{T}\right)^{\beta}]\}+\frac{1-p}{2}=\frac{1}{2}\{1+p[1-\exp[-\left(\frac{t}{T}\right)^{\beta}]]\}$$

where P(t) is the measured probability of correct response; t represents the SOA levels; finger error parameter, which takes stimulus-independent errors (e.g., attention lapses, response-key confusion) into account, is a free parameter within a range ($0 < 1 - p < 1$); T is the estimated threshold for each curve, defined as the SOA for which 81.6 % of responses were correct when $p = 1$. Weibull fit was computed using a maximum likelihood method, assuming a binomial process (Wichmann & Hill, 2001).

The second method is using the psignifit4 software package (see http://bootstrap-software.org/psignifit/) to fit psychometric curves and estimate thresholds (Schütt et al., 2016). Here the psychometric function modeling was extended from the standard binomial to a beta-binomial model to enable accurate Bayesian estimation of psychometric functions even for overdispersed data. Psychometric curves for each observer were generated by fitting the data with a Weibull function. Using the chosen sigmoid family S(x; m, w), the psychometric function ψ is defined with two additional parameters λ and γ for the upper and lower asymptote, scaling the sigmoid function:

$$\psi(x;\, m,\, w,\, \lambda,\, \gamma) = \gamma + (1-\lambda-\gamma)S(x;\, m,\, w)$$

where threshold m is the stimulus level at which 81.6 % of responses were correct when $\lambda = 0$ (to maintain consistency with the first fitting method); w represents the width (difference between the stimulus levels for which the unscaled function reaches 0.05 and 0.95 respectively); λ represents the lapse rate (the difference between the upper asymptote and 1); γ represents the guess rate (the difference between the lower asymptote and 0). γ is fixed at 0.5 and the lapse rate λ is free.

For both fitting methods, to evaluate how well the psychometric curves capture the empirical data of each individual, we assess goodness-of-fit by calculating deviance which is recommended for binomial data (Wichmann & Hill, 2001):

$$= 2\sum_{i=1}\left\{n_i y_i \log\left(\frac{y_i}{p_i}\right) + n_i(1-y_i)\log\left(\frac{1-y_i}{1-p_i}\right)\right\}$$

where denotes the number of SOA levels, n_i: the number of trials in SOA level i,y_i: the observer's response accuracy in SOA level i, p_i: the response accuracy predicted by the fitted model.

For correct models, deviance for binomial data was asymptotically distributed as χ^2_k, where denoted the number of SOA levels and a χ^2 probability of < 0.05 is considered to indicate a poor fit of the model (Hietanen et al., 2022; Lasagna et al., 2020; Wichmann & Hill, 2001). In our data, both fitting methods indicated the same observer (R23) with poor goodness-of-fit on day 1 (see Figs. S1 for detailed information). As the statistical analyses produced similar results regardless of whether this observer was included or not in the analyses, we decided to keep this observer in the threshold analyses.

The learning effect was evaluated by the improvement in thresholds. Individual threshold improvement from the pre-test to the post-test was calculated as 100 %× ($\text{Threshold}_{\text{pretest}}$ − $\text{Threshold}_{\text{posttest}}$)/$\text{Threshold}_{\text{pretest}}$ and then averaged across observers to obtain the mean percent improvement (MPI). To evaluate the progress following Phase2 training, individual further threshold improvement from post-test to post-test2 was calculated as 100 % × ($\text{Threshold}_{\text{post-test}}$ − $\text{Threshold}_{\text{post-test2}}$)/$\text{Threshold}_{\text{post-test}}$, and individual total threshold improvement from pre-test to post-test2 was calculated as 100 % ×($\text{Threshold}_{\text{pretest}}$ − $\text{Threshold}_{\text{posttest2}}$)/$\text{Threshold}_{\text{pretest}}$.

