The Multimodal Dependence of the Subject-Performed Task Effect in Instruction Memory: The Moderating Role of Instruction Conventional-ity[*]

Xie Tingting¹ Yang Lianyu¹ Wu Xiaotian¹ Zhou Diandian¹ Yu Zhanyu²

2026-08-11

ChinaRxiv

Abstract

The subject-performed task (SPT) effect refers to the phenomenon whereby individuals show higher subsequent recall accuracy when learning instructions through motor encoding (performing the actions described by the instructions) rather than verbal encoding (merely listening to the instructions). This study explains the SPT effect from a multichannel perspective and examines, through three experiments, the moderating role of instruction conventionality in the multichannel dependence of the SPT effect. Experiment 1 manipulated instruction conventionality and the visibility of object pictures, finding that conventional instructions produced an SPT effect both with and without pictures, whereas unconventional instructions showed no SPT effect under either condition. Experiment 2 replaced pictures with real objects and allowed manipulation, finding that unconventional instructions produced an SPT effect under this condition. Experiment 3 orthogonally manipulated vision (wearing a blindfold) and touch (wearing gloves), finding that the SPT effect for unconventional instructions emerged only when vision was available, while fine-grained tactile information did not affect the SPT effect. The three experiments demonstrate that the channel-dependence pattern of the SPT effect varies with instruction conventionality: highly conventional instructions are independently driven by kinesthesis, low-conventionality instructions require visual-kinesthetic coordination, and fine-grained tactile information is not necessary.

Full Text

The Multimodal Dependence of the Subject-Performed Task Effect in Instruction Memory: The Moderating Role of Instruction Conventionality*

Xie Tingting1 Yang Lianyu1 Wu Xiaotian1 Zhou Diandian1 Yu Zhanyu2

(1 School of Psychology, Fujian Normal University, Fuzhou 350117)
(2 School of Educational Science, Jiangsu Normal University, Xuzhou 221116)

Abstract The subject-performed task (SPT) effect refers to the finding that, when individuals learn instructions through action encoding—performing the actions described by the instructions—rather than through verbal encoding—only listening to the instructions—their subsequent recall accuracy is higher. From a multichannel perspective, the present study explains the SPT effect and uses three experiments to examine how instruction conventionality moderates the multichannel dependence of the SPT effect. Experiment 1 manipulated instruction conventionality and the visibility of object pictures, and found that conventional instructions produced an SPT effect both with and without pictures, whereas nonconventional instructions did not produce an SPT effect in either

condition. Experiment 2 replaced pictures with real objects and allowed manipulation, and found that nonconventional instructions produced an SPT effect under this condition. Experiment 3 orthogonally manipulated vision (wearing a blindfold) and touch (wearing gloves), and found that the SPT effect for nonconventional instructions appeared only when vision was available; fine-grained tactile information did not affect the SPT effect. The three experiments indicate that the modality-dependence pattern of the SPT effect varies with instruction conventionality: highly conventional instructions are driven independently by kinesthetic information, whereas low-conventionality instructions require visual-kinesthetic coordination, and fine-grained tactile information is not necessary.

Keywords SPT effect, instruction memory, multichannel processing theory

Classification number B842

Date received: 2026-01-26

* Supported by the Humanities and Social Sciences Fund of the Ministry of Education (24YJC190035) and the Natural Science Foundation of Fujian Province (2026J008125).

Corresponding author: Yu Zhanyu, E-mail: yzhany@jsnu.edu.cn

1 Introduction

Memory for verbal instructions refers to an individual' s ability to immediately encode, maintain, and retrieve verbal instructions, such as "push the chair" (Yang et al., 2018). As a basic cognitive ability underlying classroom listening, task execution, and self-care in daily life (Wang et al., 2022; Xie et al., 2024), its level directly affects individuals' adaptive behavioral performance. Research has shown that action encoding can improve memory for instructions: when learners study a series of verbal instructions through action encoding—that is, by personally carrying out the actions described in the instructions—rather than through verbal encoding—that is, merely listening to the instructions—their subsequent recall accuracy is higher. This phenomenon is known as the Subject-Performed Task effect, or SPT effect (Allen et al., 2019; Cohen, 1981).

With regard to the mechanism underlying the SPT effect, multimodal theory proposes that action encoding is superior to verbal encoding because, in addition to the verbal channel, it additionally activates multiple sensory channels such as kinesthesia, vision (e.g., observing objects and actions), and touch (e.g., contacting the surface of objects), thereby enhancing the richness and retrievability of memory traces (Wang Lijuan & Li Guangzheng, 2014; Bäckman & Nilsson, 1984; Engelkamp & Zimmer, 1984). However, this theory has not yet clarified the relative contribution of each channel. Among the multiple sensory channels that may be activated by action encoding, vision (which provides visual information about objects) and touch (which provides contact information

such as an object' s texture and weight) are the two channels most directly and most frequently activated during interaction with objects. Clarifying their roles is therefore a key entry point for understanding the multimodal mechanism of the SPT effect.

Existing studies have examined the role of the visual channel by manipulating object visibility, but the results have been inconsistent: some studies have found that the SPT effect exists under conditions without real objects (Engelkamp, 1996; Kormi-Nouri, 2000; Nyberg et al., 1991), whereas others have reported that the effect appears only when objects are visible (Mecklenbräuker et al., 2011; Srenffens et al., 2003). In response to this inconsistency, Yu Zhanyu (2014) proposed that material familiarity may be a moderating variable. That study manipulated the visual familiarity of nouns in the instructions—that is, whether participants knew what the object denoted by the noun looked like—and found that for verbal instructions composed of low-familiarity nouns, the SPT effect appeared only when pictures of the objects were presented; whereas for verbal instructions composed of high-familiarity nouns, the SPT effect was not affected by the visibility of object pictures.

It is worth noting that in Yu Zhanyu' s (2014) study, all verbal instructions were conventional verb–object collocations, such as the high-familiarity "pull a zipper" and the low-familiarity "tune a string," in which the verb and noun had already formed a highly automatized association. When the object noun is familiar to an individual, memory for verbal instructions composed of such conventional verb–object collocations (hereafter referred to as "conventional instructions") may rely on internally activated multimodal representations, including motor programs, typical visual scenes, and tactile expectations, and may depend less on external objects. If the individual is unfamiliar with the object noun itself, however, the corresponding representations must be evoked with the aid of the object. Recent studies have shown that even under conditions without real objects, a significant SPT effect can still be observed for conventional instructions composed of familiar object nouns (Wang Lijuan & Ma Jialin, 2023; Ma et al., 2025), which to some extent supports the view that "for conventional instructions composed of familiar object nouns, internal cue-

cues are sufficient to support the SPT effect." However, when the instruction is a nonconventional collocation (e.g., "turn the clothes hanger" ; hereafter, "nonconventional instructions"), the situation may be different. In such instructions, the verb and noun lack a preexisting association and must be integrated online; their memory processing may therefore rely more strongly on external sensory input. In other words, whether the object is presented may determine whether the SPT effect emerges under such instructions. Although some studies have used nonconventional instructions as experimental materials (e.g., Nyberg et al., 1991), they did not manipulate object visibility. Thus, an unresolved question is: for instructions composed of familiar object nouns, when the verb and noun form a nonconventional collocation, is external visual input a necessary condition for the SPT effect to occur?

In addition, when examining the influence of "object presentation" on the SPT effect, existing studies have generally attributed the role of objects to visual cues. Yu Zhanyu (2014) used object pictures, which mainly provided visual cues. However, most other studies (Engelkamp, 1996; Kormi-Nouri, 2000; Mecklenbräuker et al., 2011; Nyberg et al., 1991) used real objects, thereby providing both visual and tactile input. By attributing the effect of "object presentation" to the visual channel, these studies may have overlooked the potential contribution of tactile information, such as an object' s texture, weight, and shape. Tactile input may likewise influence the SPT effect by enriching action representations. Given that multimodal processing theory explicitly identifies touch as a potential contributing channel (Bäckman & Nilsson, 1984), yet its independent role has not been isolated, it is necessary to specifically examine the unique contribution of tactile input to the SPT effect so as to test the predictions of multimodal theory more precisely.

