ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202602.00097

Superior Statistical Learning Relies on Rejecting Partwords: Postprint

Authors: Wenbo Yu, Tianlin Wang, Dandan Liang, Dandan Liang

Date: 2026-01-27T18:18:54+00:00

Abstract

Speech segmentation in statistical learning is typically assessed using a 2-alternative forced-choice (2AFC) task. Although previous analyses have revealed performance differences among learners, and have suggested that better performance may stem from larger representational differences between target words and part-words in a memory-based model, the underlying learning mechanism driving these differences remains unclear. In the present study, seventy-four participants listened to an artificial language and then completed a 2AFC task and a 7-point Likert-scale familiarity rating task. Based on participants' performance on the 2AFC task, we established a genuine-learning criterion and divided participants into a superior-level group (38% of participants) and a regular-level group. Although both groups exhibited above-chance learning in the 2AFC and familiarity rating tasks, superior learners performed significantly better than regular learners. In addition, a series of linear mixed-effects models showed that the superior- and regular-level groups provided comparable ratings for target words and nonwords; however, superior-level participants rated part-words as less familiar than regular-level participants. These patterns suggest that the only difference contributing to their overall performance on the statistical learning task was their perceived level of familiarity with part-words. This study highlights the importance of investigating statistical learning mechanisms within a memory-based framework and provides a nuanced method for examining individual differences in statistical learning tasks.

Full Text

Preamble

Journal of Psycholinguistic Research (2025) 54:60
https://doi.org/10.1007/s10936-025-10178-w

ChinaRxiv 中X

Superior Statistical Learning Relies on Rejecting Partwords

Wenbo Yu[1] · Tianlin Wang[2] · Dandan Liang[1,3]

Received: 13 September 2023 / Accepted: 31 October 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

Speech segmentation in statistical learning is typically measured by the 2-alternative-forced-choice (2AFC) task. Although previous analysis has found performance differences among learners and better performance might stem from the larger representational differences between target words and partwords given the memory-based model, the underlying learning mechanism that drives these differences remains unclear. In the current study, seventy-four participants listened to a novel language and were then asked to complete a 2AFC task and a 7-point Likert scale familiarity rating task. On the basis of participants' performance on the 2AFC task, we identified a real learning criterion and divided participants into a superior-level group (38% of participants) and a regular-level group. Though both groups exhibited above-chance learning in the 2AFC and familiarity rating tasks, superior learners performed significantly better than regular learners. In addition, a series of linear mixed effect models showed that superior- and regular-level groups produced comparable ratings on target words and nonwords; however, superior-level participants rated partwords as less familiar than regular-level participants. These patterns suggest that the only difference that contributed to their overall performance on the SL task was their perceived level of familiarity of partwords. This study highlights the importance of investigating SL mechanism from a memory-based model and provides a nuanced method for examining individual differences in SL tasks.

Keywords Statistical learning · Real learning criterion · 2-alternative forced task · Familiarity rating task

Introduction

Statistical learning (SL) refers to a set of cognitive abilities that enable learners to detect various types of regularities in the external world. A seminal set of studies by Saffran and colleagues (1996a, 1996b) showed that individuals are sensitive to and capable of processing statistical regularities in continuous speech, which helps them discover word boundaries. Over the past twenty-five years, substantial evidence has established a strong connection between statistical learning ability and a wide range of cognitive skills, especially language processing proficiency, as well as the trajectory of language development (e.g., Graf Estes et al., 2007; Gabay et al., 2015; Potter et al., 2017; Shoaib et al., 2018; Qi et al., 2019; van Witteloostuijn et al., 2021; Isbilen et al., 2022; see

reviews in Saffran & Kirkham, 2018, and Ren et al., 2023). However, what distinguishes good SL learners from other learners remains unclear, posing a challenge to fully understanding SL' s role in language development. In this study, we propose a standardized approach to identify superior statistical learners by comparing their performance with that of their counterparts and examine why these superior SL learners performed better than regular learners.