All analyses were conducted using open-source JASP software version 0.17.2.1 (Wagenmakers et al., 2018). Improvements in SOA thresholds were compared against the value of 0 through a one-sample *t*-test. Within-group comparisons were performed using a paired samples *t*-test or one-way repeated measures analysis of variance (ANOVA). Comparisons between the two groups were conducted using both classical and Bayesian independent samples t-tests.

3. Results

3.1. ain experiment_ hase1: Comparing learning effects of reacti ation ith no reacti ation and full practice

In Phase 1 covering five days, we attempted to replicate the experiments of Amar-Halpert et al. (2017) with a larger number of observers. A total of 69 observers were randomly assigned into three groups, with 23 observers in each group (Fig. 1b). To assess the impact of different fitting methods on the results, we estimated thresholds using two approaches: the same fitting method as Amar-Halpert et al. (2017) (see Figs. S1-1, S2-1, S3-1 for individuals' data fitting) and the psignifit4 fitting method (see Figs. S1-2, S2-2, S3-2 for individuals' data fitting). The results of the same fitting method were presented unless specified.

In the reactivation group, thresholds in the post-test (day 5) were significantly reduced compared to those in the pre-test (day 1) (Fig. 2a (i), $\text{mean}_{\text{pretest}} = 127.3 \pm 10.2$ ms, $\text{mean}_{\text{posttest}} = 85.8 \pm 4.6$ ms, $t_{22} = 3.99$, $p < 0.001$, Cohen's d = 0.83; psignifit4: Fig. 2c(i), $\text{mean}_{\text{pretest}} = 143.1 \pm 14.4$ ms, $\text{mean}_{\text{posttest}} = 88.9 \pm 4.8$ ms, $t_{22} = 3.92$, $p < 0.001$, Cohen's d = 0.82). The TDT performance improved significantly (Fig. 2b, MPI = 27.1 ± 4.4 %, $t_{22} = 5.93$, $p < 0.001$, Cohen's d = 1.24; psignifit4: Fig. 2d, MPI = 28.0 ± 5.4 %, $t_{22} = 5.16$, $p < 0.001$, Cohen's d = 1.08). The mean percent improvement of the reactivation group in Amar-Halpert et al. (2017) was also significant (MPI = 20.6 ± 5.5 %).

In the no-reactivation group, thresholds in the post-test (day 5) were significantly lower than those in the pre-test (day 1) (Fig. 2a(ii), $\text{mean}_{\text{pretest}} = 117.9 \pm 6.0$ ms, $\text{mean}_{\text{posttest}} = 90.2 \pm 6.2$ ms, $t_{22} = 5.46$, $p < 0.001$, Cohen's d = 1.14; psignifit4: Fig. 2c(ii), $\text{mean}_{\text{pretest}} = 128.4 \pm 9.7$ ms, $\text{mean}_{\text{posttest}} = 92.6 \pm 6.5$ ms, $t_{22} = 3.9$, $p < 0.001$, Cohen's d = 0.82), with a threshold decrease of 27.7 ± 5.1 ms (psignifit4: 35.8 ± 9.1 ms) from pre-test to post-test. The TDT performance improved significantly (Fig. 2b, MPI = 22.7 ± 3.8 %, $t_{22} = 5.91$, $p < 0.001$, Cohen's d = 1.23; psignifit4: Fig. 2d, MPI = 23.9 ± 4.5 %, $t_{22} = 5.32$, $p < 0.001$, Cohen's d = 1.11). However, the no-reactivation group (n = 7) in Amar-Halpert et al. (2017) showed little improvement, with a threshold decrease of 7.6 ± 3.3 ms from the pre-test to the post-test.