This Study

This study aims to examine the mechanism underlying the SPT effect within the framework of multimodal processing theory, with the goal of answering two questions: (1) For instructions composed of familiar object nouns (e.g., "pencil"), do conventional collocations (e.g., "sharpen a pencil") and nonconventional collocations (e.g., "poke a pencil") elicit SPT effects under action encoding that differ in their dependence on sensory channels? (2) Does tactile input make an independent contribution to the SPT effect? Clarifying these two questions will help specify the concrete manner of "multimodal synergy" in multimodal processing theory—that is, whether all channels contribute and the effect increases as more channels are involved, or whether the role of a channel differs depending on whether the instruction can evoke existing action experience. It may also provide more targeted memory-support strategies for populations with limited visual or tactile function, such as children with autism and individuals with visual impairment or tactile damage.

To answer these two questions, the present study designed three experiments. Experiment 1 used familiar objects commonly encountered in daily life, manipulated the conventionality of instructions through the verb, and, as in the study by Yu Zhanyu (2014), manipulated the visual channel through the presence or absence of object pictures, in order to test whether conventional and nonconventional instructions differ in the extent to which the SPT effect depends on the visual channel. Because the preexisting association between the verb and noun is strong in conventional instructions, we hypothesized that the SPT effect for such instructions does not depend on the visual channel and can be induced by kinesthetic input alone. For nonconventional instructions, because the verb and noun lack a preexisting association, individuals must temporarily integrate the verb and noun to construct an action representation, and may therefore need the visual channel for the SPT effect to occur. However, Experiment 1 used object pictures as the carrier of visual cues, and research has shown that

object pictures are weaker than real objects both in capturing individuals' attention and in activating object names (Gomez et al., 2018; Snow et al., 2023). Accordingly, object pictures may, because of vis-

because the strength of the party-line cue was insufficient, it could not effectively promote the temporary integration of the verb and noun, and thus could hardly induce the SPT effect; the individual may need to engage in actual interaction with the object in order to effectively construct an action–object linkage and thereby produce the SPT effect. To test the latter hypothesis, Experiment 2 focused on non-habitual instructions, replacing them with real objects (objects that participants were required actually to manipulate), and examined whether object interactivity is a necessary condition for the emergence of the SPT effect under non-habitual instructions. If the results of Experiments 1 and 2 show that "the SPT effect for habitual instructions does not depend on the presentation of objects, whereas the SPT effect for non-habitual instructions appears only under real-object conditions," then it can be preliminarily inferred that the SPT effect under habitual instructions does not depend on the visual channel, whereas the SPT effect under non-habitual instructions depends on sensory input provided by interaction with objects (other than kinesthesia). Because object interaction involves both vision and touch, Experiment 2 cannot determine which sensory channel plays the role. Experiment 3 will separate the contributions of vision and touch: under the premise that real objects are presented, it will orthogonally manipulate vision (blindfold vs. no blindfold) and fine touch (wearing gloves vs. not wearing gloves), and set a purely verbal encoding baseline, to examine the independent contributions of vision and fine touch to the SPT effect under non-habitual instructions. Through these three experiments, the present study will reveal how the sensory-channel dependence of the SPT effect changes with instruction habituality.

2 Experiment 1: Whether Instruction Habituality Modulates the Influence of Object-Picture Presentation on the SPT Effect

Experiment 1 examined whether, for verbal instructions composed of familiar object nouns, instruction habituality modulates the influence of object visual cues on the SPT effect. As in the study by Gan Yu (2014), this experiment manipulated whether object pictures were presented and whether visual cues were provided. In addition, familiar object nouns were used, and verb manipulation was adopted to form habitual instructions (e.g., "sharpen the pencil") and non-habitual instructions (e.g., "saw the pencil").

2.1 Method

2.1.1 Participants Based on the interaction effect size reported by Xie et al. (2024) ($\eta_p^2 = 0.082$), an a priori power analysis was conducted using G*Power 3.1 ($\alpha = 0.05$, power = 0.95, repeated-measures ANOVA, within–between inter-

action), estimating a required total sample size of 56 (14 participants per group). To obtain more stable effect estimates and reserve room for data exclusion, before data collection we decided to increase the sample size to 50 participants per group, recruiting a total of 200 undergraduate students aged 18–22 years ($M_{age} = 20.14 \pm 0.93$ years; 81 males). All participants self-reported no history of psychiatric illness, normal or corrected-to-normal vision, right-handedness, and normal intelligence. According to the pre-established criteria, data were to be excluded if they deviated from the sample mean by more than 3 standard deviations, or if participants reported distraction during the encoding phase and obtained a recall score of 0. No participants met the exclusion criteria, and all were included in the final analysis. After the experiment, participants received a small cash payment.

2.1.2 Materials Based on the instruction-memory materials database of Wang et al. (2022) and Xie et al. (2024), all 16 object nouns were selected from it, and 13 self-compiled common object nouns were added, forming 29 candidate nouns. For each candidate word, one matching “符”

verbs that conform to everyday habits (e.g., matching “switch” with “press”) to form habitual instructions (e.g., “press the switch”). Because some verbs appeared repeatedly in the matching process, duplicate items were removed to prevent repeated verbs from interfering with the experimental results. Ultimately, 20 nouns were retained, and each noun was assigned one unique, non-repeated habitual verb, yielding 20 habitual instructions. In addition, for these 20 nouns, 20 verbs were matched that are usually not collocated with them but whose actions are executable (e.g., although “press the mirror” is semantically unconventional, an individual can complete the action of “pressing”), thereby forming non-habitual instructions; the 20 verbs within this set did not repeat.

To test the suitability of the materials, 10 undergraduates who did not participate in the formal experiment were invited to complete two rating tasks ($M_{\text{age}} = 19.90 \pm 0.74$ years; 5 males): (1) they rated the familiarity of 20 object nouns, 20 habitual verbs, and 16 non-habitual verbs on a 1–7 scale (4 non-habitual verbs overlapped with habitual verbs and were not included in the ratings; 1 = extremely unfamiliar, 7 = extremely familiar); (2) they rated the collocational plausibility of 40 verb–noun phrases on a 1–7 scale (1 = completely implausible, 7 = completely plausible). The results showed that the familiarity scores for the nouns, habitual verbs, and non-habitual verbs all ranged from 4 to 7. Taking 4 as the criterion for “relatively familiar,” all vocabulary items received scores no lower than 4, indicating that they were all relatively familiar to the participants. In terms of collocational plausibility, habitual collocations scored significantly higher ($M = 6.67$, $SD = 0.69$, range 4–7) than non-habitual collocations ($M = 1.38$, $SD = 0.70$, range 1–3), $t(199) = 79.78$, $p < 0.001$.

The 20 habitual instructions ultimately used were: sharpen the pencil, move the chair, press the switch, carry the plate, tie the shoelace, open the umbrella, hang the clothes hanger, brush with the toothbrush, comb with the comb, wring the

towel, squeeze the toothpaste, rub the soap, lift the bucket, look in the mirror, grip the chopsticks, swipe on the mobile phone, flip through the book, wipe the table, wear the glasses, and cover oneself with the quilt.

The corresponding 20 non-habitual instructions were: poke the pencil, tear the chair, push the switch, break the plate, knock the shoelace, turn over the umbrella, pick at the clothes hanger, cut the toothbrush, pinch the comb, touch the towel, comb the toothpaste, wring the soap, pat the bucket, press the mirror, pull the chopsticks, flick the mobile phone, scrape the book, swing the table, shake the glasses, and rotate the quilt.

2.1.3 Design

A 2 (instruction habituality: habitual vs. non-habitual) ×2 (object-picture presentation: presented vs. not presented) ×2 (encoding mode: verbal vs. action) mixed design was adopted. Encoding mode was a within-participant variable, and the remaining variables were between-participant variables. The dependent variable was the number of instructions correctly recalled under each experimental condition.

