Reliability of Statistical Learning Ability Measurement: Evidence from Modality, Material Characteristics, and Test Tasks

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

Date: 2025-01-02T00:00:00+00:00

Abstract

Statistical learning ability is frequently employed as an independent variable to predict the development of individual language abilities; however, experimental tasks designed to target group differences generally suffer from low reliability, making it difficult to satisfy the fundamental requirements of psychometrics. The present study synthesized learning materials using target structures of mixed lengths, utilized two test tasks—two-alternative forced-choice and familiarity rating—and designed two modalities: auditory speech and visual graphics. The internal consistency coefficient and split-half reliability of the tests were computed. The results demonstrated that the reliability of experimental tasks constructed using mixed-length target structures was generally superior to that of previous studies; concurrently, reliability under the visual modality was higher than under the auditory modality, and the reliability of the forced-choice task surpassed that of the familiarity rating task. In conclusion, this study recommends employing learning materials synthesized with mixed-length target structures under the visual modality and assessing participants' statistical learning ability using forced-choice tasks.

Full Text

A Reliability Study of Statistical Learning Ability Measurement: Evidence from Modality, Material Features, and Testing Tasks

YU Wenbo[1], QI Hetong[1], WANG Tianlin[2], LIANG Dandan[1,3]

[1]School of Chinese Language and Culture, Nanjing Normal University, Nanjing 210097
[2]School of Education, University at Albany, State University of New York, New

York 12222
[3]Interdisciplinary Research Center for Linguistic Science, University of Science and Technology of China, Hefei 230026

Abstract

Statistical learning (SL) ability is frequently employed as an independent variable to predict individual differences in language development. However, experimental tasks originally designed to detect group differences typically suffer from low reliability, failing to meet basic psychometric standards. To investigate optimal measurement approaches for SL ability, this study synthesized learning materials using mixed-length target structures and compared two testing tasks: a two-alternative forced-choice task and a familiarity rating task. Additionally, we implemented both auditory speech and visual graphic modalities. Reliability was assessed through internal consistency (Cronbach's alpha) and split-half reliability. Results demonstrated that tasks employing mixed-length target structures achieved higher reliability than previous studies using uniform-length structures. Furthermore, reliability was superior in the visual modality compared to the auditory modality, and the forced-choice task yielded higher reliability than the familiarity rating task. In summary, this study recommends assessing SL ability using forced-choice tasks with mixed-length target structures presented in the visual modality.

Keywords: statistical learning ability; reliability; forced-choice task; familiarity rating task

1. Introduction

Using basic cognitive abilities as independent variables to predict higher-level cognitive functions represents a common research approach in psychology. With increasing demands for rigorous methodology, the scientific validity and measurement precision of experimental paradigms have drawn growing scholarly attention. Researchers have noted that traditional cognitive experiments often exhibit low reliability when measuring specific cognitive abilities, falling short of psychometric requirements for validity and reliability (Hedge et al., 2018). In the field of psycholinguistics, statistical learning ability is considered a fundamental cognitive capacity closely linked to language acquisition processes such as spoken word segmentation and lexical-semantic learning (Xu et al., 2020; Bogaerts et al., 2020; Estes et al., 2007, 2015; Newport, 2016; Saffran & Kirkham, 2018; Siegelman, 2020). Traditional statistical learning paradigms were designed from a group-differences perspective, focusing on whether group means exceed a certain criterion (e.g., one-sample t-tests) or differ significantly between groups (e.g., independent samples t-tests). When applied to individual-differences research (e.g., regression or correlation analyses), these paradigms suffer from low test reliability and unstable ability estimation, leading to inconsistent findings

regarding whether SL ability predicts language development (Lammertink et al., 2020). Several studies in the SL literature have examined the reliability of using traditional SL tasks to predict language development from a psychometric perspective (Siegelman et al., 2017; Siegelman et al., 2018a). The present study modifies traditional measurement approaches and validates their effectiveness, aiming to improve SL ability assessment and draw broader attention to the psychometric properties of cognitive experiments.