In the full-practice group, thresholds in the post-test (day 5) were significantly lower than those in the pre-test (day 1) (Fig. 2a(iii), $\text{mean}_{\text{pretest}} = 125.0 \pm 8.7$ ms, $\text{mean}_{\text{posttest}} = 88.0 \pm 5.3$ ms, $t_{22} = 5.8$, $p < 0.001$, Cohen's d = 1.21; psignifit4: Fig. 2c(iii), $\text{mean}_{\text{pretest}} = 135.4 \pm 12.1$ ms, $\text{mean}_{\text{posttest}} = 89.4 \pm 6.3$ ms, $t_{22} = 5.39$, $p < 0.001$, Cohen's d = 1.1). The TDT performance improved significantly (Fig. 2b, MPI = 27.3 ± 3.4 %, $t_{22} = 8.04$, $p < 0.001$, Cohen's d = 1.68; psignifit4: Fig. 2d, MPI $=$ 30.5 $\pm$ 3.6 %, t_{22} $=$ 8.36, p $<$ 0.001, Cohen's d $=$ 1.74). The learning progress from day 1 to day 2 in our full-practice group was also significant (MPI $=$ 12.5 $\pm$ 4.2 %, t_{22} $=$ 2.96, p $=$ 0.007, Cohen's d $=$ 0.62; psignifit4: MPI $=$ 12.4 $\pm$ 4.0 %, t_{22} $=$ 3.09, p $=$ 0.005, Cohen's d $=$ 0.65). The total learning effect of the full-practice group in Amar-Halpert et al. (2017) was also significant (MPI $=$ 26.6 $\pm$ 5.9 %), but their day1-to-day2 improvement was insignificant (MPI $=$ 2.9 $\pm$ 5.8 %).

A classical independent sample t-test revealed no significant difference in learning improvements between the reactivation group and the no-reactivation group (t_{44} $=$ 0.75, p $=$ 0.46, Cohen's d $=$ 0.22; psignifit4: t_{44} $=$ 0.58, p $=$ 0.57, Cohen's d $=$ 0.17). A Bayesian independent-sample t-test also supports the null hypothesis, with a Bayes factor ($_{10}$) of 0.37 (psignifit4: $_{10}$ $=$ 0.34), representing the ratio of the likelihood of the observed data under the alternative hypothesis to the likelihood under the null hypothesis. A $_{10}$ of 0.37 or 0.34 indicated anecdotal evidence in favor of the null hypothesis, according to the interpretation of the Bayes factor magnitude (Johnson et al., 2022; Wagenmakers et al., 2018). The finding suggested that brief

1.34, anecdotal evidence for the alternative hypothesis).

Similarly, in the no-reactivation group, thresholds in post-test2 were also significantly reduced compared to those in the post-test (Fig. 3c, $\text{mean}_{\text{\_post-test}} = 89.0 \pm 6.7$ ms, $\text{mean}_{\text{\_post-test2}} = 71.3 \pm 5.7$ ms, $t_{19} = 2.93$, $p = 0.009$, Cohen's d = 0.66), and the TDT performance improved significantly during Phase2 (Fig. 3d(i), $\text{MPI}_{\text{\_Phase2}} = 17.0 \pm 5.6$ %, $t_{19} = 3.01$, $p = 0.007$, Cohen's d = 0.67; psignifit4: Fig. 3d(ii), $\text{MPI}_{\text{\_Phase2}} = 19.1 \pm 5.7$ %, $t_{19} = 3.37$, $p = 0.003$, Cohen's d = 0.75). The full-practice improvements over the two phases were significantly greater than the Phase 1 improvements (Fig. 3d(i), $\text{MPI}_{\text{\_Phase1}} = 24.1 \pm 4.0$ %, $\text{MPI}_{\text{\_total}} = 38.1 \pm 4.3$ %, $t_{19} = 3.23$, $p = 0.004$, Cohen's d = 0.72; psignifit4: Fig. 3d (ii), $\text{MPI}_{\text{\_Phase1}} = 25.1 \pm 4.8$ %, $\text{MPI}_{\text{\_total}} = 40.0 \pm 4.8$ %, $t_{19} = 3.40$, $p = 0.003$, Cohen's d = 0.76). Notably, both classical and Bayesian independent samples t-tests showed that the no-reactivation group's full-practice improvements were also greater than those of the

