

MIT Press • Open Encyclopedia of Cognitive Science

Statistical Learning

Jenny Saffran¹

¹Department of Psychology and Waisman Center, University of Wisconsin - Madison

MIT Press

Published on: Jul 24, 2024

DOI: <https://doi.org/10.21428/e2759450.7a74eed8>

License: [Creative Commons Attribution-NonCommercial 4.0 International License \(CC-BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/).

Statistical learning refers to the ability to track patterns in the environment. These patterns occur in a wide array of domains (e.g., speech, scenes, melodies). Humans track statistics over stimuli ranging from simple units (e.g., phonemes, tones, geometric shapes) to abstract categories (e.g., patterns of nouns and verbs) and also connect stimuli across multiple modalities. A wide range of statistical regularities are detected by learners, including but not limited to frequency distributions of individual elements, frequencies of co-occurrence, and probabilities of co-occurrence. Statistical learning may be bottom-up (driven by the input itself) or integrated with prior knowledge about the structure of the input.

History

In the 1980s and 1990s, researchers in several language-related disciplines converged on the potential of statistical regularities as potentially informative cues for language learners [see [Language Acquisition](#)]. This convergence included quantitative analyses of language corpora, natural language processing approaches in AI, computational models of language learning (notably, Jeff Elman's work on recurrent networks), and experiments in which adult learners were exposed to artificial languages designed to simulate key aspects of natural languages. At the same time, advances in infant research methods were leading researchers like Dick Aslin, Peter Jusczyk, Pat Kuhl, Jim Morgan, and Janet Werker to go beyond earlier questions about what infants know at a given age, turning instead to questions about how infants learn.

Research focused on human statistical learning emerged at the nexus of two interrelated theoretical issues: nature versus nurture and domain specificity versus domain generality [see [Cognitive Development](#)]. The first issue concerns what types of knowledge and abilities are needed to explain developmental phenomena above and beyond environmental input. The second issue concerns the degree to which such knowledge and abilities are tailored to solve a specific learning problem (e.g., language development) versus available for learning more generally. Beginning in the 1990s, researchers found that infants were sensitive to statistical regularities in stimulus streams including syllables, musical notes, and visual scenes (for review, see [Saffran & Kirkham, 2018](#)). Over the ensuing decades, the literature has broadened to include participants across the human lifespan, including both neurotypical and neurodiverse individuals. A wide range of tasks and methods have been applied in this area, including studies with nonhuman animals, neuroimaging methods, and computational modeling.

Core concepts

Questions about human statistical learning abilities have been shaped by considerations of the kinds of challenges that infants face in real-world learning. For example, research focused on the role of frequency of occurrence was motivated by the challenge of acquiring native language phoneme categories ([Maye, Gerken, & Werker, 2002](#)). Similarly, research focused on the role of co-occurrence statistics was motivated by observations about the problem of word segmentation: How do infants figure out where one word ends and the next one begins in the absence of acoustic cues such as pauses (e.g., [Saffran, Aslin, &](#)

[Newport, 1996](#))? We see similar motivations for research focused on learning of nonadjacent statistical patterns, which are important for grammatical structure in language (e.g., [Gómez, 2002](#)), and for studies examining how infants track associations between labels and objects, which are important for understanding word learning [see [Word Learning](#)] (e.g., [Yu & Smith, 2007](#)).

For these reasons, many of the tasks used to interrogate statistical learning were designed to simulate infants' real-world learning. In the language domain, studies often employ miniature artificial languages that are stripped down versions of real languages designed to manipulate the statistics of interest. Similarly, experiments in the visual domain employ miniature visual worlds, such as patterns of shapes that co-occur in space (e.g., [Fiser & Aslin, 2002](#)) or time (e.g., [Kirkham, Slemmer, & Johnson, 2002](#)).

Questions, controversies, and new developments

One important open question pertains to the degree to which statistical learning processes are tailored to specific domains. Some evidence suggests dissimilarities in learning across domains, even in infancy (e.g., [Emberson et al., 2019](#)). Other studies suggest similarities in the kinds of statistical patterns infants prioritize across domains (e.g., [Santolin & Saffran, 2019](#)). Research focused on the neural substrates of statistical learning suggests that both domain-general brain areas (like the hippocampus) and modality-specific brain areas (such as the auditory or visual cortices) are implicated.

