COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13198

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13198

Statistical Learning of Language: A Meta-Analysis Into 25 Years of Research

Erin S. Isbilen,[a,b] Morten H. Christiansen[a,b,c]

[a]*Department of Psychology, Cornell University*
[b]*Haskins Laboratories*
[c]*Interacting Minds Centre and School of Communication and Culture, Aarhus University*

Received 20 July 2021; received in revised form 16 August 2022; accepted 22 August 2022

Abstract

Statistical learning is a key concept in our understanding of language acquisition. Ample work has highlighted its role in numerous linguistic functions—yet statistical learning is not a unitary construct, and its consistency across different language properties remains unclear. In a meta-analysis of auditory-linguistic statistical learning research spanning the last 25 years, we evaluated how learning varies across different language properties in infants, children, and adults and surveyed the methodological trends in the literature. We found robust learning across stimuli (syllables, words, etc.) in infants, and across stimuli and structures (adjacent dependencies, non-adjacent dependencies, etc.) in adults, with larger effect sizes when multiple cues were present. However, the analysis also showed significant publication bias and revealed a tendency toward using a narrow range of simplified language properties, including in the strength of the transitional probabilities used during training. Bayes factor analyses revealed prevalent data insensitivity of moderators commonly hypothesized to impact learning, such as the amount of exposure and transitional probability strength, which contradict core theoretical assumptions in the field. Methodological factors, such as the tasks used at test, also significantly impacted effect sizes in adults and children, suggesting that choice of task may critically constrain current theories of how statistical learning operates. Collectively, our results suggest that auditory-linguistic statistical learning has the kind of robustness needed to play a foundational role in language acquisition, but that more research is warranted to reveal its full potential.

Keywords: Statistical learning; Auditory-linguistic statistical learning; Cross-situational learning; Language; Language acquisition; Meta-analysis

Correspondence should be sent to Erin S. Isbilen, Haskins Laboratories, 300 George Street, #900, New Haven, CT 06511, USA. E-mail: erin.isbilen@yale.edu

1. Introduction

Statistical learning, the ability to track recurring distributional patterns in the environment, is an integral component of cognition in humans and other species. It has been implicated as a key theoretical construct in numerous cognitive abilities, spanning the auditory (e.g., Saffran, Johnson, Aslin, & Newport, 1999), visual (e.g., Fiser & Aslin, 2001; 2005), and tactile modalities (e.g., Conway & Christiansen, 2005), in both human and non-human animals (Hauser, Newport, & Aslin, 2001; Newport, Hauser, Spaepen, & Aslin, 2004). This versatile ability is cited as a cornerstone of many sophisticated behaviors, including the development of social inference in young children (e.g., Kushnir, Xu, & Wellman, 2010), the classification of novel objects into distinct categories (e.g., Folstein, Gauthier, & Palmeri, 2010), and even the learning of action sequences (Baldwin, Andersson, Saffran, & Meyer, 2008) and emotional expressions (Mermier, Quadrelli, Turati, & Bulf, 2022). Of particular interest, the last 25 years have seen a surge of studies exploring the role of statistical learning in one of our most unique and complex abilities as a species: human language.

Perhaps the best-known example of statistical learning in the linguistic domain comes from the formative study by Saffran, Aslin, and Newport (1996). This study demonstrated that 8-month-old infants could track the transitional probabilities between one syllable and the next after mere minutes of exposure to an artificial language (e.g., for the sequence *AB*, the probability of *B* given *A*). Sensitivity to such statistical information is found to be useful for segmenting continuous speech, even when controlling for the frequency of syllable co-occurrences (i.e., how often they appear in the language; Aslin, Saffran, & Newport, 1998). This kind of statistical learning may thus help individuals break down linguistic input into words and phrases using simple, domain-general cognitive mechanisms (Saffran, 2003). This groundbreaking finding sparked an explosion of studies devoted to testing the potential of statistical learning in explaining behavior across different domains, age groups, and structures (for a review, see Frost, Armstrong, & Christiansen, 2019). Since then, it has been shown that infants, children, and adults can all leverage statistical regularities to discriminate words that they have been exposed to in fluent speech from foil items that were not present in the input, and that this ability extends to non-adjacent regularities (items that do not occur directly next to one another in a sequence; R. L. Gómez, 2002; Newport & Aslin, 2004). This ability can even predict individual differences in language proficiency across development (e.g., R. L. A. Frost et al., 2020; Gabay, Thiessen, & Holt, 2015; Isbilen, McCauley, & Christiansen, 2022; Mirman, Magnuson, Estes, & Dixon, 2008; see Mirman, Graf Estes, & Magnuson, 2010, for computational modeling of this effect), suggesting that these simple, laboratory-based experiments may capture fundamental aspects of language learning in the real world.