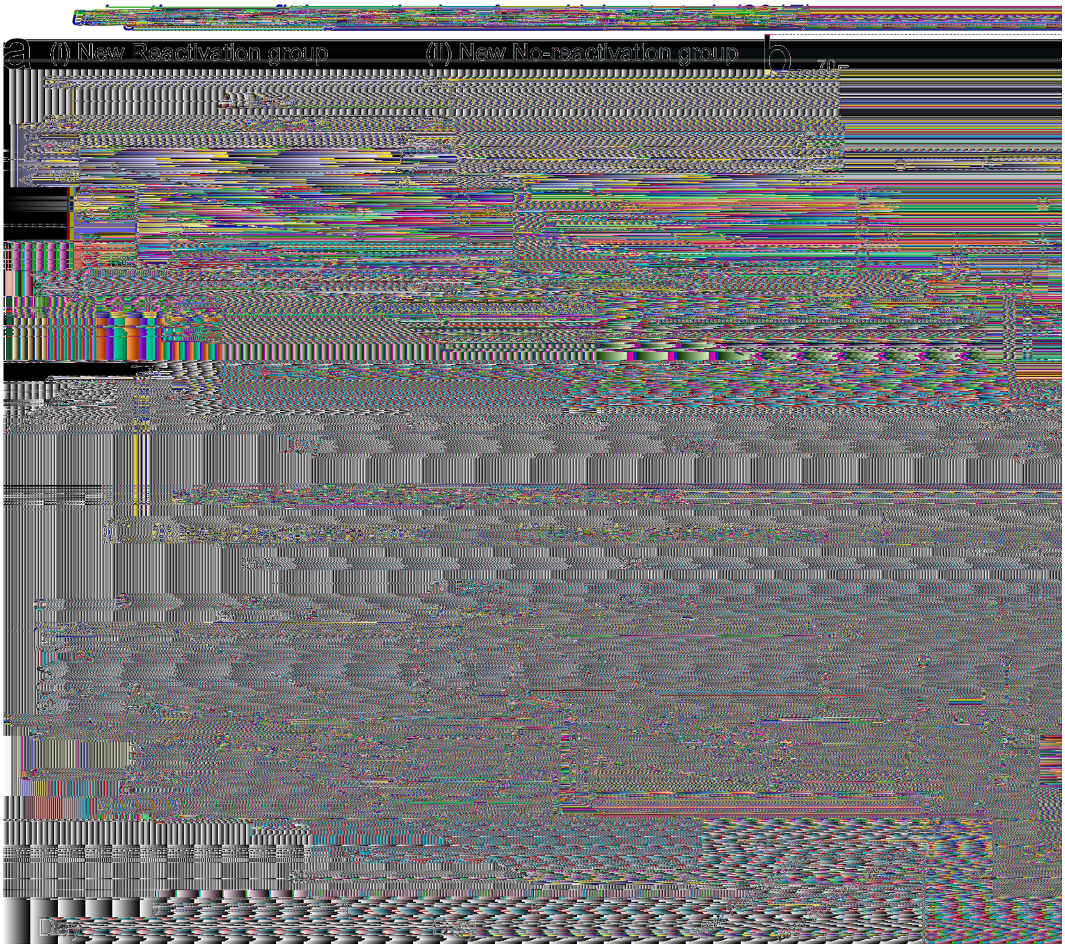

Fig. 4. Perceptual learning of TDT from pre-test (day 1) to post-test (day 5) in the new reactivation and no-reactivation groups