When assessing individuals' SL performance, researchers traditionally adopted the 2-alternative-forced choice (2AFC) task that requires participants to choose the more familiar item from two alternatives: a target word and a partword (or nonword) (Mirman et al., 2008; Palmer et al., 2019). If the mean accuracy is above chance, it is assumed that learning has occurred. Since the correct item (i.e., target word) must be included as one of the two items in each trial of the 2AFC task, higher-than-chance accuracy can be achieved by either encoding statistical information from the exposure phase or making random judgements at slightly higher than chance level during the test phase. To determine if one responds via knowledge obtained from the learning phase, several researchers have suggested that we should identify the individuals who are able to segment speech only by statistical knowledge with a stricter criterion: a real learning criterion (Batterink et al., 2015a; Siegelman et al., 2017). This criterion should be significantly higher than chance in a one-tailed t-test with a hypothesized direction, thus excluding the possibility of guessing at an above-chance level. Its formula is shown in (1); p represents the chance of correctly answering a trial by chance (0.5), n represents the number of trials in a 2AFC task and $Z_{0\cdot05}$ is commonly set as 1.645 in one tail t-test.

$$\text{real learning criterion} = np + Z_{0.05}$$

Recent re-analyses on SL data have conferred the utility of employing the real learning criterion for grouping –that it separates superior learners from the others. Siegelman and colleagues (2017) reanalysed the data from Siegelman and Frost (2015) and found that nearly 60% of participants were below the criterion, indicating that there was no real learning effect for most participants. Similarly, we re-examined our experimental data in a verbal SL task, where two groups of 30 participants completed the task with different TP levels (Yu et al., 2021). Although both groups showed significant learning outcomes at the group level, only a small proportion of participants (i.e., nine participants in each condition) performed above the criterion of real learning. Moreover, the mean accuracy of this subgroup of participants was around 0.8, which was significantly higher than the rest. These analyses illustrate that individual differences may be ubiquitous in 2AFC-assessed SL research, and that the real learners outperformed others. However, it remains unclear what behavioral differences drive such a distinction –what are the differentiating factors between superior SL learners and others when they participate in such a SL task?

Many SL tasks, including the 2AFC task, are off-line tasks primarily designed to

test participants'ability to recognize words segmented during the exposure phase. Related studies have introduced some memory-based models to explain how participants process and store the syllabic units (Thiessen et al., 2013; Thiessen, 2017; Saffran & Kirkham, 2018). Recently, Isbilen et al. (2020) proposed a chunking-model, suggesting that the statistical learning effect is a consequence of chunking mechanisms. Since both a target and a foil are presented in each test trial in a 2AFC task, participants need to compare their familiarity or memory representation of each alternative item in order to decide. The forced-choice paradigm may require meta-cognitive decision processing (Christiansen, 2019), thus successful SL performance is the function of both memory representation of target words and partwords. Based on these analyses, it is plausible that larger representational differences between target words and partwords in the exposure phase are encoded by superior learners.

In line with this idea, the current study aimed to explore the underlying memory patterns associated with two types of words. Specifically, we sought to determine whether the advanced accuracy displayed by superior learners stems from a more robust memory representation of target words, a weaker memory representation of partwords, or a combination of both factors. Any of these three scenarios could lead to a more significant distinction between target words and partwords among superior learners. To examine these behavioral differences across various levels of learners, one approach involves considering human participants as decision-makers tasked with categorizing presented test alternatives into either target words or foils. For superior learners, there are three plausible pathways for making such decisions: First, they may exhibit a greater familiarity with target words compared to regular learners. Consequently, they are more precise and confident when identifying target words in a 2AFC task; Second, and conversely to the first possible pathway, it is possible that they have a significantly lower level of familiarity with partwords compared to regular learners. This lower familiarity enables them to reject partwords with higher accuracy, even if their familiarity with the target word is on par with that of regular learners. The third possibility combines the first two scenarios: superior learners may demonstrate both a stronger familiarity with target words and a lower familiarity with partwords. In summary, this study delves into the memory patterns underpinning word recognition and decision-making abilities among learners of different proficiency levels.

Additionally, most previous studies tested SL with languages consisting of equiterminous words (e.g., all words are disyllabic or trisyllabic). This type of material, however, has been criticized as simple and monotonous artificial languages with words of the same length do not accurately reflect learners'experience in the real world (Erickson & Thiessen, 2015; Frost et al., 2020). Johnson and Tyler (2010) found that infants are not able to segment a continuous speech which was created with different length of word even after a long exposure time. Moreover, adult participants' learning effect in 2AFC task decreased dramatically in a mixed length of language compared with equal length artificial language (Hoch et al., 2013). Considering these aspects, the current study investigates how su-

perior learners versus regular learners encode regularities from a mixed-length artificial language created with trisyllabic and disyllabic words.