Statistical learning has been implicated in a stunning breadth of linguistic abilities, from the acquisition of phonological regularities to the learning of grammatical patterns. However, the question of how multipurpose versus stimulus-specific such learning is remains a central topic of theoretical debate (for a review, see Frost, Armstrong, Siegelman, & Christiansen, 2015). Indeed, some researchers have questioned whether statistical learning may be an overly broad term that links disparate learning phenomena (Thiessen, 2017): given the considerable differences among statistical learning tasks, do they in fact all depend on shared cognitive processes

or different mechanisms? In this paper, we set out to test how the efficacy of statistical learning might vary across different language properties by performing a meta-analysis on a large sample of the published research on auditory-linguistic statistical learning. We focus on this particular form of learning as it is one of the most widely studied subareas and is arguably the most relevant for understanding the contribution of statistical learning to language acquisition in hearing populations.[1] As statistical learning is hailed as a cornerstone of language, it is, therefore, imperative to understand its strengths and limitations given the current state of the field and whether key theoretical assumptions uphold across the literature.

To motivate our analysis, we start by discussing the nature of statistical learning. We review the auditory-linguistic statistical learning literature to pinpoint current theoretical questions and assumptions about how this phenomenon operates. We then explain how meta-analyses enable researchers to empirically test these questions and assumptions by pooling data from numerous samples to deduce which effects are most robust in the literature. In the present case, we also use the meta-analysis to identify methodological trends in statistical learning research. Identifying these trends can elucidate what areas of the literature are well explored, what areas require additional inquiry, and how current methods might constrain our understanding of statistical learning behavior. Finally, we conclude the introduction by presenting our moderators of interest, all of which are commonly assumed to significantly influence statistical learning, and language learning at large: the properties of the input languages, how participants were trained, how they were tested, and how these features influence performance across development (in infants, children, and adults). Given the theorized centrality of statistical learning to numerous aspects of language acquisition, the analysis of these features enables us to gauge the degree to which it is truly general purpose, and how it might vary across different linguistic inputs.

1.1. The multifaceted nature of statistical learning

Statistical learning has been documented across a broad range of domains and modalities, lending weight to the idea that such learning may serve as a mainspring of cognition in humans and other species. Yet despite its generality, this line of research also poses several theoretical challenges: though multipurpose, statistical learning is not uniform. For instance, auditory and visual statistical learning seem to follow different developmental trajectories, with the learning of auditory regularities outpacing the learning of visual regularities (Raviv & Arnon, 2018). Moreover, the question of whether statistical learning operates uniformly across the different surface properties of stimuli even *within* a single domain and modality is vigorously debated (see R. Frost et al., 2019, for a review). For example, individuals' capacity to learn adjacent auditory-linguistic regularities does not reliably predict their ability to learn non-adjacent auditory-linguistic regularities, even though both structures share the same modality and domain (Siegelman & Frost, 2015). Such findings have led researchers to question what makes a "good" statistical learner when the learning of one kind of structure does not necessarily predict sensitivity to other similar structures (Bogaerts, Siegelman, Christiansen, & Frost, 2022). As the staggering diversity of the world's languages suggests, natural languages are richly varied, sporting a panoply of phonological, grammatical, and

morphological structures (Evans & Levinson, 2009). To date, research on statistical learning has greatly enriched our understanding of how individuals overcome what is commonly cited as a key initial hurdle for novice learners: finding the words in continuous speech. However, the segmentation of individual words is far from the only aspect of language acquisition—learning in the real world is considerably more complex. In order for statistical learning to cash in on its promise as a central component of language acquisition, its utility should generalize at least to some degree to other structures and contexts, beyond the segmentation of individual words.