(Fig. 4a(ii), mean$_{pretest}$ = 122.9 ± 14.1 ms, mean$_{post\text{-}test}$ = 85.9 ± 7.4 ms, t_7 = 4.16, p = 0.004, Cohen's d = 1.47; psignifit4: Fig. 4c(ii), mean$_{pretest}$ = 146.5 ± 24.1 ms, mean$_{post\text{-}test}$ = 86.8 ± 8.0 ms, t_7 = 3.24, p = 0.01, Cohen's d = 1.15). The TDT performance also improved significantly (Fig. 4d, MPI = 28.4 ± 3.9 %, t_7 = 7.20, $p < 0.001$, Cohen's d = 2.54; psignifit4: Fig. 4d, MPI = 36.0 ± 5.3 %, t_7 = 6.75, $p < 0.001$, Cohen's d = 2.39).

Both classical and Bayesian independent samples t-tests indicated no significant difference in learning improvements between the two groups (t_{14} = 0.50, p = 0.63, Cohen's d = 0.25; BF_{10} = 0.47, anecdotal evidence for the null hypothesis; psignifit4: t_{14} = 1.0, p = 0.33, Cohen's d = 0.50; BF_{10} = 0.60, anecdotal evidence for the null hypothesis). Therefore, the results of the control experiment further indicated that reactivation did not yield additional gains in learning improvement compared to a no-reactivation condition, which is consistent with the results of the main experiment.

3.4. Finger error/lapse rate and goodness-of-fit under two fitting methods

Finger errors or lapse rates reflected stimulus-independent errors (e.g., attention lapses, response-key confusion). Their values in each fitting method were reported in Fig. 5 (also see Supplementary Tables S2 ~ S6). One can readily see that finger error/lapse rate values showed a trend of decrease across sessions (along with a decrease in thresholds), suggesting increasingly reliable judgments as training progressed.

Statistical analyses of the finger errors/lapse rate values across days were conducted for each group. When using the same fitting method as Amar-Halpert et al. (2017), a one-way repeated measures ANOVA showed that finger error values showed a significant decrease from day 1 (pre-test) to the last three days (reactivation group: Fig. 5a(i), $ps < 0.02$; full-practice group: Fig. 5a(iii), $ps < 0.04$) and a significant reduction from day 1 to the other four days (no-reactivation group: Fig. 5a(ii), $ps < 0.001$). Paired samples t-tests indicated that finger error values decreased significantly from day 1 to day 5 for both the new reactivation and new no-reactivation groups (Fig. 5a(iv), $ps < 0.01$).

Similarly, when using the psignifit4 fitting method, a one-way repeated measures ANOVA showed that the lapse rate values in the reactivation group showed no significant main effect of days (Fig. 5b(i), p = 0.06), but a paired samples t-test showed a significant decrease from day 1 to day 8 (p = 0.04). In the no-reactivation group, lapse rate values decreased significantly from day 1 to the other four days (Fig. 5b(ii), $ps < 0.03$). In the full-practice group, lapse rate values exhibited a significant reduction from the first two days to the last day (Fig. 5b(iii), $ps < 0.03$). Additionally, the lapse rate values of the new reactivation and new no-reactivation groups showed no significant changes from day 1 to day 5 (Fig. 5b(iv), $ps > 0.1$).

To evaluate how well the psychometric curves capture the empirical data of each individual, we assess goodness-of-fit by calculating deviance (Wichmann & Hill, 2001), in which smaller deviance indicates better goodness-of-fit (Haynes et al., 2024; Su et al., 2024). In particular,