Statistical learning refers to the process by which individuals discover statistical regularities from temporal and spatial input to acquire new knowledge (Yu et al., 2021a, 2021b; Frost et al., 2020; Isbilen & Christiansen, 2022; Saffran et al., 1996). The classic SL task, introduced by Saffran et al. (1996), employs a learning-test paradigm where learning materials consist of four equi-length target words (e.g., three syllables each, with uppercase letters representing syllables as shown in [Figure 1: see original paper]) concatenated pseudo-randomly, with each target word appearing 45 times. During testing, participants listen to target words and part-words, with learning inferred from differential attention times. Subsequent research with children and adults has adapted this learning-phase material, typically using forced-choice tasks in the test phase (see Isbilen & Christiansen, 2022). Each trial presents a target word paired with a part-word (e.g., CJK) or non-word (BHE), requiring participants to identify the basic units comprising the learning material. Since part-words occur at boundaries between target words, they lack strong memory traces, whereas syllable combinations within target words remain consistently co-occurring, yielding more robust memory representations. Significant learning is inferred when group accuracy exceeds chance (0.5).

In recent years, individual-differences research has used accuracy rates from forced-choice tasks as indices of SL ability to predict language development in typically developing children and explain language deficits in clinical populations (Erickson et al., 2016; Isbilen et al., 2022; Kidd & Arciuli, 2016; Kidd et al., 2020; von Koss Torkildsen, 2019). Despite numerous significant findings, tasks designed for group comparisons exhibit low reliability. As summarized in , most studies fail to meet the minimum psychometric standard of 0.8 (Nunnally & Bernstein, 1994). Siegelman (2017) identified two key problems with group-difference paradigms: (1) insufficient test trials (typically only 16), and (2) consistent difficulty across trials because they always pair part-words with target words. These factors constrain score variance, resulting in low reliability for correlation-based analyses. Moreover, forced-choice tasks require repeated presentation of the same options to counterbalance order effects, reducing sensitivity and affecting reliability. While some SL studies report internal consistency coefficients, systematic comparisons across modalities and tasks are scarce, and no comprehensive assessment protocol for SL ability exists. Arnon (2020) calculated reliability indices for multiple SL tasks in adults and children using classic paradigms, finding moderate reliability for adults but unacceptably low reliability for children. Siegelman (2017) modified visual SL tasks, substantially improving reliability but at the cost of increased testing duration and multiple

item formats, making them impractical for infants and clinical populations.

Beyond these methodological issues, researchers have noted that overly idealized assumptions also compromise reliability. SL tasks presuppose a "tabula rasa" hypothesis (Elazar et al., 2022), assuming participants have no prior exposure to the artificial language and that all learning arises from the experimental exposure. However, in auditory SL, syllable combinations often have traces in participants' native language, and heterogeneous language experience across participants introduces variability that reduces internal consistency. While syllable-based SL studies predominate, other materials such as tones (Saffran et al., 1999), sounds (Siegelman et al., 2018a), and visual shapes (Siegelman et al., 2018b) have been used (see meta-analysis by Frost et al., 2020). Visual graphics are less susceptible to prior experience and better satisfy the "tabula rasa" assumption, potentially yielding higher reliability (Siegelman et al., 2018a).

These measurement challenges have constrained research on the relationship between SL and language ability. The present study modified traditional tasks by manipulating trial difficulty, test format, and material modality. Specifically, unlike previous studies using uniform-length target structures, we synthesized learning materials with mixed-length targets, which create varied transitional probabilities and memory representations, enriching difficulty gradients and increasing score variance while preventing rhythmic expectancy effects (Hoch et al., 2013). Second, we examined familiarity rating tasks (Batterink et al., 2015), where participants rate the familiarity of target, part-word, and non-word structures, avoiding repeated option presentation that reduces sensitivity. Third, we implemented both visual and auditory modalities for comparison. We anticipated significantly improved reliability, particularly for visual familiarity rating tasks. Given constraints on sample size and task complexity, we did not include a uniform-length control condition, instead using mixed-length structures exclusively. Rather than parametric tests like t-tests, we adopted a meta-analytic approach, comparing reliability across eight conditions with previous studies () to evaluate our assessment protocol.