Some headway has been made into these issues, both within the statistical learning literature and under the banners of artificial grammar learning or implicit learning (for reviews on the historical separations in the literature, see Christiansen, 2019; Perruchet & Pacton, 2006). Scores of studies have now tested the acquisition of artificial grammars involving multiple types of regularities (e.g., Getz, Ding, Newport, & Poeppel, 2018; R. L. Gómez & Gerken, 1999; Saffran, 2001), illustrating that statistical learning can capture the acquisition of linguistic structures beyond individual words. Others have successfully extended the statistical learning framework to model facets of multimodal integration and semantic acquisition, demonstrating that individuals can map newly acquired auditory words onto visual referents based on their statistical co-occurrence with one another during training (e.g., Benitez, Yurovsky, & Smith, 2016; Graf Estes, 2012). The boundaries of this ability have been further probed by experiments investigating how learners aggregate cross-situational statistics in the auditory and visual modalities. Cross-situational learning paradigms simulate language learning in the real world by presenting participants with words or phrases that can potentially map onto multiple competing objects or scenes in the environment (e.g., an infant hears the word "ball," which can map onto numerous toys in their field of vision). Over time, learners can capitalize on these co-occurrences to acquire words, phrases, and their meanings (e.g., Monaghan, Schoetensack, & Rebuschat, 2019; Smith & Yu, 2008; Yurovsky, Fricker, Yu, & Smith, 2014; Yurovsky, Yu, & Smith, 2013), just as they do in natural settings (e.g., the infant consistently hears the word "ball" in relation to small round objects and eventually surmises that such objects are the referents for this word).

However, while evidence of statistical learning has been cataloged across several different types of language properties, considerable differences within auditory-linguistic statistical learning have also been observed. As mentioned above, individuals' ability to acquire adjacent linguistic dependencies does not significantly correlate with their ability to acquire non-adjacent linguistic dependencies (e.g., Misyak & Christiansen, 2012; Siegelman & Frost, 2015). Prior language experience might also affect learning differentially (e.g., Trecca et al., 2019), with patterns that deviate from one's native language being considerably harder to learn. Furthermore, some studies find that adjacent word dependencies appear to be easier to acquire than non-adjacent word dependencies, with the learning of non-local regularities only occurring under particular conditions (e.g., only when the intervening units display sufficient variability; R. L. Gómez, 2002; R. L. Gómez & Maye, 2005). However, others suggest that when probabilities are controlled, both types of regularities can be learned simultaneously and at the same rate (Vuong, Meyer, & Christiansen, 2016). This was revealed by online serial reaction time (RT) data, whereas offline grammaticality judgment data suggested that

adjacent dependencies were better learned. These disparities give rise to the question of whether statistical learning might work equally well across all linguistic features, or if it is better equipped to support certain kinds of learning over others. It also suggests that the results of statistical learning studies may be critically influenced by the tasks used to measure it, in some cases leading to weak correlations between tasks that are designed to tap into the same kinds of auditory regularity (e.g., Erickson, Kaschak, Thiessen, & Berry, 2016; Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018).

Researchers have also explored whether statistical learning might be facilitated by additional cues in the environment, which appear to be an important feature of language acquisition in the real world. Natural languages are rife with cues beyond the statistical probabilities between linguistic elements, and the interactive nature of language introduces a host of factors that may facilitate acquisition. The presence of multiple cues can enrich the learning environment by drawing learners' attention to salient features of the input, which may help them gain a foothold in the language system (e.g., Bates & MacWhinney, 1989; Gleitman & Wanner, 1982; Morgan, 1996). For instance, social cues such as eye gaze can help direct attention to the correct referent for novel labels in cross-situational learning tasks, enabling individuals to identify the target object among a multitude of competitors (MacDonald, Yurovsky, & Frank, 2017). This closely resembles what has been observed in infant–caretaker interactions in natural settings, with infants attending to caretaker gaze and other referential actions during labeling events (e.g., Baldwin, 1993; Yu & Smith, 2012). Prosodic cues, such as the exaggerated pitch contours frequently present in child-directed speech, can also help learners determine the boundaries of words in statistical segmentation tasks (e.g., Endress & Mehler, 2009; Jusczyk, Houston, & Newsome, 1999; Morgan, Meier, & Newport, 1987; Mueller, Bahlmann, & Friederici, 2010; Thiessen & Saffran, 2003). Additionally, others have shown that more complex, probabilistic non-adjacent patterns can only be learned when additional visual or auditory cues are provided (Van den Bos, Christiansen, & Misyak, 2012). These studies provide important insights into how statistical learning interfaces with aspects of natural language acquisition, suggesting that additional cues can be integrated with the statistics of the language to improve learning outcomes.