2.1.4 Procedure

The experimental flowchart is shown in Figure 1. Participants were randomly assigned to one of four groups: (1) habitual-picture-presented group; (2) habitual-picture-not-presented group; (3) non-habitual-picture-presented group; (4) non-habitual-picture-not-presented group. "Habitual" refers to semantically plausible instructions (e.g., "sharpen the pencil"); "non-habitual" refers to instructions that are semantically unconventional but whose actions are executable (e.g., "poke the pencil"). All participants were required to complete both action-encoding and verbal-encoding tasks, and each condition included learning and recall tasks for 10 instructions. The presentation order of the two conditions was balanced across participants. The specific task procedure was as follows:

(1) Action-encoding condition: while the experimenter orally stated an instruction (lasting approximately 3 seconds), participants in the picture-presented group would

saw a color photograph of the actual object corresponding to the noun in the instruction (held by the experimenter and shown for 7 s); in the no-picture condition, they saw a blank sheet of A4 paper (held by the experimenter and shown for 7 s). During this period, participants were required to perform the corresponding action according to the instruction they heard (for example, when hearing "sharpen the pencil," they made the action of sharpening a pencil with their hands). After all 10 instructions had been presented, participants were required to orally report, within 2 min and in any order, the 10 instructions they had heard.

(2) **Language-encoding condition:** Color photographs of actual objects or blank paper were likewise presented (7 s each), but participants kept their hands still and only listened to the instructions. After all 10 instructions had been presented, participants completed a 2-min oral free-recall task.

Participants' oral recall was recorded and independently scored by a non-psychology student who was unaware of the purpose of the experiment. The scoring criteria were as follows: the instruction orally reported by the participant had to be consistent with the instruction presented during the learning phase; it was counted as correct only when both the verb and the noun were correct (e.g., if the learning-phase instruction was "sharpen the pencil," the participant had to say "sharpen the pencil"). No requirement was imposed on order. Responses that included only the noun, only the verb, or a mismatch between the verb and the noun were all counted as incorrect.

Random assignment → Oral presentation of 1 instruction + presentation of picture/blank paper → {Action encoding: participant performs the action; Language encoding: participant remains still} → after 10 instructions → 2-min oral free recall

Figure 1. Flowchart of Experiment 1. Random assignment means that participants were randomly assigned to the conventional-picture-presented group, the conventional-picture-not-presented group, the nonconventional-picture-presented group, or the nonconventional-picture-not-presented group. The conventional-picture-presented group was presented with conventional instructions and object pictures; the conventional-picture-not-presented group was presented with conventional instructions and blank paper; the nonconventional-picture-presented group was presented with nonconventional instructions and object pictures; and the nonconventional-picture-not-presented group was presented with nonconventional instructions and blank paper.

2.2 Results

The four groups of participants did not differ significantly in age ($F(3, 196) = 0.78, p = 0.504$) or gender ratio ($\chi^2(3) = 3.88, p = 0.275$). A three-factor repeated-measures analysis of variance was conducted on the number of correctly recalled items for all participants; the descriptive statistics are shown in Figure 2.

The main effect of instruction conventionality was significant, $F(1, 196) = 185.42, p < 0.001, \eta_p^2 = 0.49$. The main effect of object-picture presentation was significant, $F(1, 196) = 144.96, p < 0.001, \eta_p^2 = 0.43$. The main effect of encoding mode was significant, $F(1, 196) = 42.47, p < 0.001, \eta_p^2 = 0.18$. The interaction between instruction conventionality and object-picture presentation was significant, $F(1, 196) = 22.67, p < 0.001, \eta_p^2 = 0.10$. The interaction between instruction conventionality and encoding mode was significant, $F(1, 196) = 21.93, p < 0.001, \eta_p^2 = 0.10$. The interaction between object-picture presentation and encoding mode was not significant, $F(1, 196) = 0.39, p =$

$0.531, \eta_p^2 = 0.002$. The three-way interaction among instruction conventionality, object-picture presentation, and encoding mode was not significant, $F(1, 196) = 0.67, p = 0.413, \eta_p^2 = 0.003$.

According to statistical conventions, simple-effects analyses based on a priori theoretical hypotheses do not depend on the significance of the overall interaction effect (Clark-Carter, 2018; Howell, 2013). Given that the present study had clear theoretical predictions regarding the pattern of the SPT effect under conventional and nonconventional instructions—namely, that conventional instructions would show an SPT effect regardless of whether pictures were present, whereas nonconventional instructions would not—we conducted Bonferroni-corrected simple-effects analyses of the differences between action encoding and language encoding under the four conditions (instruction conventionality × object-picture presentation). The results showed that, for conventional instructions, regardless of whether pictures were presented, the action

under action encoding than under verbal encoding, regardless of whether pictures were presented (picture present: $M_{\text{action}} = 8.22 \pm 0.98$ vs. $M_{\text{verbal}} = 7.08 \pm 1.37$, $p < 0.001$, 95% CI = [0.73, 1.55]; picture absent: $M_{\text{action}} = 7.04 \pm 1.18$ vs. $M_{\text{verbal}} = 5.87 \pm 1.20$, $p < 0.001$, 95% CI = [0.77, 1.59]); that is, the SPT effect was present in both cases. For non-habitual instructions, regardless of whether pictures were presented, there was no difference in the number of instructions recalled under action encoding versus verbal encoding (picture present: $M_{\text{action}} = 6.36 \pm 1.63$ vs. $M_{\text{verbal}} = 6.02 \pm 1.41$, $p = 0.102$, 95% CI = [-0.07, 0.75]; picture absent: $M_{\text{action}} = 3.44 \pm 1.64$ vs. $M_{\text{verbal}} = 3.40 \pm 1.49$, $p = 0.847$, 95% CI = [-0.37, 0.45]); that is, the SPT effect was absent in both cases. A further comparison of the magnitude of the SPT effect for habitual instructions under the picture-present and picture-absent conditions (i.e., performance in the action-encoding condition minus performance in the verbal-encoding condition; see also Allen et al., 2019; Cohen, 1981; Xie et al., 2024) showed no significant difference between the two, $M_{\text{picture present}} = 1.14 \pm 1.37$ vs. $M_{\text{picture absent}} = 1.18 \pm 1.47$, $t(98) = -0.14$, $p = 0.888$, Cohen' s $d = 0.03$, 95% CI = [-0.60, 0.52].

Figure 2. Means and standard errors of the number of correctly recalled instructions in each condition in Experiment 1. $p < 0.05$

Figure labels: y-axis, "Number of correctly recalled instructions" ; legend, "verbal encoding" and "action encoding"; x-axis conditions, "habitual–picture present," "habitual–picture absent," "non-habitual–picture present," and "non-habitual–picture absent."

2.3 Discussion

Experiment 1 found that the influence of presenting object pictures on the SPT effect did not vary as a function of instruction habituality. Specifically, for habitual instructions, the SPT effect was present both when object pictures were

present and when they were absent. This indicates that when an instruction involves an action–object pairing that is common in daily life, such as "sharpen the pencil," external visual cues are not necessary; motor input is sufficient to support effective memory processing. This result is consistent with the findings of Yu Zhanyu (2014). By contrast, for non-habitual instructions, such as "saw the pencil," the SPT effect did not emerge regardless of whether object pictures were presented. This finding suggests that when the verb and noun in an instruction lack a pre-existing association, relying solely on simulated action in the absence of a physical object may make it difficult to establish a stable action–language integrated representation, thereby failing to facilitate memory.