Another open question pertains to age and whether there are developmental differences in the use of statistical regularities. Some studies suggest remarkable similarities between infants and adults ([Choi et al., 2020](#)), whereas others suggest developmental changes (for review, see [Forest et al., 2023](#)). It seems likely that the particular structures to be learned interact with age and experience, such that there are developmental differences for some statistical regularities but not others (e.g., [Newport, 2020](#)).

A third open question is whether statistical learning accounts can explain the acquisition of complex aspects of human language. Some studies suggest that insufficient information exists in the input to support learning without innate knowledge (e.g., [Han, Musolino, & Lidz, 2016](#)). However, contemporary large language models generate text that adheres to syntactic structure without access to innate knowledge. These AI successes support the view that natural language input contains exceptionally rich statistical cues. The degree to which these statistical regularities are available to human learners remains an open question (for a recent infant-scaled large language model of a statistical language learning problem, see [Vong et al., 2024](#)) [see [Large Language Models](#)].

Broader connections

In addition to these links between developmental psychology, cognitive neuroscience, computational models, and linguistics, statistical learning research connects with several additional fields. One of these is communication disorders. Challenges in statistical learning have been observed in individuals with a range of developmental disabilities, particularly those that involve language, such as developmental language disorder (e.g., [Lammertink et al., 2017](#)) and dyslexia (e.g., [Lee, Cui, & Tong, 2022](#)). These issues

dovetail with broader questions about individual differences in statistical learning and how they may relate to language skills in the population more generally (e.g., [Bogaerts, Siegelman, & Frost, 2021](#)).

Another field with intriguing connections to statistical learning is comparative psychology. Statistical learning research with nonhuman animals has demonstrated both similarities and dissimilarities with human learners (for review, see [Santolin & Saffran, 2018](#); [Wilson et al., 2020](#)). Better understanding of these confluences and divergences may help us to gain traction on key questions concerning human evolution.

Further reading

- Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1–2), e1373.
- 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. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 145(12), 1128.
- Saffran, J. R. (2020). Statistical language learning in infancy. *Child Development Perspectives*, 14(1), 49–54.

References

- Bogaerts, L., Siegelman, N., & Frost, R. (2021). Statistical learning and language impairments: Toward more precise theoretical accounts. *Perspectives on Psychological Science*, 16(2), 319–337.

↩

- Choi, D., Batterink, L. J., Black, A. K., Paller, K. A., & Werker, J. F. (2020). Preverbal infants discover statistical word patterns at similar rates as adults: Evidence from neural entrainment. *Psychological Science*, 31(9), 1161–1173.

↩

- Emberson, L. L., Misyak, J. B., Schwade, J. A., Christiansen, M. H., & Goldstein, M. H. (2019). Comparing statistical learning across perceptual modalities in infancy: An investigation of underlying learning mechanism(s). *Developmental Science*, 22(6), e12847.

↩

- Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24), 15822–15826.

↩

- Forest, T. A., Schlichting, M. L., Duncan, K. D., & Finn, A. S. (2023). Changes in statistical learning across development. *Nature Reviews Psychology*, 2(4), 205–219.

↩

- Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5), 431–436.



- Han, C. H., Musolino, J., & Lidz, J. (2016). Endogenous sources of variation in language acquisition. *Proceedings of the National Academy of Sciences of the United States of America*, 113(4), 942–947.



- Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83(2), B35–B42.



- Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical learning in specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research: JSLHR*, 60(12), 3474–3486.



- 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.



- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101–B111.



- Newport, E. L. (2020). Children and adults as language learners: Rules, variation, and maturational change. *Topics in Cognitive Science*, 12(1), 153–169.



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



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



- Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, 22(1), 52–63.



- Santolin, C., & Saffran, J. R. (2019). Non-linguistic grammar learning by 12-month-old infants: Evidence for constraints on learning. *Journal of Cognition and Development*, 20(3), 433–441.



- Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682), 504–511.

[↩](#)

- Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., van der Kant, A., Smith, K., & Rey, A. (2020). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science*, 12(3), 843–858.

[↩](#)

- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414–420.

[↩](#)