Collectively, the last quarter of a century of statistical learning research suggests that acquired sensitivity to distributional patterns may play a role in many aspects of language learning, from the acquisition of individual words from fluent speech to the generalization of grammatical regularities. It can support the formation of word-referent mappings and can work in tandem with additional environmental cues to bolster learning. While numerous studies showcase the ubiquity of statistical learning across many aspects of language acquisition, what remains unclear is its relative *strength* across different linguistic features. As statistical learning is cited as a central feature of language acquisition, it is crucial to assess what theoretical assumptions hold across the literature and isolate which ones fail to reach significance, in order to hone future theories and guide the next decades of research.

1.2. A meta-analytic approach to statistical learning

In this paper, we take stock of a large sample of studies published on auditory-linguistic statistical learning since the seminal Saffran et al. (1996) study. We evaluate the contribution

of this form of statistical learning to different aspects of language acquisition by conducting a meta-analysis of a large subset of papers published over the last 25 years. Meta-analyses offer a powerful tool for analyzing a phenomenon across labs, methodologies, and populations, allowing researchers to isolate which effects are most reliable in the literature. They are typically conducted by analyzing the effect sizes of published studies, enabling scientists to explore the impact of various moderators on the strength of learning. Here, we additionally use our meta-analysis to reveal the most common methodological practices in the literature. While a few prior meta-analyses on statistical learning have been conducted (infant statistical learning: Black & Bergman, 2017; artificial grammar learning across species: Trotter, Monaghan, Beckers, & Christiansen, 2020; statistical learning in specific language impairment: Lammertink, Boersma, Wijnen, & Rispens, 2017), they have typically worked with smaller sample sizes and adopted different foci than the present analysis. To our knowledge, the meta-analysis presented here is the largest analysis of auditory-linguistic statistical learning to date.

Synthesizing the results from 175 papers and 636 studies on auditory-linguistic statistical learning, we ask: How is statistical learning impacted by different language and training properties? What are the kinds of methods that have been used to test statistical learning and are some tasks more adept at capturing this behavior? Do different age groups vary in their ability to learn certain language properties and are they differentially impacted by methodological factors? And importantly, where is further work needed?

Based on the methods laid out in Trotter et al. (2020), which outlines a comprehensive guide for conducting large-scale meta-analyses related to artificial grammar learning, we examined the influence of an extensive collection of moderators on statistical learning effect sizes in infants, children, and adults. We take an in-depth look at several key factors that are hypothesized to impact learning: the language properties, the training methods, and the testing methods.

The language properties describe the nature of the input that participants were tasked with learning. These moderators include stimulus type (e.g., whether the input presents syllable-level dependencies, word-level dependencies, phoneme dependencies, etc.), structure type (e.g., adjacent dependencies, non-adjacent dependencies, or multiple-regularity languages[2] that involve a combination of adjacent, non-adjacent, or other grammatical structures), and the strength of the transitional probabilities utilized in the language. These moderators allow us to gauge how statistical learning fluctuates across inputs, to determine the degree to which it is general purpose versus stimulus specific. It also allows us to survey what language properties are well represented in the literature, and what warrants further examination.

The training methods describe how participants were trained during the exposure phase of each experiment. We evaluate whether the number of training items shapes the strength of statistical learning, and, specifically, whether increasing the number of items significantly decrements effect sizes. In addition, we gauge how the number of exposures to each training item impacts effect size. These moderators address how task load and exposure influence learning in the different age groups. Indeed, studies have shown that the amount of exposure, the number of words in a language, and the length of those items all significantly impact statistical learning, with fewer exposures, more items, and longer items hindering learning

(Frank, Goldwater, Griffiths, & Tenenbaum, 2010). These analyses on training methods dovetail with questions raised in the language acquisition literature at large, such as how the quantity and variability of input influence developmental outcomes (e.g., Hart & Risley, 2003). In addition, we analyzed how the presence of multiple cues (prosodic cues, speaker cues, visual cues, etc.) impacts statistical learning, building upon a long lineage of studies demonstrating the importance of multiple cue integration to language. We also assessed whether prior knowledge (e.g., whether the trained languages are constructed to be congruent or incongruent with the phonotactics of the learner's native language; Finn & Hudson Kam, 2008; 2015) determines the strength of statistical learning, for studies that manipulated these dimensions.