In the current study, we investigated two groups of learners' performance on target words and partwords by adding a familiarity rating task. If superior learners rely on high familiarity of target words, they should produce higher rating scores on target words, but similar scores on partwords compared with less successful learners. In contrast, a different pattern should be observed if superior learners have lower familiarity of foils: they should rate target words similarly compared to less successful learners; whereas lower familiarity rating on partwords should be reported by superior learners compared to their less successful counterparts. Finally, a pattern of both higher score for target words and lower scores for partwords from superior participants can be predicted by the third hypothesis.

Method

Participants

Seventy-four native Mandarin speakers (56 females, mean age = 21.00 years, range: 18–27) were recruited from a Mandarin-speaking region in Southeast China. Participants were randomly assigned to one of two orders of presentation (order A: 2AFC task first, n = 36; order B: Rating task first, n = 38). Two additional participants in order A were excluded due to technical issues during testing. Participants were screened for foreign language as well as musical experience. Only those who self-reported as not having had any significant exposure to a foreign language or musical training were included in the experiment. After the test, all participants were compensated with monetary payment for their time. The experiments were approved by the Institutional Review Board (NNU 2,022,060,023).

Materials

The artificial language consisted of both syllabic and tonal tiers. For SL studies that employ artificial languages, it is widely accepted that the word units should be nonsensical in order to avoid semantic interference (Palmer et al., 2019; Saffran et al., 1996a, 1996b). To achieve this, in the current study, we wanted to ensure that the syllables are both phonologically legal and semantically empty to the participants.

The Mandarin phonological system includes a total of 413 syllables and 4 lexical tones (i.e., Tone 1, Tone 2, Tone 3, and Tone 4). They can then combine into about 1522 tonal syllables. For instance, the syllable /ma/ can be combined with all four tones and results in ma1 (mother), ma2 (hemp), ma3 (horse), and ma4 (scold). However, not all syllables will result in meaningful tonal syllables when combined with a certain tone. For example, /se/ can be combined with Tone 4 (i.e., se4, colour), but when it is combined with Tone 1, the tonal syllable does

not have any semantic meaning. Following this logic, we identified 10 syllables that can be combined with Tone 1 and result in a nonsensical tonal syllable. This form of design can also ensure that the participants' Mandarin language background does not hinder their processing of an unfamiliar language, and it also removes any potential interference of TP information at the tonal tier (see Gómez et al., 2017 for a discussion). Syllables were recorded in a sound-attenuating room to digital format at 44,100 Hz with 16bit precision.

A female native Mandarin speaker produced all syllables with natural speaking rate and monotone speaking style. In order to preserve coarticulation and ensure that no boundary cues such as pause or stress were unintentionally inserted, the speaker was asked to produce a list of syllable triplets with the target syllable in the middle position. The target syllables were then isolated and normalized for duration (300 ms), mean pitch (266 Hz), and intensity (70 dB) via Praat software (http://www.praat.org/).

Unlike classic SL research which tends to utilize nonsensical words of equal length (e.g., Sohail & Johnson, 2016; Wang & Saffran, 2014), we created four nonsensical words (two trisyllabic words and two disyllabic words) of different lengths using the 10 tonal syllables. Two counterbalanced versions of the artificial language were created to control for arbitrary preferences in learning languages. The test items which were target words in Language A were partwords in Language B, and vice versa. Table 1 shows the target word, partword and nonword test items for each language. In each experiment order, half of the participants learned language A and the other half learned language B.

To create the two versions of artificial language, each word token was played 120 times, resulting in a total of 480 words. The language was concatenated by Praat script in a pseudorandom sequence, which ensured that there were no immediate repetitions of the same word. The transitional probability between syllables forming a target word was 1.0 and the transitional probability between syllables spanning word boundaries was 0.33. In addition, the number of each word was balanced between the first and second half of artificial language. In doing so, we hope to make sure the TPs between words through the whole language would be more well-distributed, which prevented the constant repetition of certain words at the beginning or end of the language. The resulting artificial language lasted about 6 min.

In the test phase, there were four partwords and nonwords for each artificial language. Only partwords paired with target words in the forced-choice task, while both partwords (the same as forced-choice task) and nonwords appeared in the familiarity rating task. Like target words, there were two types of lengths of partwords (i.e., trisyllabic and disyllabic test items), and a partword was paired with a target word of equal length in each trial in order to rule out any preference for word length in the test phase. Partwords contained syllables from the adjacent target words. For example, partword tediafo was made up of the last syllable of the target nueruote and the first two syllables of another target diafolai, leading to a TP of 0.33 between the first two syllables. Nonwords

contained syllables from different words (e.g., nuemeilai, diasete, refo, rouruo) that never co-occurred during the exposure. The within word TPs for nonwords were therefore always 0 since these syllable sequences were novel as far as the familiarization speech was concerned.