It should be noted that "the absence of an SPT effect for non-habitual instructions whether or not object pictures were presented" is not equivalent to "visual cues are ineffective." If action encoding itself cannot initiate effective memory processing (e.g., because of a lack of perception of the object' s function or manner of operation···

···the potential gain from visual input cannot be manifested. In other words, the absence of the effect may have stemmed from a failure of baseline processing rather than from the ineffectiveness of visual cues. In Experiment 1, pictures were used as visual cues; their informational strength—no three-dimensional structure and no action–object interaction—may have been insufficient to support online action planning for non-habitual instructions. In addition, Experiment 1 used a between-participants design to test the role of object pictures; individual differences may have increased error variance, thereby relatively reducing statistical power and limiting the detection of weak effects. Experiment 2 replaced object pictures with real objects, allowing participants to engage in genuine action–object interaction, and tested, in a within-participants design, whether non-habitual instructions could elicit an SPT effect through action encoding.

3 Experiment 2: The Influence of Real-Object Presentation on the SPT Effect Under Non-Habitual Instructions

Experiment 1 found that the SPT effect did not emerge for non-habitual instructions under either the object-picture or no-object-picture condition. To examine whether this absence of the effect was due to insufficient visual cues, Experiment 2 replaced pictures with interactable real objects and explored whether, when richer sensory input was provided—including visual, tactile, and action-related physical information—action encoding under non-habitual instructions could elicit an SPT effect.

3.1 Method

3.1.1 Participants Based on the interaction-effect size reported by Xie et al. (2024) ($\eta_p^2 = 0.082$), an a priori power analysis was conducted using G*Power 3.1 ($\alpha = 0.05$, power $= 0.95$; repeated-measures ANOVA, within-subject factor).

The required sample size was 26. To obtain more stable effect estimates and to allow room for data exclusion, we decided before data collection to increase the sample size to 64 participants. In total, 64 undergraduates were recruited ($M_{\text{age}} = 19.59 \pm 1.07$ years; 23 males). Following the exclusion criteria of Experiment 1, Experiment 2 excluded data from 2 participants who reported mind-wandering during the encoding phase and had a recall score of 0, as well as 4 extreme observations (deviating from the sample mean by 3 standard deviations). The final sample included data from 58 participants ($M_{\text{age}} = 19.55 \pm 1.06$ years; 20 males). All participants self-reported no history of mental illness, normal or corrected-to-normal vision, right-handedness, and normal intelligence.

3.1.2 Materials Because some of the objects used in Experiment 1 (e.g., a quilt) were difficult to manipulate physically in a tabletop experimental setting, the present experiment did not use those real-object materials. Following the matching logic of Experiment 1, we selected from the material pools of Wang et al. (2022) and Xie et al. (2024) 13 highly familiar object nouns suitable for tabletop manipulation: scissors, key, pencil, comb, towel, cup, tissue, umbrella, spoon, chopsticks, mirror, mask, and toothbrush. These were combined with 10 high-frequency verbs—turn, touch, push, shake, knock, flip, pull, press, cut, and pinch—to generate 130 non-habitual instructions, such as "cut the towel."

Ten undergraduates who had participated in the ratings for Experiment 1 (5 males and 5 females) were invited to conduct two ratings of the materials: (1) they rated the familiarity of the 13 nouns and 10 verbs on a 1–7 scale (1 = extremely unfamiliar, 7 = extremely familiar); and (2) they rated the matching plausibility of the 130 instructions on a 1–7 scale (1 = completely implausible, 7 = completely plausible). The results showed that all noun

The familiarity of both the nouns and the verbs was relatively high (range: 5–7); the appropriateness of the noun–verb pairings in the instructions was relatively low (range: 1–3, below the midpoint of 4 on the 7-point scale), indicating clear semantic unconventionality. From these, 40 non-routine instructions were selected. The screening criterion was that verbs and nouns did not repeat within each sequence, so as to control for mutual interference among materials. The 40 selected instructions formed 4 sequences, with 10 instructions in each sequence. All instructions could be completed by participants with one hand in a desktop setting.

3.1.3 Design

A 2 (object presentation: present vs. absent) × 2 (encoding mode: verbal vs. action) within-subjects design was adopted. The dependent measure was the number of instructions correctly recalled under each condition.

3.1.4 Procedure

The experimental flowchart is shown in Figure 3. Each participant was required to complete four experimental conditions, with condition order counterbalanced across participants. Each condition included encoding and recall phases for 10 non-routine instructions. The tasks in the encoding phase for each condition were as follows:

(1) Object-present action-encoding condition: while orally presenting an instruction (about 3 s), the experimenter placed the corresponding object at the center of the desktop (presented for 7 s); during this period, participants were required to perform on the object an action consistent with the verb (e.g., upon hearing "cut the towel," they made a "cutting" action toward the towel with their hand). (2) Object-absent action-encoding condition: there was no object on the desktop; when participants heard the instruction, they were required to perform the corresponding action without a physical object during this period. (3) Object-present verbal-encoding condition: while orally presenting an instruction (about 3 s), the experimenter placed the corresponding object at the center of the desktop (7 s); participants were required to keep their hands still during this period, only listen to the instruction, and not perform any action. (4) Object-absent verbal-encoding condition: there was no object on the desktop; when participants heard the instruction (3 s), they were required to keep their hands still and only listen to the instruction.

Under all conditions, after the encoding of the 10 instructions was completed, participants orally recalled as many of the 10 instructions as possible within 2 min, in any order. Participants' oral recall was recorded, and scoring was conducted in the same way as in Experiment 1.

Object-present action encoding: [orally state 1 instruction + present object]
Object-absent action encoding: [orally state 1 instruction + do not present object]
→ (action encoding) [participant performs the action]

Object-present verbal encoding: [orally state 1 instruction + present object]
Object-absent verbal encoding: [orally state 1 instruction + do not present object]
→ [participant remains still] (verbal encoding)

After 10 instructions → [2-min oral free recall]

Figure 3. Flowchart of Experiment 2. Each participant was required to complete four condition tasks: object-present action encoding, object-absent action encoding, object-present verbal encoding, and object-absent verbal encoding.

3.2 Results

A two-factor repeated-measures analysis of variance was conducted on the number of correctly recalled items in each condition; descriptive statistics are shown in Figure 4. The main effect of encoding mode was significant, $F(1, 57) = 28.29$, $p < 0.001$, $\eta_p^2 = 0.33$. The main effect of object presentation was significant, $F(1, 57) = 247.15$, $p < 0.001$, $\eta_p^2 = 0.81$. The interaction between object presentation and encoding mode was significant, $F(1, 57) = 16.65$, $p < 0.001$, $\eta_p^2 = 0.23$. Bonferroni-corrected simple-effects analyses showed that when objects were presented, the number recalled in the action-encoding condition ($M = 6.69$, $SD =$

1.52) was greater than the number recalled in the verbal-encoding condition ($M = 5.03, SD = 1.77$), $p < 0.001$, 95% CI = $[1.09, 2.22]$. When no object was presented, the number recalled in the action-encoding condition ($M = 3.09, SD = 1.48$) did not differ from that in the verbal condition ($M = 2.81, SD = 1.28$), $p = 0.188$, 95% CI = $[-0.14, 0.69]$. Under both action encoding and verbal encoding, the number recalled in the object-present condition was greater than that in the no-object condition, $ps < 0.001$, and none of the 95% CIs included 0.