2. Method

2.1 Participants One hundred fifty-nine participants (47 male) aged 18–27 years participated, all native Mandarin speakers. Forty-one participants completed Auditory Material A, 39 completed Auditory Material B, 40 completed Visual Material A, and 39 completed Visual Material B. Participants provided informed consent and received modest compensation. The study was approved by the university ethics committee (NNU2022060023 and NNU202302010).

2.2 Experimental Design We employed a learning-test paradigm with a 2 (modality: visual graphics vs. auditory speech) × 2 (test task: familiarity rating vs. forced-choice) × 2 (counterbalanced material: Set A vs. Set B) mixed design. Modality and material were between-subjects variables; notably, target structures in Set A served as part-words in Set B and vice versa, ensuring effects

were not due to specific material combinations. Test task was a within-subjects variable. Participants were randomly assigned to the four between-subjects conditions, with half completing the familiarity rating task first and half completing the forced-choice task first to counterbalance order effects. Dependent variables were reliability indices: Cronbach's alpha and split-half reliability.

2.3 Materials

2.3.1 Auditory Speech Materials Learning materials were adapted from Yu et al. (2021b). We selected 10 Mandarin syllables (CV and CVV structures, the most frequent forms in Mandarin) from a syllable inventory. All syllables used the first tone to avoid tonal statistical information and had no corresponding Chinese characters to minimize associations. A female native Mandarin speaker recorded the syllables in a professional studio at 44,100 Hz sampling rate. To eliminate recording artifacts, target syllables were embedded between filler syllables (e.g., nve1-ruo1-gei1, with ruo1 as the target). Syllables were then isolated using Praat software and normalized for duration (300 ms), mean pitch (266 Hz), and intensity (70 dB). The 10 syllables were randomly combined into two sets of nonsense words: two disyllabic and two trisyllabic target words. Learning materials A and B were synthesized using Praat scripts, with constraints preventing immediate repetition of the same word and ensuring equal transition probabilities (1/3) to other words. Each target word appeared 120 times, totaling 480 words over 6 minutes.

In the forced-choice task, each trial paired one target word with one part-word of equal length to control for word-length effects. Each target word was paired with two part-words, with target words appearing first in half the trials and part-words first in the other half to counterbalance order effects. There were 8 disyllabic and 8 trisyllabic pairs (16 trials total). In the familiarity rating task, participants rated target words, part-words, and non-words on a 7-point familiarity scale. Trial order was randomized for both tasks. Example stimuli are shown in .

2.3.2 Visual Graphic Materials Visual shapes were adapted from Siegelman et al. (2018b). Ten meaningless shapes comprised the learning materials. Each shape was presented for 800 ms followed by a 100 ms blank interval (SOA = 900 ms) to ensure robust learning. The same construction principles applied as in the auditory modality, with each target shape combination appearing 28 times. Forced-choice and familiarity rating tasks paralleled the auditory modality, with target, part-word, and non-word structures presented as unified visual configurations. Example stimuli are shown in [Figure 2: see original paper].

2.3.3 Experimental Procedure The experiment was programmed in E-Prime. Auditory tasks were completed with headphones at a fixed computer volume of 30%. Both modalities included practice and formal phases. After

instructions, participants were exposed to 5 seconds of learning material, followed by forced-choice and familiarity rating tasks. Practice materials did not appear in the formal phase. The auditory task required approximately 15 minutes, while the visual task required about 10 minutes. The experimental flow is illustrated in [Figure 3: see original paper]. Materials, data, and code are available at: https://github.com/wenboyu0803/reliability-of-SL.

3. Results

Data were analyzed using R (4.3.1). Cronbach's alpha and split-half reliability were calculated using the reliability function from the psych package. Results are presented in .