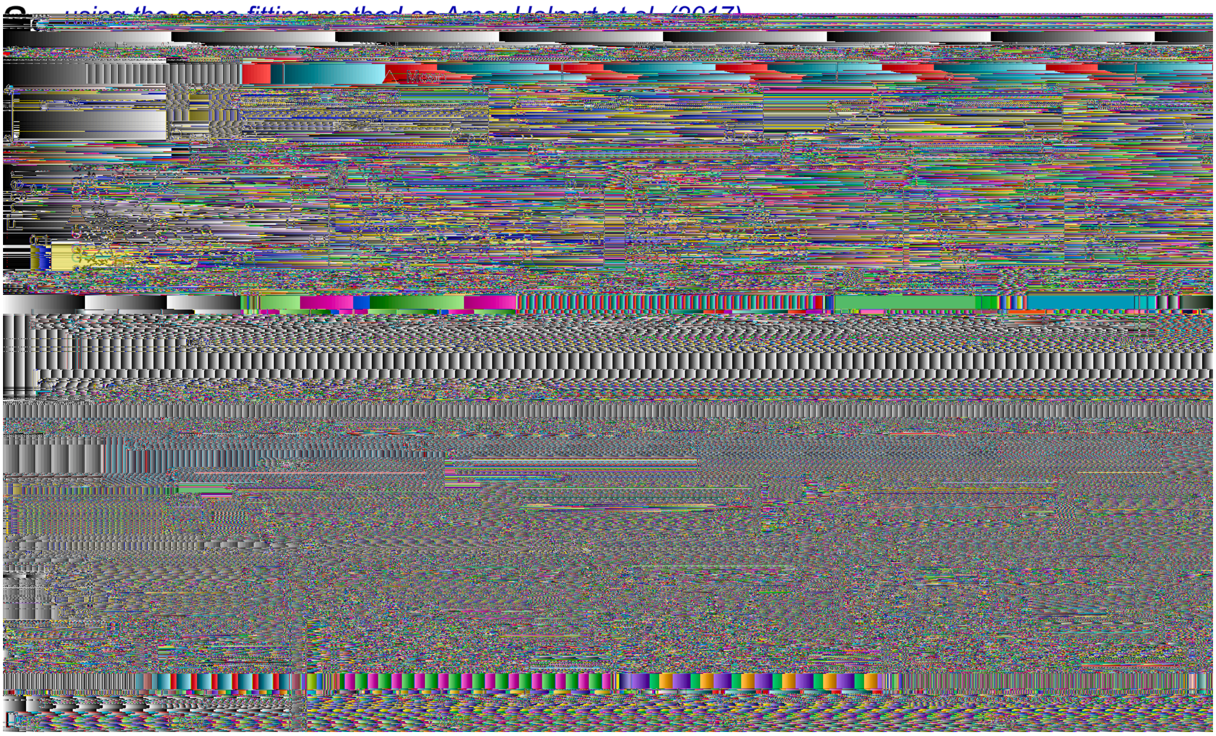

Fig. 5. Finger error/lapse rate across days in each group under two fitting methods. **a & b.** Finger error in the same fitting method as Amar-Halpert et al. (2017) (a) or lapse rate in the psignifit4 fitting method (b) changed as days in the reactivation group (i), the no-reactivation group (ii), the full-practice group (iii), the new reactivation group, and the new no-reactivation group (iv). Solid triangles and hollow circles represented mean and individual values, respectively. Error bars indicated ± 1 standard error of the mean. Note: 20 observers in the no-reactivation group finished the two phases of the operation.