Figure 4. Mean and standard error of the number of correctly recalled instructions under each condition in Experiment 2. $*p < 0.05$

3.3 Discussion

Experiment 2 found that, for non-habitual instructions, an SPT effect appeared only when an object was presented. This contrasts with the finding in Experiment 1 that there was no SPT effect in the picture condition, suggesting that interaction with real objects may be key to producing an SPT effect for non-habitual instructions. However, Experiments 1 and 2 differed in experimental design and materials. To rule out the possibility that these differences confounded the visual-cue effect under non-habitual instructions, we added a picture experiment with the same structure as Experiment 2. The participants were 26 undergraduates from the same university as those in Experiment 2, including 10 males; a within-participants design was used; the materials were the same as in Experiment 2, with pictures of objects replacing real objects, and participants performed simulated actions toward the pictures. The results showed that the interaction between object-picture presentation and encoding method was not significant ($F(1, 25) = 0.84, p = 0.368, \eta_p^2 = 0.03$); neither the main effect of encoding method ($F(1, 25) = 1.60, p = 0.218, \eta_p^2 = 0.06$) nor the main effect of object-picture presentation ($F(1, 25) = 1.74, p = 0.20, \eta_p^2 = 0.07$) was significant either. The number recalled under picture/action encoding was $M = 3.92, SD = 1.32$; under picture/verbal encoding, $M = 4.08, SD =$ 1.32; under no-picture/action encoding, $M = 3.38, SD = 1.53$; and under no-picture/verbal encoding, $M = 3.96, SD = 1.69$. This indicates that, even in a within-participants design with greater statistical power, the picture condition

still could not produce an SPT effect for non-habitual instructions.

Taken together, the results of Experiments 1 and 2 and the supplementary experiment indicate that, for non-habitual instructions, the emergence of the SPT effect depends on the presence of real objects with which participants can interact. When there are only object pictures or no objects, mere action simulation may, because it lacks access to the physical attributes of the object

(such as shape, texture, and manipulability), making it difficult to effectively integrate action and verbal information; real objects, by simultaneously providing visual, tactile, and action feedback, allow action coding to be successfully initiated.

From the perspective of multisensory channels, the results of Experiment 1 indicate that the visual channel is not necessary for producing the SPT effect under habitual instructions. The results of Experiment 1, Experiment 2, and the supplementary experiment together show that, for non-habitual instructions, kinesthesia alone is insufficient to produce the SPT effect, and the participation of sensory channels other than kinesthesia is still required. However, real objects simultaneously activate vision and touch, and the present results cannot determine whether the key role is played by vision, touch, or both. Therefore, Experiment 3 adopted an orthogonal design, manipulating the presence or absence of visual and fine tactile input to examine which sensory channels the SPT effect under non-habitual instructions depends on.

4 Experiment 3: Independent Contributions of Visual and Tactile Channels to the SPT Effect

Experiment 3 separated the contributions of vision and fine touch to the SPT effect under non-habitual instructions by manipulating their availability. If a given sensory input plays a key role, then depriving it should weaken or eliminate the SPT effect; if the two work jointly, then the effect should be greatest only when both vision and fine touch are available.

4.1 Method

4.1.1 Participants

Based on the interaction effect size reported by Xie et al. (2024) ($\eta_p^2 = 0.082$), an a priori power analysis was conducted using G*Power 3.1 ($\alpha = 0.05$, power = 0.95, repeated-measures ANOVA, within-subject factor), yielding a required sample size of 23. To obtain more stable effect estimates and allow room for data exclusion, before data collection we decided to increase the sample size to 54 participants, and recruited 54 undergraduates ($M_{\text{age}} = 19.80 \pm 1.17$ years; 25 males). The exclusion criteria were the same as in Experiment 1. No participants were excluded, and all were included in the final analysis. All partic-

ipants reported no history of neurological or psychiatric disorders, had normal or corrected-to-normal vision, and were right-handed.

4.1.2 Materials

The experimental materials were the same as in Experiment 2. Ten non-habitual instructions formed one sequence, with no repetition of verbs or nouns within each sequence. A total of five sequences were prepared, corresponding to the five experimental conditions.

4.1.3 Design

A single-factor within-subjects design was adopted. The independent variable was coding mode, comprising five levels: action coding with vision and touch, action coding with vision but without fine touch, action coding without vision but with touch, action coding without vision and without fine touch, and verbal coding. The dependent measure was the number of instructions correctly recalled under each condition.

4.1.4 Procedure

The experimental procedure is shown in Figure 5. Each participant completed five experimental conditions, with the order of conditions counterbalanced across participants. Each condition included the encoding and recall phases for 10 non-habitual instructions. The tasks for each condition during the encoding phase were as follows:

(1) **Action-encoding condition with vision and touch:** While the experimenter orally stated one instruction (about 3 s), the corresponding object was placed at the center of the tabletop (7 s); during this period, the participant was required to perform on the object an action consistent with the verb (e.g., when hearing "cut the towel," the participant used the hand to make a "cutting" action on the towel). (2) **Action-encoding condition with vision but without fine tactile information:** The participant put on thick cotton gloves in advance (blocking perception of texture and temperature, while retaining coarse shape and resistance); while the experimenter orally stated the instruction (about 3 s), the corresponding object was placed at the center of the tabletop (7 s); during this period, the participant was required to perform on the object an action consistent with the verb. (3) **Action-encoding condition without vision but with touch:** The participant put on a fully light-blocking silicone eye mask in advance (ensuring no light leakage); while the experimenter orally stated the instruction (about 3 s), the corresponding object was placed at the center of the tabletop (7 s); during this period, the participant was required to perform on the object an action consistent with the verb. (4) **Action-encoding condition without vision and without fine tactile information:** The participant put on thick cotton

gloves in advance (blocking perception of texture and temperature, while retaining coarse shape and resistance) and a fully light-blocking silicone eye mask (ensuring no light leakage); while the experimenter orally stated the instruction (about 3 s), the corresponding object was placed at the center of the tabletop (7 s); during this period, the participant was required to perform on the object an action consistent with the verb. (5) **Verbal-encoding condition:** No object was present on the tabletop; when the participant heard the instruction (3 s), they could only listen to the instruction and were not allowed to make any action.

Under all conditions, after encoding the 10 instructions, the participant orally recalled as many of the 10 instructions as possible within 2 minutes, in any order. Participants' oral recall was recorded, and scoring was conducted in the same way as in Experiment 1.

Text in Figure 5 flowchart:

- Orally state 1 instruction + present object
  - Participant performs the action (action with vision and touch)
  - Participant wears gloves and performs the action (action with vision but without fine tactile information)
  - Participant wears an eye mask and performs the action (action without vision but with touch)
  - Participant wears gloves and an eye mask and performs the action (action without vision and without fine tactile information)
- Orally state 1 instruction + do not present object
  - Participant remains still (verbal encoding)
- After 10 instructions → 2-minute oral free recall

Figure 5. Flowchart of Experiment 3. Each participant was required to complete tasks under five conditions: action encoding with vision and touch, action encoding with vision but without fine tactile information, action encoding without vision but with touch, action encoding without vision and without fine tactile information, and verbal encoding. When there was no vision, participants were required to wear an eye mask; when fine tactile information was unavailable, participants were required to wear gloves.

4.2 Results

A one-way repeated-measures analysis of variance was conducted on the number of correctly recalled items under each condition; the descriptive statistics are shown in Figure 6. The main effect of encoding method was significant, $F(4, 212) = 19.26$, $p < 0.001$, $\eta_p^2 = 0.27$. Bonferroni-corrected post hoc multiple comparisons showed that the number of recalled items in the action-encoding condition with vision and touch ($M = 6.89$, $SD = 1.69$) did not differ from that in the action-encoding condition with vision but without fine tactile information ($M = 6.83$, $SD = 1.66$) ($p > 0.999$); both were greater than the numbers recalled in the action-encoding condition without vision but with touch ($M = 5.13$,

Figure 6. Bar chart showing the number of correctly recalled instructions in five conditions: visual–tactile action encoding, visual/no fine tactile action encoding, nonvisual–tactile action encoding, nonvisual/no fine tactile action encoding, and verbal encoding. The y-axis is labeled "Number of correctly recalled instructions"; significance brackets are marked with asterisks.

Figure 1: Figure 6. Bar chart showing the number of correctly recalled instructions in five conditions: visual–tactile action encoding, visual/no fine tactile action encoding, nonvisual–tactile action encoding, nonvisual/no fine tactile action encoding, and verbal encoding. The y-axis is labeled "Number of correctly recalled instructions"; significance brackets are marked with asterisks.