shows Cronbach's alpha and split-half reliability coefficients for each condition. The relationship between reliability indices from the current study and previous research is depicted in [Figure 4: see original paper], revealing that mixed-length target structures produced reliability comparable to or better than prior studies. Learning effects were significant across all tasks: In the auditory forced-choice task, accuracy exceeded chance levels ($t(79) = 5.18$, $p < 0.001$, 95% CI: [0.07, 0.16], $d = 0.58$). In the auditory familiarity rating task, target words received higher ratings than part-words ($t(79) = 4.57$, $p < 0.001$, 95% CI: [0.33, 0.84], $d = 0.51$) and non-words ($t(79) = 10.64$, $p < 0.001$, 95% CI: [1.25, 1.83], $d = 1.54$). In the visual forced-choice task, accuracy also exceeded chance ($t(73) = 5.93$, $p < 0.001$, 95% CI: [0.10, 0.20], $d = 0.68$). In the visual familiarity rating task, target structures were rated higher than part-structures ($t(76) = 6.58$, $p < 0.001$, 95% CI: [0.84, 1.57], $d = 1.20$) and non-structures ($t(76) = 12.77$, $p < 0.001$, 95% CI: [2.29, 3.14], $d = 2.72$). These results demonstrate robust learning effects across tasks. Finally, significant correlations emerged between forced-choice and familiarity rating tasks within each modality: $r = 0.46$ ($p < 0.001$, 95% CI: [0.25, 0.61]) for auditory, and $r = 0.45$ ($p < 0.001$, 95% CI: [0.23, 0.61]) for visual.

4. Discussion

When investigating the relationship between SL ability and language development, measurement quality is paramount. Previous studies' suboptimal reliability has generated considerable debate. By employing mixed-length target structures and comparing forced-choice versus familiarity rating tasks across modalities, the present study identified a more effective assessment protocol. Visual modality tasks achieved high internal consistency, generally meeting psychometric standards, with superior split-half reliability intervals. Forced-choice tasks outperformed familiarity rating tasks, particularly in the visual modality.

4.1 Impact of Mixed-Length Learning Materials on Reliability In the auditory modality, we combined disyllabic and trisyllabic target words; in the visual modality, we combined two-shape and three-shape sequences. According to memory chunking mechanisms in SL (Isbilen et al., 2020; Perruchet, 2019),

participants' test performance reflects memory representations formed during learning. Trisyllabic target words and part-words share syllable combinations with only one boundary cue, creating similar memory strengths and greater selection difficulty. In contrast, disyllabic targets and part-words share no syllables but have one boundary feature, producing larger memory strength differences and easier discrimination. Mixed-length structures thus create a finer difficulty gradient in forced-choice tasks. Familiarity rating tasks further enhance difficulty variation by including non-structures that never appeared during learning, representing the easiest discrimination. As predicted, mixed-length structures yielded reliability equal to or exceeding previous uniform-length studies.

4.2 Impact of Material Modality on Reliability Cronbach's alpha exceeded 0.70 for all visual modality tasks, reaching 0.86 and 0.74 for forced-choice tasks—(nearly) meeting psychometric standards (Nunnally & Bernstein, 1994). Conversely, three of four auditory conditions fell below 0.70. Visual modality split-half reliability intervals were higher and narrower than auditory intervals, confirming that visual materials enhance SL task reliability. These findings align with Siegelman et al. (2018a): visual SL is less susceptible to individual language experience, yielding more consistent judgments across participants. Siegelman's revised visual tasks (2017, 2018a) presented each target structure only 24 times with 42 test trials combining familiarity ratings and pattern completion, achieving alphas of 0.84 and 0.78—comparable to our results. However, our study employed the more common two-alternative forced-choice format with only 16 trials, offering greater efficiency for use with children and clinical populations while maintaining psychometric quality.

Although some research suggests auditory SL predicts reading skills (Gabay et al., 2015; Qi et al., 2019), most studies use visual materials (Tong et al., 2019; Lee et al., 2022). Theoretically, visual shapes may have stronger associations with logographic scripts like Chinese characters. Our findings demonstrate that visual materials provide more stable SL assessment, potentially advancing research on SL and reading skills in Chinese children. In Mandarin, only about 20 meaningless syllables conform to phonotactic rules with the first tone, creating material construction challenges. Considering psychometric requirements and practical feasibility, we recommend visual modality tasks for assessing SL ability. Finally, significant within-modality correlations between forced-choice and familiarity rating tasks emerged. While such correlations in the auditory modality are well-documented (Erickson et al., 2016), visual modality task correlations have rarely been reported and warrant further investigation.