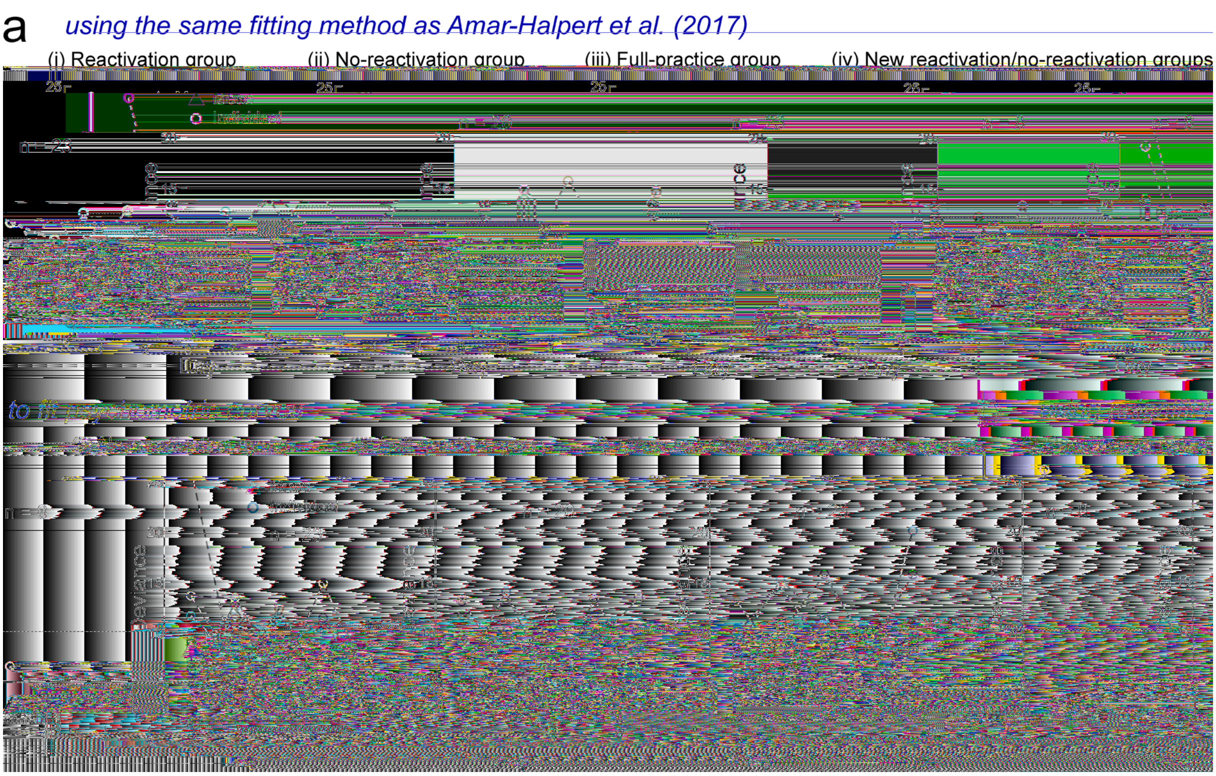

Fig. 6. Goodness-of-fit (deviance values) across days for each group under two fitting methods. **a & b.** Deviance values changed as days in the reactivation group (i), the no-reactivation group (ii), the full-practice group (iii), the new reactivation group, and the new no-reactivation group (iv) under the same fitting method as Amar-Halpert et al. (2017) (a) and the psignifit4 fitting method (b). Solid triangles and hollow circles represented mean and individual values, respectively. Error bars indicated ± 1 standard error of the mean.

both fitting methods indicated the same observer (R23) in the reactivation group with poor goodness-of-fit on day 1 in the chi-square test. After excluding this observer's deviance values (which were retained in Fig. 6(i)), we performed a statistical analysis of the deviance values across days within each group. A one-way repeated measures ANOVA showed that, in the reactivation group, deviance values decreased significantly from day 1 (pre-test) to day 8 (post-test2) under the same fitting method (Fig. 6a(i), $p = 0.02$), but not under the psignifit4 fitting method (Fig. 6b(i), $p = 0.10$). In the no-reactivation group, the main effect of days is not significant (Fig. 6a(ii), $p = 0.35$; psignifit4: Fig. 6b (ii), $p = 0.65$). In the full-practice group, the main effect of days is also not significant (Fig. 6a(iii), $p = 0.35$; psignifit4: Fig. 6b(iii), $p = 0.36$). Additionally, paired samples t-tests indicated no significant change in deviance values from day 1 to day 5 for both the new reactivation and no-reactivation groups (Fig. 6a(iv), $ps > 0.1$; psignifit4: Fig. 6b(iv), $ps > 0.2$). These findings indicated that goodness-of-fit did not change across days in most groups, reflecting the stability of the fitting models.