$SD = 1.94$), the action-encoding condition without vision and without fine tactile information ($M = 5.24$, $SD = 1.76$), and the verbal-encoding condition ($M = 4.78$, $SD = 1.93$) ($ps < 0.001$). The numbers recalled in the latter three conditions did not differ from one another ($ps \geq 0.999$). This result indicates that the SPT effect appeared only in the two vision conditions (i.e., the action-encoding condition with vision and touch and the action-encoding condition with vision but without fine tactile information).

effect.

To further examine the existence and strength of the SPT effect, Experiment 3 calculated the gain of each action-encoding condition relative to the verbal baseline (i.e., performance in the action-encoding condition minus performance in the verbal-encoding condition) as an index of the SPT effect size. A one-way repeated-measures ANOVA was conducted on the four SPT effect sizes. The results showed a significant main effect, $F(3, 159) = 19.25$, $p < 0.001$, $\eta_p^2 = 0.27$. Bonferroni-corrected post hoc multiple comparisons showed that, under conditions with vision (regardless of whether touch was available), the SPT effect sizes were relatively large (visual–tactile action-encoding condition: $M = 2.11$, $SD = 2.29$; visual/no fine tactile action-encoding condition: $M = 2.06$, $SD = 2.39$), and the two did not differ ($p > 0.999$). Under conditions without vision (regardless of whether touch was available), the SPT effect sizes were relatively small (nonvisual–tactile action-encoding condition: $M = 0.35$, $SD = 2.75$; nonvisual/no fine tactile action-encoding condition: $M = 0.46$, $SD = 2.60$), and the two did not differ ($p > 0.999$). The SPT effect sizes in the visual–tactile action-encoding and visual/no fine tactile action-encoding conditions were both larger than those in the nonvisual–tactile action-encoding and nonvisual/no fine tactile action-encoding conditions ($ps < 0.001$).

Figure 6. Means and standard errors of the number of correctly recalled instructions in each condition in Experiment 3. * $p < 0.05$

4.3 Discussion

By manipulating the availability of vision and fine touch, Experiment 3 separated their respective contributions to the SPT effect for nonroutine instructions. The results showed that, regardless of whether vision was available, deprivation of fine tactile information (such as texture and temperature) did not affect the emergence of the SPT effect, indicating that fine tactile information is not necessary for the SPT effect. In contrast, vision played a crucial role in the SPT effect: only when the visual channel was open did the SPT effect size reach a relatively high level (mean $\approx$ 2.0), whereas visual deprivation significantly reduced the effect size (mean $<$ 0.5). Taken together, the three experiments indicate that, for nonroutine instructions composed of familiar object nouns, the generation of the SPT effect requires participants to interact with real objects (Experiment 2 vs. Experiment 1 and the supplementary experiment), and in this interaction the key···

it is the visual channel, rather than fine tactile input, that plays the key role (Experiment 3).

In addition to dissociating the contributions of vision and fine touch, the results of Experiment 3 also help address a potential alternative explanation for Experiment 2: real objects, compared with pictures, provide stronger affordances and multimodal perception (Gomez et al., 2018; Snow et al., 2023). Thus, the SPT effect observed for nonconventional instructions in the real-object condition of Experiment 2 might not have originated from action execution, but instead from these properties of real objects. If the SPT effect were driven entirely by the affordances of real objects, then under the object-present condition, performance in the action-encoding and verbal-encoding conditions should not differ, because participants in both conditions saw the real objects and the degree of affordance activation should have been the same. However, the data from Experiment 2 showed that recall in the object-present action-encoding condition ($M = 6.69$) was significantly higher than in the object-present verbal-encoding condition ($M = 5.03$). This indicates that action execution provided an additional boost to memory on the basis of affordances. With respect to multimodal perception, in Experiment 3 both the vision-plus-touch condition and the vision-without-fine-touch condition used real objects, and participants executed actions in both; the only difference was the availability of fine tactile information. The SPT effects in the two conditions did not differ significantly. If the advantage of real objects came entirely from their multimodal perception, including the tactile component, then weakening tactile information should have weakened the effect; however, this was not the case. Taken together, these two points indicate that the experimental data do not support using affordances or multimodal perception alone to explain the SPT effect under real-object conditions in Experiment 2; action execution itself is indispensable to this effect. As to which specific components of action execution are at work, Experiment 2 cannot provide a direct answer. Although Experiment 3 separated the sensory components involved in action execution, because Experiment 2 did not manipulate

the visual channel, applying the findings of Experiment 3 retrospectively to Experiment 2 is a matter of logical inference, and its explanatory power is limited. Future research could directly separate vision from kinesthesis—for example, by adding a condition in which real objects are presented but participants only perform simulated actions—to test whether the results of Experiment 2 were caused by visual–kinesthetic coordination or by other processing related to interaction with real objects.

5 General Discussion

Through three experiments, this study examined how the sensory channels on which the SPT effect depends vary with the conventionality of the instructions, focusing on two questions: (1) For instructions composed of familiar object nouns, such as "pencil," do conventional pairings, such as "sharpen the pencil," and nonconventional pairings, such as "poke the pencil," elicit SPT effects under action encoding that show different dependence on sensory channels? (2) Does fine tactile input make an independent contribution to the SPT effect? Experiment 1 found that, for conventional instructions, the SPT effect was present regardless of whether object pictures were presented, indicating that when the verb and noun have a preexisting association, kinesthetic input is sufficient to support the emergence of the SPT effect, without the need for external visual cues. For nonconventional instructions, however, the SPT effect did not appear regardless of whether object pictures were present, indicating that when the verb and noun lack a preexisting association, action execution alone is insufficient to produce a memory advantage. Experiment 2 focused on nonconventional instructions and found that when real, interactable objects were present, the SPT effect emerged; in the supplementary picture condition, which used a within-subject design and the same materials as Experiment 2, no SPT effect was observed. This ruled out the confound of differences in experimental design and further confirmed the key role of object interactivity in the encoding of nonconventional instructions. Experiments 1 and 2 therefore show that the SPT effect for conventional instructions does not depend on visual cues from objects, whereas the effect for nonconventional instructions does. Experiment 3 orthogonally manipulated vision (blindfold) and fine touch (gloves)

input and found that, for nonconventional instructions, the presence of the SPT effect depends on the visual modality (the SPT effect size is larger when vision is available and is markedly reduced when vision is absent), but not on fine-grained tactile information (such as texture and temperature). Taken across all experiments, the sensory-modality dependence of the SPT effect varies with the conventionality of the instruction: the SPT effect for conventional instructions can be driven independently by kinesthesia; for nonconventional instructions, however, the SPT effect requires the coordinated action of kinesthesia and vision, whereas fine-grained tactile information is not necessary in this process.

Why does the SPT effect for nonconventional instructions depend on the coordination of vision and kinesthesia, while fine-grained touch plays almost no role?