Procedure

After the consent procedure, participants were tested individually in a quiet room. Prior to the familiarization phase, participants were instructed to concentrate on listening to an alien language for about 6 min and told that they would be later tested on their knowledge of the language. Additionally, a short practice session with an artificial language fragment which is similar to the exposure material and a concise version of both the 2AFC and the familiarity rating tasks were carried out before the formal experiment. Participants then began to listen to auditory stimuli via a headphone at 30 dB. Immediately after listening to the artificial language, they entered the test phase and were asked to finish a 2AFC task and a 7-point Likert-type scale rating task. The order of these two tasks was counterbalanced across participants.

In each trial of the 2AFC task, a partword and a target word were sequentially presented auditorily, with a 500 ms pause between them. As soon as a trial was presented, participants saw a written instruction that prompted them to judge which item was more familiar by pressing key "1" (indicating the first item) or "2" (indicating the second item) using the keyboard with no time limit. The next trial began after a response was recorded. There were 16 trials in total for the 2AFC task comprising of target words and their corresponding partwords. The pairing and order of presentation of the target word-partword dyads were counterbalanced across participants, with eight trials for disyllabic contrasts and eight for trisyllabic contrasts.

In the familiarity rating task, participants rated how familiar a sound sequence was using a seven-point Likert-type scale (with "7"indicating "extremely familiar" and "1" indicating "not at all familiar"). There were 12 test items in total, with 4 items for each type of word (i.e., target words, partwords and nonwords). The order of presentation was randomized across participants, and participants were asked to assess the familiarity of each item with no time limit. The next trial began after a response was recorded. The entire experiment took about 20 min to finish (see Fig. 1 [Figure 1: see original paper]).

Grouping: Superior Versus Regular Learners

Our primary motivation was to investigate whether there exists a different memory representation in participants' SL ability as measured by word segmentation tasks. In order to identify superior learners, we divided participants based on their performance on the 2AFC task. Combining the suggestion from Batterink et al. (2015a) and the trials presented in our 2AFC task, the real learning criterion for an individual was set to 12 correct responses or more[1] –that only

when participants answered at least 12 items correctly could we determine that their performance was not based on guessing. As a result, 28 out of 74 (38%) participants performed over this real learning criterion and were referred to as superior learners; the rest of the participants were referred to as regular learners (see Fig. 2 [Figure 2: see original paper]).

Results

Effect of Task Order

In order to rule out order effects (order A: 2AFC task first, order B: Rating task first), a generalized linear mixed-effects model (GLMM, Jaeger, 2008) for the 2AFC task and a linear mixed-effect model (LMM) for the familiarity rating task were carried out from the afex package in R (R Development Core Team, 2018). While there were no order effects on the 2AFC task ($F(1, 72.04) = 15.84$, $p < 0.001$), there was a significant order effect on the rating task, with participants finishing the rating task first rated higher familiarity scores on all types of words ($F(2, 21.01) = 46.47$, $p < 0.001$) and a significant word type effect ($F(2, 21.01) = 46.47$, $p < 0.001$); there was no significant interaction between experimental order and word type ($F(2,798.54) = 1.26$, $p = .28$). Taken together, although experimental order affected the familiarity ratings, the differences were not a function of word type, implying little influence on the aim of this study.

Overall Performance in 2AFC Task and Familiarity Rating Task

Then we conducted a series of planned one-sample t tests between the participants' average accuracy and chance level (0.5). Results revealed a significant learning effect: $t(73) = 4.84$, $p < .001$, $d = 0.56$, $M = 0.61$, 95%CI: [0.57,0.66]. As a group, participants performed above chance and chose the correct targets more often than the foils. Superior learners chose target words more than partwords ($t(27) = 25.73$, $p < .001$, $d = 4.86$, $M = 0.81$, 95%CI: [0.79, 0.84]), while regular learners showed no preference for target words over partwords ($t(45) = -0.60$, $p = .56$, $d = -0.09$, $M = 0.49$, 95%CI: [0.44, 0.53]). Importantly, the accuracy of superior learners was significantly higher than that of regular learners, $t(68.07) = 13.59$, $p < 0.001$, $d = 2.77$, 95%CI = [0.37, 0.28]. In sum, when separated into the superior versus regular learners based on the criterion for real learning, only the superior learners exhibited real SL learning (see Fig. 3a [Figure 3: see original paper]).