4.3 Impact of Test Task Type on Reliability In addition to forced-choice tasks, we examined familiarity rating tasks. Forced-choice tasks require repeated option presentation to counterbalance order, incorporating both initial learning and secondary learning effects during testing, which reduces sensitivity. However, results showed that visual forced-choice tasks produced higher Cronbach's alphas and narrower split-half reliability intervals than familiarity rating tasks,

indicating that forced-choice tasks provide better psychometric assessment of SL ability.

4.4 Limitations and Future Directions Several limitations warrant consideration. Regarding task format, construct validity analyses suggest both forced-choice and familiarity rating tasks are reflective tasks that assess not only statistical information extraction but also metacognitive abilities (Isbilen & Christiansen, 2022; Ordin & Polyanskaya, 2021), potentially confounding pure SL measurement. Future research should examine alternative task formats. Regarding material parameters, learning effects depend on exposure duration and frequency in addition to probabilistic information (Bogaerts et al., 2016). Future studies targeting infants should systematically investigate these factors.

To meet basic psychometric standards, this study modified SL measurement protocols and found that combining mixed-length target structures with forced-choice tasks in the visual modality yields stable reliability indices.

References

Xu, G., Fan, R., & Jin, H. (2020). The cognitive and neural mechanisms of statistical learning and its relationship with language. *Advances in Psychological Science*, 28(9), 1525-1538.

Yu, W., Wang, L., Cheng, X., Wang, T., Zhang, J., & Liang, D. (2021a). The influence of language experience on probabilistic word segmentation. *Advances in Psychological Science*, 29(5), 787-795.

Yu, W., Wang, L., Qu, X., Wang, T., Zhang, J., & Liang, D. (2021b). The effects of transitional probability and word length expectancy on speech statistical learning. *Acta Psychologica Sinica*, 53(6), 565-574.

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.

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

Bogaerts, L., Siegelman, N., & Frost, R. (2016). Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychonomic Bulletin & Review*, 23, 1250-1256.

Bogaerts, L., Frost, R., & Christiansen, M. H. (2020). Integrating statistical learning into cognitive science. *Journal of Memory and Language*, 115, 104167.

Elazar, A., Alhama, R. G., Bogaerts, L., Siegelman, N., Baus, C., & Frost, R. (2022). When the "Tabula" is anything but "Rasa:" What determines performance in the auditory statistical learning task?. *Cognitive Science*, 46(2), e13102.

Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra*, 2(1), 14.

Estes, K. G., 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.

Estes, K. G., Gluck, C. W., & Bastos, C. (2015). Flexibility in statistical word segmentation: Finding words in foreign speech. *Language Learning and Development*, 11(3), 252-269.

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. L. (2015). Impaired statistical learning in developmental dyslexia. *Journal of Speech, Language, and Hearing Research*, 58(3), 934-945.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50, 1166-1186.

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., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*, 44(7).

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

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

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, 87(1), 184-193.

Kidd, E., Arciuli, J., Christiansen, M. H., Isbilen, E. S., Revius, K., & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology*, 200, 104964.

Lammertink, I., Boersma, P., Rispens, J., & Wijnen, F. (2020). Visual statistical learning in children with and without DLD and its relation to literacy in children with DLD. *Reading and Writing*, 33(6), 1557-1589.

Lee, S. M. K., Cui, Y., & Tong, S. X. (2022). Toward a model of statistical learning and reading: Evidence from a meta-analysis. *Review of Educational*

Research, 92(4), 651-691.

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, 8(3), 447−461.

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Ordin, M., & Polyanskaya, L. (2021). The role of metacognition in recognition of the content of statistical learning. *Psychonomic Bulletin & Review*, 28, 333-340.

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

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).

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, 69(1), 181–203.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1), 27-52.

Siegelman, N. (2020). Statistical learning abilities and their relation to language. *Language and Linguistics Compass*, 14, e12365.

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., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018a). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, 177, 198-213.

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018b). Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, 42(S3), 692-727.

Tong, X., Leung, W. W. S., & Tong, X. (2019). Visual statistical learning and orthographic awareness in Chinese children with and without developmental dyslexia. *Research in Developmental Disabilities*, 92, 103443.

von Koss Torkildsen, J., Arciuli, J., & Wie, O. B. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences*, 69, 60-68.

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