According to multimodal processing theory, action enactment adds additional sensory channels on the basis of verbal encoding, thereby deepening the memory trace of the instruction and making it easier to retrieve (Bäckman & Nilsson, 1984). For conventional instructions (e.g., "sharpen a pencil"), the association between the verb and the noun already exists, and action enactment can directly deepen the memory trace of the entire phrase. However, for nonconventional instructions (e.g., "pull the chopsticks"), the verb and the noun lack a preexisting association. Although action enactment can deepen the trace of the verb (e.g., "pull"), the trace of the noun (e.g., "chopsticks") is difficult to strengthen through the action itself and may require the aid of visual information. Studies have found that the visual system can rapidly acquire objects' spatial features and functional information (Contier et al., 2024; Ruotolo et al., 2020). It may therefore be inferred that, during encoding, vision can help individuals identify the long-axis orientation, manipulation endpoints, and functional surfaces of the chopsticks, thereby forming guidance for "how to operate the object" (e.g., identifying the direction and manner of "pulling"). When individuals actually perform the action according to this guidance, visual information and kinesthetic information are integrated with each other; the noun trace may be strengthened by object information provided by vision, the verb trace may be deepened by kinesthetic enactment, and the memory trace of the entire verb–noun phrase may be deepened through integration. If vision is blocked, clear operational guidance may fail to form (e.g., not knowing where to pull from or in which direction to pull); the noun trace, lacking the support of object information, becomes difficult to strengthen, and memory performance is consequently weakened. Unlike vision, fine-grained touch may be neither critical nor sufficient in the process of trace deepening. On the one hand, research shows that when vision is available, the brain, based on long-term experience, has established efficient "vision → touch" predictions (Li et al., 2019). When actual tactile input (e.g., an object' s hardness or texture) is consistent with visual predictions, that signal may be regarded as redundant and suppressed to improve processing efficiency. On the other hand, even in the absence of vision, although touch can convey local physical properties, it cannot provide the object' s overall spatial orientation or manipulation axis, and therefore may have difficulty guiding the setting of action parameters (e.g., being unable to determine the starting point or direction of "pulling"). Research on blind individuals indirectly supports this inference: even among blind individuals whose tactile function, through long-term compensation, far exceeds that of typical individuals, the SPT effect remains weaker than that of sighted individuals (Kormi-Nouri, 1995). This conversely indicates that, even under optimal conditions, touch is still difficult to substitute for the core role of vision in deepening memory traces. Considering that the above explanation is only an inference about the results of Experiment 3 based on multimodal processing theory, future studies could provide more conclusive evidence by testing "whether enhancement of the noun trace necessarily requires the involvement of the visual system" and "whether the visual system is sufficient to independently guide the setting of action parameters."

In sum, the present study reveals that the contribution of sensory modalities to the SPT effect is not constant, but differs with the conventionality of the instruction. These findings help refine the classic multimodal processing theory (Bäckman & Nilsson, 1984). This theory emphasizes

Multichannel information can provide richer encoding ("rich encoding") to enhance memory, but it remains unclear which channels play a key role in the SPT effect and which have only limited effects. In Experiment 1 of the present study, habitual instructions produced SPT effects that did not differ between the no-picture condition (kinesthetic channel only) and the picture-present condition (dual kinesthetic and visual channels); that is, no stronger effect was observed under the dual-channel condition. This result differs from the prediction of multichannel theory that "the richer the channels, the stronger the encoding." This suggests that, for habitual instructions with preexisting associations (e.g., "sharpen a pencil"), the SPT effect does not depend on an increase in the number of channels. By contrast, for non-habitual instructions lacking preexisting associations (e.g., "pull chopsticks"), the results of Experiments 2 and 3 revealed another pattern: kinesthesia alone could not produce an SPT effect; vision and kinesthesia had to work together. However, not every added channel was effective, and the addition of fine tactile information did not further enhance the effect. Thus, for non-habitual instructions, multichannel theory can explain why multiple channels need to work together, but it cannot explain which channels are critical when added. Taken together, the results of the three experiments show that the contribution of a channel is not determined simply by the number of channels, but is related to whether the instruction has a preexisting association: when an instruction has such an association, the kinesthetic channel is sufficient; when it lacks such an association, additional channel support is required. Therefore, multichannel theory needs to move from the view that "multiple channels bring richer encoding" toward a more differentiated view that "channel contributions vary with experience."

At the practical level, this study provides refined application guidelines for the strategy of "action encoding to promote memory." When individuals need to process non-habitual instructions in which there is no preexisting association between the verb and the noun (such as operating a new device or understanding an unfamiliar procedure), bodily action alone is unlikely to effectively improve memory. Visible objects (physical objects or high-fidelity visual presentations) must be provided simultaneously in order to establish visual-spatial anchors and support the online construction of temporary associations. It is worth noting that fine tactile perception is not a necessary condition. In situations where tactile input is restricted, such as when gloves are worn, self-performed actions can still promote memory as long as the visual channel is available. This finding has important applied value: in hygiene-sensitive settings (such as laboratories and kitchens) or among groups with tactile sensitivity/avoidance characteristics, action-encoding interventions can be implemented effectively, and instruction memory improved, simply by ensuring that visual information is visible and that objects are interactive, without relying on fine tactile feedback.

This study also has limitations. First, the present study formed habitual and non-habitual instructions through manipulation of the verb, but this manipulation was accompanied by differences across multiple dimensions, such as semantic plausibility and action naturalness. Although these dimensions were constant within the habitual instructions and within the non-habitual instructions in Experiment 1, and therefore do not affect the conclusions from the comparison between the picture-present and no-picture conditions within each type of instruction, they may affect the generalizability of the results of Experiment 1. In addition, there was also a certain degree of heterogeneity within the non-habitual instructions: some instructions, such as "touch a towel," can occur in specific contexts, whereas "cut a toothbrush" is closer to a semantic anomaly. This internal heterogeneity was balanced between the picture-present and no-picture conditions through random assignment, and therefore does not affect the statistical conclusions of that comparison, but it may limit the conceptual unity of "non-habitual." Future research could further distinguish subtypes of non-habitual instructions and systematically examine the effects of these dimensions on the SPT effect. Second, the present study treated object familiarity as a constant variable (all three experiments used highly familiar everyday objects) in order to examine the independent role of instruction habituality; accordingly, the core conclusions are limited to familiar objects. For unfamiliar objects, whether habituality plays the same role remains to be tested. Future research could, on the basis of the present study, manipulate object familiarity and systematically examine whether familiarity

the role of habituality in multichannel encoding. Third, in Experiment 1 and Experiment 2, the mode of object presentation (picture vs. real object) and the mode of operation (simulated action vs. real interaction) varied simultaneously, making it impossible to fully disentangle the respective contributions of visual intensity under the real-object condition and genuine object–action interaction. Although this does not affect the core conclusions of the present study—that "the interaction between kinesthesis and vision is critical for producing the SPT effect under non-habitual instructions" and that "fine-grained tactile information is not necessary" —it limits a more refined separation of the respective roles of vision and operation on this basis. Future research could add a condition in which "real objects are presented but only simulated actions are performed" to clarify this issue further. Fourth, the present study focused on vision and touch; object presentation may also simultaneously elicit other sensory inputs, such as olfaction and audition (e.g., sounds produced by object collisions), whose potential contributions have not yet been tested. Future research could use paradigms such as odor-controlled rooms and sound-attenuated environments to broaden the dimensions examined in multichannel interaction. Finally, the present study revealed the necessity of vision and kinesthesis for the SPT effect; however, owing to the design of behavioral experiments, it was unable to characterize the collaborative process between the two during real-time processing. Future work could combine eye tracking, action kinematic analysis, or computational modeling to characterize in detail the temporal dynamic coupling between

visual attentional allocation and action execution, thereby further revealing the internal processes by which vision and kinesthesis jointly facilitate memory.

6 Conclusion

(1) The SPT effects of habitual and non-habitual instructions have different sensory-channel dependencies: habitual instructions can be independently driven by kinesthesis, whereas non-habitual instructions require the coordination of vision and kinesthesis.

(2) Fine-grained tactile information makes no independent contribution to the SPT effect.

References

Allen, R. J., Hill, L. J. B., Eddy, L. H., & Waterman, A. H. (2019). Exploring the effects of demonstration and enactment in facilitating recall of instructions in working memory. *Memory & Cognition, 48*(12), 400–410. https://doi.org/10.3758/s13421-019-00978-6

Bäckman, L., & Nilsson, L. G. (1984). Aging effects in free recall: An exception to the rule. *Human Learning: Journal of Practical Research & Applications, 3*(1), 53–69.

Clark-Carter, D. (2018). Subsequent analysis after ANOVA or χ^2. In *Quantitative psychological research: The complete student' s companion* (4th ed.). Routledge.

Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology, 22*(1), 267–281. https://doi.org/10.1111/j.1467-9450.1981.tb00402.x

Contier, O., Baker, C., & Hebart, M. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour, 8*, 2179–2193. https://doi.org/10.1038/s41562-024-01980-y

Engelkamp, J. (1996). Organisation and recall in verbal tasks and in subject-enactment tasks. *European Journal of*

Cognitive Psychology, 8(3), 257–274.