To examine the learning effect in the rating task, we ran a LMM where word type (target word, partword and nonword) was entered as a fixed-effect factor and both participant and item were included as random intercepts. The model yielded a significant main effect for word type ($F(2,21.59) = 44.78$, $p < .001$). Specifically, target words elicited higher familiarity ratings than partwords, which in turn were associated with higher ratings than nonwords (see Fig. 3b). Pairwise comparisons were adjusted using the Bonferroni correction (Table 2).

To test the correlation between the scores from two task, the score difference between the target word and the partword in familiarity rating task, was calculated for each participant. Then Pearson correlation analysis showed that the scores of the two tasks correlated significantly, with a moderate effect size, r(74) = 0.45, 95%CI = [0.25, 0.62], p < 0.001 (see Fig. 3c). Finally, three correlation analyses were conducted for each of the three types of words between their 2AFC accuracy and rating scores: accuracy only showed a marginal significance with the ratings of target words (r(74) = 0.24, 95%CI = [0.02, 0.45], p = 0.07) and partwords (r(74) = -0.23, 95%CI = [-0.44, -0.01], p = 0.07), after FDR correction.

Different Learning Mechanisms Between Superior and Regular Learners

To examine learning outcomes between the two groups, a LMM with maximal random effect structure was fitted to the data. The model included groups and word types as fixed factors, and only random intercepts for participants and items as the random effects; the model revealed a significant main effect of word type ($F(2,21.93) = 47.46$, $p < .001$), a non-significant main effect of group ($F(2,71.97) = 1.65$, $p > .05$) and an interaction between word type and group ($F(2,793.20) = 7.57$, $p < .001$). Results of simple effect analysis after Bonferroni correction are summarized in Table 3 .

Critically, superior learners rated partwords significantly less familiar in comparison to the ratings generated by regular learners on the same items, while there were no significant rating differences on target words. These results suggest that the key component that distinguishes the 2AFC performance and learners' level is partwords (see Fig. 3d [Figure 3: see original paper]).

Better Discrimination Ability for Superior Learners

An alternative account to the notion that superior SL learners produced lower familiarity ratings on partwords is that they were more cautious when making choices in a 2AFC task and employed a more stringent criterion for decision making. To rule out this possibility, we computed the index of discriminability and likelihood ratio based on Signal Detection Theory (SDT). In Batterink and Paller' s (2017) study, partwords and nonwords rated as 3 or 4 (on a 4-point Likert-type scale) were tagged as false alarms; those rated as 1 or 2 were deemed as hits. Similarly, target words rated 3 or 4 were seen as hits while those rated as 1 or 2 were false alarms. In reference to this classification criterion, we applied four as dividing point in our 7-point Likert-type scale target words rated as 4 to 7 were regarded as hits while those rated as 1 to 3 were false alarms; and foils (i.e., partwords and nonwords) rated as 1 to 3 were seen as hits while those rated as 4 to 7 were seen as false alarms. An index of discriminability (d'), which reflects a participant' s ability to distinguish between signal and noise, and likelihood ratio (β), which reflects the participant's subjective criteria

when making judgements, were computed respectively for each participant in the study. The formats are as followed:

$$d' = Z_{hit} - \beta = O_{hit} - Z_{falsealarm}O_{falsealarm}$$

If superior learners succeeded in word segmentation because they were better at discriminating the items from each other, they should show significant higher d' value but not β value; if superior learners used a different decision-making criterion from regular learners, they should have higher β value but not d' value. Independent samples t-test only revealed a significant difference on d' (t(50.78) = -3.10, p < .01, M(superior) = 1.05, M(regular) = 0.17, 95%CI: [-1.47, -0.31]), but not on β (t(69.91) = 1.40, p = .16, M(superior) = 1.19, M(regular) = 2.25, 95%CI: [-0.45, 2.58]). These results thus ruled out the explanation that the advance performance on the 2AFC task comes from a stricter criterion for making choices for the superior learners.

Discussion

The current study employed a 2AFC task and a familiarity rating task to examine individuals' SL ability when learning a complex artificial language. Results revealed that participants with higher scores in the 2AFC task produced lower familiarity ratings on partwords than those with lower accuracy in the 2AFC task, while the scores of both groups are comparable on target words and non-words. This pattern of familiarity rating score implies that learning effect on partwords plays a potential role in distinguishing good and poor SL learners.