Besides, both fitting methods indicated a significantly smaller deviance in the control experimental groups compared to the main experimental groups. Specifically, classical independent samples t-tests showed the new reactivation group had significantly lower deviance values compared to the reactivation group on day 1 ($p = 0.002$; psignifit4: $p < 0.001$) and day 5 ($p = 0.002$; psignifit4: $p < 0.001$). Similarly, the new no-reactivation group had significantly lower deviance values than the no-reactivation group on day 1 ($p = 0.005$; psignifit4: $p = 0.004$) and day 5 ($p < 0.001$; psignifit4: $p < 0.001$). These results suggested that our control experiment with the modified constant stimuli method produced better goodness of fit.

3. . erformance on the fixation task

It is speculated that the large improvement of TDT in the no-reactivation group might be related to a strategy change in this group, where the observers shifted their focus of attention away from the fixation task towards the eccentric texture target discrimination task, thereby there might be a reduction in correctly reported fixation targets from the pre-test to the post-test. To answer this question, we analyzed the performance of the fixation task from the first day to the last day for each group. A one-way repeated measures ANOVA showed that the accuracies of the fixation task exhibited a significant increase from day 1 (pre-test) to the other four days (reactivation group: Fig. 7a, $ps < 0.001$; no-reactivation group: Fig. 7b, $ps < 0.01$). In the full-practice group, the accuracies increased significantly from day 1 (pre-test) to the last three days (Fig. 7c, $ps < 0.001$), and the accuracies on the second day were significantly lower than those on the last day (

comparable. Our results suggested that the improvement in the TDT task associated with brief reactivation (5 trials per day over 3 days) did not surpass that of the no-reactivation group with no exposure to the TDT task over the same d that the exact conditions of measurement play an important role in learning and transfer (Manning et al., 2018; Xiong et al., 2016; Zhang & Yu, 2018). Further evidence is needed to determine whether the current results are specific to the particular psychophysical method. Second, although effective foveal fixation was shown as the high accuracy in the central tumbling T/L task, the observers could shift (without being aware of it) their gaze towards the trained quadrant with the TDT target by 1 to 2 degrees to gain resolution of the texture elements and reduce crowding. Further exploration combining eye tracking could be used to determine any changes in fixation behavior across practice sessions and to test whether the tumbling T/L task is completed by extra-foveal vision. Third, although our findings of the null effect of reactivation resonate with previous research in visual and other domains (Chalkia et al., 2021; Chen & de Beeck, 2021; Luyten & Beckers, 2017), the positive findings of Amar-Halpert et al. (2017) are supported by reactivation effects in related paradigms like orientation detection (Bang et al., 2018). In the domain of motor learning, length of reactivation was identified as a crucial boundary condition determining whether human motor memories can undergo reconsolidation (de Beukelaar et al., 2014). Similar boundary conditions have not been reported in visual perceptual learning, more replications and attempts are needed to confirm the existence of the reconsolidation phenomenon in the field of vision science.

CRediT authorship contribution statement

Jun-Ping Zhu: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Jun-Yun Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Natural Science Foundation of China grant 32371079 (JYZ; Beijing, China). We thank Cong Yu for his insightful comments and discussions. We thank CaiXia Chen and KunYun Wang for helping with the data analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.visres.2025.108543.

Data availability

Data will be made available on request.

References

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *ature, 38* (6631), 401–406. https://doi.org/10.1038/387401a0

Amar-Halpert, R., Laor-Maayany, R., Nemni, S., Rosenblatt, J. D., & Censor, N. (2017). Memory reactivation improves visual perception. *ature euroscience, 20*(10), Article 1325-+. https://doi.org/10.1038/nn.4629

Auber, A., Tedesco, V., Jones, C. E., Monfils, M. H., & Chiamulera, C. (2013). Post-retrieval extinction as reconsolidation interference: Methodological issues or boundary conditions? *syc :*