Engelkamp, J., & Zimmer, H. D. (1984). Motor programme information as a separable memory unit. *Psychological Research*, 46(3), 283–299. https://doi.org/10.1007/BF00308889

Gomez, M. A., Skiba, R. M., & Snow, J. C. (2018). Graspable objects grab attention more than images do. *Psychological Science*, 29(2), 206–218. https://doi.org/10.1177/0956797617730599

Howell, D. C. (2013). *Statistical Methods for Psychology* (8th ed.). Wadsworth Cengage Learning

Kormi-nouri, R. (1995). The nature of memory for action events: An episodic integration view. *European Journal of Cognitive Psychology*, 7(4), 337–363.

Kormi-Nouri, R. (2000). The role of movement and object in action memory: A comparative study between blind, blindfolded and sighted subjects. *Scandinavian Journal of Psychology*, 41(1), 71–75.

Li, Y., Zhu, J., Tedrake, R., & Torralba, A. (2019). *Connecting Touch and Vision via Cross-Modal Prediction.* 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10601–10610. https://doi.org/10.1109/cvpr.2019.01086

Ma, J. L., Wang, L., & Li, Y. (2025). Imagery processing is necessary for the subject-performed task effect: Evidence from event-related potentials. *Learning and Motivation*. https://doi.org/10.1016/j.lmot.2024.102091

Mecklenbräuker, S., Steffens, M. C., Jelenec, P., & Goergens, N. K. (2011). Interactive context integration in children? Evidence from an action memory study. *Journal of Experimental Child Psychology*, 108(4), 747–761. https://doi.org/10.1016/j.jecp.2010.12.002

Nyberg, L., Nilsson, L. G., & Bäckman, L. (1991). A component analysis of action events. *Psychological Research*, 53(3), 219–225.

Ruotolo, F., Kalénine, S., & Bartolo, A. (2020). Activation of manipulation and function knowledge during visual search for objects. *Journal of Experimental Psychology: Human Perception and Performance*, 46(1), 66–90. https://doi.org/10.1037/xhp0000696

Snow, J. C., Gomez, M. A., & Compton, M. T. (2023). Human memory for real-world solid objects is not predicted by responses to image displays. *Journal of Experimental Psychology: General*, 152(10), 2703–2712. https://doi.org/10.1037/xge0001387

Wang, L. J., & Li, G. Z. (2014). A review on the theories of subject performed task effect in action memory. *Journal of Psychological Science*, 37(4), 998–1001. [Wang Lijuan, Li Guangzheng. (2014). A theoretical exploration of the SPT effect in action memory. *Psychological Science*, 37(4), 998–1001.]

Wang, L. J., & Ma, J. L. (2023). Imagery encoding in action memory of art students and non-art students.

Psychological Development and Education, 39(4), 457–464. [Wang Lijuan, Ma Jialin. (2023). Representational encoding of action memory in art students and non-art students. *Psychological Development and Education*, 39(4), 457–464.]

Wang, L., J., Xie, T. T., Ma, H., Xu, M., & Xie, X. C. (2022). Subject-performed task effect on working memory performance in children with autism spectrum disorder: The role of intelligence. *Autism Research, 15*(9), 1698–1709. https://doi.org/10.1002/aur.2710

Xie, T. T., Ma, H., Wang, L., J., & Du, Y. F. (2024). Can enactment and motor imagery improve working memory for instructions in children with autism spectrum disorder and children with intellectual disability? *Journal of Autism and Developmental Disorders, 54*(1), 131–142. https://doi.org/10.1007/s10803-022-05780-z

Yang, T. X., Jia, L., Zheng, Q., Allen, R. J., & Ye, Z. (2018). Forward and backward recall of serial actions: Exploring the temporal dynamics of working memory for instruction. *Memory & Cognition, 47*(2), 279–291. https://doi.org/10.3758/s13421-018-0865-x

Yu, Z. Y. (2014). *The influence of physical features on the subject performed task advantage effects in action memory.* (Master's thesis, Jilin University). [Yu Zhanyu. (2014). The role of physical attributes in the SPT advantage effect in action memory (Master's thesis, Jilin University).]

Modality Dependence in the Subject-performed Task Effect for Verbal Instructions: The Moderating Role of Action-Object Pair Conventionality

XIE Tingting[1] YANG Lianyu[1] WU Xiaotian[1] ZHOU Diandian[1] YU Zhanyu[2]

([1]School of Psychology, Fujian Normal University, Fuzhou 350117, China)

([2]School of Education Science, Jiangsu Normal University, Xuzhou 221116, China)

Abstract

The subject-performed task (SPT) effect refers to the finding that action encoding (physically performing described actions) yields better recall of verbal instructions than verbal encoding (merely listening to them). Multimodal theory proposes that this advantage arises because action encoding recruits additional sensory channels beyond the verbal modality, including kinesthetic (motor execution), visual (e.g., observing objects and actions), and tactile input (e.g., touching object surfaces), thereby enriching memory traces and enhancing their retrievability (Bäckman & Nilsson, 1984; Engelkamp & Zimmer, 1984).

Building on this framework, the present study examined how the conventionality of action–object pairings in verbal instructions moderates the multimodal dependencies of the SPT effect, with the aim of distinguishing critical from auxiliary sensory contributions. Three experiments were conducted using familiar everyday objects.

In Experiment 1 ($N = 200$), we manipulated instruction conventionality (conventional, e.g., "sharpen a pencil"; non-conventional, e.g., "poke a pencil")

and the presence versus absence of object pictures (providing visual cues without physical interaction). Results showed that conventional instructions elicited a SPT effect (i.e., performance in the action encoding condition was superior to that in the verbal encoding condition) regardless of whether pictures were present. In contrast, non-conventional instructions failed to produce the SPT effect even when object pictures were available. We hypothesized that static visual input alone may be insufficient to support the online integration of verbs and nouns for non-conventional pairings, and that physical interaction with objects might be necessary. To test this hypothesis, Experiment 2 ($N = 58$) replaced pictures with real, manipulable objects, enabling participants to physically interact with them during encoding. Under these conditions, non-conventional instructions yielded a significant SPT effect when objects were present but not when absent, indicating that sensorimotor interaction, rather than mere visibility, is essential for the SPT effect with non-conventional instructions. However, because real objects concurrently engage both vision and touch, Experiment 2 could not isolate which sensory channel drives this effect. Experiment 3 ($N = 54$) addressed this by orthogonally manipulating visual access (blindfolded vs. unblindfolded) and fine tactile fidelity (wearing gloves vs. bare-handed) during object interaction, with a verbal-only baseline. Results showed that the SPT effect emerged only when vision was available, irrespective of fine tactile condition. Specifically, recall performance was significantly better in both "vision + touch" and "vision + no-fine-touch" action conditions than in the verbal baseline ($ps < .001$), whereas "no-vision + touch" and "no-vision + no-fine-touch" conditions did not differ from verbal encoding ($ps > .999$). These findings demonstrate that vision, rather than fine tactile input, is the necessary channel for the SPT effect with non-conventional instructions.

Collectively, these findings reveal that the sensory channel dependence of the SPT effect varies with instruction conventionality. For conventional instructions, kinesthetic input alone, without visual or tactile support, suffices to drive the SPT effect. For non-conventional

instructions, the SPT effect requires the coordination of kinesthetic and visual input, while fine tactile information plays no independent role. These results refine multimodal theory by demonstrating that channel contributions are not simply additive, where more channels yield greater benefit, but are instead flexibly determined by whether pre-existing action–object associations are available. This advances the understanding of multimodal memory and offers practical guidance for memory support strategies in populations with sensory or neurodevelopmental differences (e.g., autism spectrum disorder), where optimizing visual–motor integration may enhance instruction memory without relying on tactile precision.

Keywords: SPT effect, instruction memory, multimodal processing theory

Note: Figure translations are in progress. See original paper for figures.