Real Learning Criterion

The use of 2AFC tasks has proven effective in testing SL effects at the group average level. However, due to the influence of factors such as guessing, evaluating individual differences in learning effects becomes challenging (Batterink et al., 2015a; Isbilen & Christiansen, 2022; Siegelman et al., 2017). While some studies have mentioned the use of real learning criterion to identify different levels of SL learners, none have explored the different learning mechanisms using this as a grouping standard. In this study, regular learners' accuracy in the 2AFC task was not significantly higher than chance level and was notably lower than that of superior learners. By analysing the participants distribution, a total of 28 participants in this study were identified as superior learners, accounting for approximately 40% of all participants. This finding is consistent with previous research, which also reported that superior learners accounted for around 30% to 40% of all participants. For example, in the study by Yu et al. (2021), 33% of participants (9 of whom were superior learners, with a total of 30 participants) exceeded the real learning criterion; the research results of Siegelman et al. (2017) are also inline with the current study, with approximately 40% of participants meeting or exceeding the criterion. In conclusion, based on the average learning effect of the group and the distribution of participants in the two

groups, the current study adds to the existing evidence that the real learning criterion can, to some extent, be used to distinguish superior statistical learners from regular statistical learners.

The Mechanism Underlying Statistical Learning of Superior Learners

In the familiarity rating task, superior learners rated partwords as less familiar than regular learners, whereas superior learners rated target words as equally familiar as regular learners. Meanwhile, the significant differences of index of discriminability (d') and non-significant likelihood ratio (β) between the two groups provided further evidence for a sensitivity account rather than a stringent standard account —the superior learners did not employ a stricter standard in making decisions in the 2AFC task. Hence, the dissociable ratings between target words and partwords across the two groups pointed to the second hypothesis that superior learners succeeded in a verbal SL task because of their weaker familiarity to partwords rather than a stronger representation of target words.

Statistical learning performance is often seen as an automatic process that involves breaking continuous speech and storing word candidates in memory. Model simulations have show that learners, while encoding the syllables of an artificial language, also segment speech into small and distinct parts of varying lengths (Perruchet & Vintner 1998). According to Perruchet (2019), these provisional units are temporarily stored in a lexicon and subject to memory constraints, where the strength of their memory representation is positively correlated with the number of occurrences in the exposure phrase. Units that do not recur will decay in the memory system. Building on these arguments, relevant research suggests that learning performance on target words is a core variable that reflects the nature of SL as an implicit memory mechanism (Christiansen, 2019; Isbilen et al., 2020). Consequently, the initial hypothesis was that superior learners would rate target words as more familiar, indicating a more efficient implicit learning process than regular learners, leading to higher accuracy in 2AFC task.

However, the equal familiarity rating on target words suggests that both groups of participants relied on a similar implicit learning process, and the distinguishing factor between superior and regular learners lies in their rating scores of partwords. These findings suggest that familiarity ratings on partwords, ranging from low to high, might indicate individual differences in SL ability. As discussed in the introduction, the learning effect represents the memory process during exposure time. Thus, the different rating patterns on partword imply that the two groups of learners differ in their cognitive processes for this type of words. Based on these considerations, the present study aimed to determine the unique contribution of the learning effect of partwords in distinguishing good and bad SL learners.

To the best of our knowledge, only a few studies have explored the learning effect and underlying mechanism of partwords in SL. Traditionally, SL learner' s repre-

sentation of target words, partwords, and nonwords should gradually decrease in memory strength, as what Batterink and Paller (2017) revealed. However, one study has found that participants' scores of 2AFC task (words compared with partwords) was not significant higher than chance level (Lukács et al., 2023). Therefore, the underlying mechanisms of partwords is more complex than previously thought. When explaining the reason of different learning performances on partwords between superior and regular learners, we hypothesized that it involves an explicit inhibition mechanism in at least verbal SL.

Firstly, prior research has indeed demonstrated that explicit mechanisms significantly contribute to SL performance. For instance, the ability to segment based on statistical regularities is greatly compromised when attentional resources are depleted, such as when performing a concurrent task like two-back rhyme (Palmer & Mattys, 2016; Toro et al., 2005). Additionally, in studies where participants learned the same auditory material, those in the explicit group showed a larger P300 response to predictable targets compared to participants in the implicit group (Batterink et al., 2015b). These findings suggest that SL utilizes domain-general resources to actively update the syllable units segmented during the exposure phase. Moreover, inhibitory control has been identified as a key cognitive ability that allows individuals to resist interference during the statistical learning of a new language (Bartolotti., et al., 2011; Guo et al., 2021). Superior learners in the current study might be skilled in utilizing this explicit mechanism, enabling them to recognize that partwords are not the learning goal, resulting in a lower familiarity rating for those items. In other words, their ability to inhibit the processing of irrelevant information might contribute to their superior performance in the SL task.

Secondly, the 2AFC task implemented in our study also arguably encompasses and reflects explicit knowledge of SL through deliberate recognition since it asks participants to make an explicit choice between a target word that embodies the relevant statistical regularities, and a foil that does not (Isbilen, et al., 2020). This explicit instruction would equip participants with information that the familiarization stream contained discrete constituents. Especially for superior learners, they are not only capable of locating the target words, but might also be able to identify the partwords as implausible tokens, thus lowering their familiarity consciously. Hence, the way we grouped the participants possibly reflects different amounts of explicit knowledge and suggests that superior learners have enough cognitive resource to eliminate partword-related memory representation intentionally and therefore reject partwords to a greater degree in 2AFC trials consisting of a target word and a partword. While the focus of this study is to shed light on learners' different sensitivity to TPs and not on attention allocation or metacognition, a follow-up study using methods such as that of Ordin and Polyanskaya (2021) could potentially provide a direct answer regarding the cause of rating patterns.

Segmenting Complex Speech

Artificial languages comprised of equal-length, nonsensical, words are typically utilized to assess participants' SL ability. However, this design method has drawn criticism regarding its lack of ecological validity vis-à-vis reflection of the complex linguistic environment and whether results from these oversimplified and highly artificial designs can illuminate humans' real-world SL ability. As a step towards creating a more complex set of learning materials, we mixed nonsensical words of two lengths in a synthesized speech stream and measured the learning performance. Results from both tasks showed a significant learning effect, suggesting that learners were able to tolerate this more challenging artificial language and that future design can take advantage of learners' capability by utilizing mixed-length materials. This significant learning effect in the present study is consistent with the two other adult studies that investigated SL ability with artificial languages of mixed length words (Yu et al., 2021; Hoch et al., 2013). It is important to note that the effect size of all learners' SL effect in the 2AFC, as measured by the t-test, is approximately 0.6. This finding is consistent with the results of a recent meta-analysis study (see Fig. 3, Isbilen & Christiansen, 2022). These results from the current study provide further support of the learnability of mix-length words in an auditory SL task as assessed by word segmentation. Moreover, they emphasize the feasibility and potential of incorporating such features in future research.

Limitation and Directions

One limitation of the current study is related to the experiment design. Originally, we intended to split all learners into two groups using the real learning criterion and then ask them to finish a completely new SL task. However, the scarcity of nonsensical syllables in Mandarin presented a challenge, making it difficult to find 20 nonsensical syllables with tone 1. Furthermore, the low correlation between different modalities of SL, as pointed out by Siegelman and Frost (2015), made it impractical to combine SL tasks from two modalities. As a result, the current design was chosen to reveal the rating pattern across target words and partwords among the two levels of learners.

Based on the findings, one area of great value for future research could be to explore ways of measuring SL ability in a more concise manner. Although recent SL research has shorten the exposure time for child participants, such as using 3-min exposure in the study by Arnon (2020), the trials in the test phase still consisted of 25 items. If familiarity with partwords can be shown to be a reliable indicator of SL ability, this could significantly reduce the test time and improve the efficiency of SL task. In the current study, we only detected the different patterns of rating scores between the two groups of learners, but we did not explore whether rating scores on partwords could serve as a unique index for evaluating individual' SL ability. This is an important issue that requires further investigation in future research.

Conclusion

In sum, our results offered a nuanced view into the encoding mechanisms of SL: the learning effect of partwords may differ for superior- vs. regular-level learners. This study also suggests that partwords is likely to be an effective index to discriminate SL ability, highlighting an avenue for future research on individual differences as well as statistical learning mechanisms.

Funding

The article was supported by Social Science Foundation of Jiangsu Province Higher Education Institutions (2022SJYB2051) and Initial Scientific Research Fund of Nanjing Normal University (184080H202A121).

Declarations

Conflict of interest The authors have no financial or non-financial interests to disclose.

Ethical Approval The experiments were approved by the Institutional Review Board (NNU2022060023).

Informed Consent All participants understanded the purpose of the experiment before the experiment and sign informed consent form.

References

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, 52, 68–81.

Bartolotti, J., Marian, V., Schroeder, S. R., & Shook, A. (2011). Bilingualism and inhibitory control influence statistical learning of novel word forms. *Frontiers in Psychology*, 2, 324.

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, 90, 31–45.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015a). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, 83, 62–78.

Batterink, L. J., Reber, P. J., & Paller, K. A. (2015b). Functional differences between statistical learning with and without explicit training. *Learning & Memory*, 22(11), 544.

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, 11, 468–481.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, 37, 66–108.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2020). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 145(12), 1128–1153.

Gabay, Y., Thiessen, E. D., & Holt, L. (2015). Impaired statistical learning in developmental dyslexia. *Journal of Speech, Language, and Hearing Research*, 58, 934–945.

Gómez, D. M., Mok, P., Ordin, M., Mehler, J., & Nespor, M. (2017). Statistical speech segmentation in tone languages: The role of lexical tones. *Language and Speech*, 61(1), 84–96.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? statistical segmentation and word learning. *Psychological Science*, 18(3), 254–260.

Guo, C., Tsegaye, A., Arató, J., & Logemann, H. N. A. (2021). The role of attention, inhibition and statistical learning in Chinese character recognition by novices. *Current Research in Behavioral Sciences*, 2(2021), Article 100012.

Hoch, L., Tyler, M. D., & Tillmann, B. (2013). Regularity of unit length boosts statistical learning in verbal and nonverbal artificial languages. *Psychonomic Bulletin & Review*, 20(1), 142–147.

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical learning of language: A meta-analysis into 25 years of research. *Cognitive Science*, 46(9), Article e13198.

Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition*, 225(2022), Article 105123.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*. https://doi.org/10.1111/cogs.12848

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Development Science*, 13(2), 339–345.

Lukács, Á., Dobó, D., Szőllősi, Á., Németh, K., & Lukics, K. S. (2023). Reading fluency and statistical learning across modalities and domains: Online and offline measures. *PLoS ONE*, 18(3), Article e0281788.

Mirman, D., Magnuson, J. S., Estes, K. G., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, 108(1), 271–280.

Ordin, M., & Polyanskaya, L. (2021). The role of metacognition in recognition of the content of statistical learning. *Psychonomic Bulletin & Review*, 28, 333–

340.

Palmer, S. D., Hutson, J., White, L., & Mattys, S. L. (2019). Lexical knowledge boosts statistically-driven speech segmentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(1).

Palmer, S. D., & Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *Quarterly Journal of Experimental Psychology*, 69(12).

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. *Topics in Cognitive Science*, 11(3), 520–535.

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of Memory & Language*, 39(2), 246–263.

Potter, C. E., Wang, T., & Saffran, J. R. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cognitive Science*, 41(S4), 913–927.

Qi, Z., Sanchez Araujo, Y., Georgan, W. C., Gabrieli, J. D., & Arciuli, J. (2019). Hearing matters more than seeing: A cross-modality study of statistical learning and reading ability. *Scientific Studies of Reading*, 23(1), 101–115.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Ren, J., Wang, M., & Arciuli, J. (2023). A meta-analysis on the correlations between statistical learning, language, and reading outcomes. *Developmental Psychology*. Advance online publication.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, 69(1).

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory & Language*, 35(4), 606–621.

Shoaib, A., Wang, T., Hay, J. F., & Lany, J. (2018). Do infants learn words from statistics? Evidence from English-learning infants hearing Italian. *Cognitive Science*, 42(8), 3083–3099.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, 49(2), 1–15.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81, 105–120.

Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, 12(2), 105–115.

Thiessen, E. D., & Erik, D. (2017). What' s statistical about learning? insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2016.0056

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, 139(4), 792–814.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, 97(2), B25–B34.

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2021). The contribution of individual differences in statistical learning to reading and spelling performance in children with and without dyslexia. *Dyslexia*, 27(2), 168–186.

Wang, T. L., & Saffran, J. R. (2014). Statistical learning of a tonal language: The influence of bilingualism and previous linguistic experience. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2014.00953

Yu, W., Wang, L., Qu, X., Wang, T., Zhang, J., & Liang, D. (2021). Transitional probabilities and expectation for word length impact verbal statistical learning. *Acta Psychologica Sinica*, 53(6), 565–574.

Publisher' s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

---

[1] Based on formula calculation learning criterion $0.5 + 1.645 \times \sqrt{(0.5)} = 11.29$ trials

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

Source: ChinaXiv —Machine translation. Verify with original.