Finally, the testing methods describe the way that participants' knowledge of the trained languages is tested. In recent years, it has come to light that different tests of statistical learning vary in their reliability and sensitivity to individual differences in learning (e.g., Arnon, 2020; Siegelman et al., 2017; Siegelman, Bogaerts, Christiansen, & Frost, 2017), which suggests that the choice of task that experimenters employ may at least in part affect the strength of the observed results. In addition, the nature of the tasks used and the computations they rely upon have recently been called into question. Reflection-based measures, such as the classic two-alternative forced-choice (2AFC) task, require participants to deliberate over learned material, thereby potentially adding noise to such measures due to individual variation in people's ability to introspect about what they have learned (Christiansen, 2019). These tasks are argued to recruit cognitive processes that are not directly relevant for learning and may inevitably capture individual differences in reflective and decision-making abilities in addition to the studied phenomenon.

This shortcoming has led to the development of what Christiansen (2019) and others now refer to as processing-based measures—tasks that more directly tap into the mechanisms involved in the processing of statistical regularities. These measures include variants on RT tasks (e.g., Batterink, 2017; Batterink & Paller, 2017; Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015; D. M. Gómez, Bion, & Mehler, 2011; Karuza, Farmer, Fine, Smith, & Jaeger, 2014; Poulin-Charronnat, Perruchet, Tillmann, & Peereman, 2017; Qi, Sanchez, Georgan, Gabrieli, & Arciuli, 2019; Siegelman, Bogaerts, Kronenfeld, & Frost, 2018) as well as those that utilize serial recall to test learning (e.g., Conway, Bauernschmidt, Huang, & Pisoni, 2010; Isbilen, Frost, Monaghan, & Christiansen, 2022; Isbilen, McCauley, Kidd, & Christiansen, 2020; Kidd et al., 2020). We, therefore, evaluated whether test type (processing vs. reflection based) significantly influenced effect sizes in adults and children, where such tasks are commonly used. Taking an even finer-grained approach, we also assessed the influence of specific tasks (AFC, grammaticality/familiarity ratings, RT tasks, recall, etc.). These analyses of testing methods may inform the theoretical considerations investigated in the meta-analysis by highlighting better methods of testing and identifying where choice of task might influence the theoretical deductions made about statistical learning.

In addition, we evaluated several other methodological factors, examining whether the number of test trials affected test outcomes, to determine whether longer tests resulted in poorer performance (e.g., due to fatigue effects or motivational factors). Lastly, as interest in individual differences has been burgeoning in the study of statistical learning (for a review, see Siegelman et al., 2017), we assessed whether studies that self-report as taking an

individual differences approach (as opposed to a group-level approach) generated stronger results. We also investigated how all of these factors influence cross-situational learning, where participants are tasked with matching newly acquired words to referents across different visual contexts.

By evaluating the impact of these mediating factors—language properties, training methods, and testing methods—we can not only get insight into the current state of the art in the field but also illuminate where future experimentation might be needed so that the full theoretical impact of the role of statistical learning in language acquisition can be better appraised.

2. Method

2.1. Literature search

For this meta-analysis, we were interested in pinpointing behavioral studies on auditory-linguistic statistical learning. Studies were included in the analysis if they trained participants on novel auditory-linguistic materials (either artificial or non-native speech) in an exposure phase, where statistical regularities served as the main cue to learning. Following the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher, Liberati, & Altman, 2009), the literature search was conducted using the SCOPUS database (Scopus, 2019), on journal articles published in English from January 1996 when Saffran et al. (1996) was published through December 2020. Both the search criteria and analyses were pre-registered (https://aspredicted.org/rr49u.pdf).

To target studies on auditory-linguistic statistical learning, the database was searched for the keywords[3] "'language' AND 'statistical' OR 'distributional' AND 'learning,'" where all terms were required in an article's title, keywords, and/or abstract. This returned 3,633 total articles, which were further refined by filtering out studies that did not meet certain criteria through the filters in the Scopus search portal. To pinpoint behavioral studies, further exclusions were done on the basis of subject area (the list for each subject area was gone through by hand to ensure that no relevant articles would be lost). All subject areas other than the following were excluded: arts and humanities, psychology, social sciences, multidisciplinary, and computer science (which contained several relevant records, but mostly modeling papers which were then manually excluded). This resulted in 1,925 articles that were then individually screened (1,924 after one duplicate paper was removed). Of these, 622 were excluded because they were computer science papers, computational simulations, or corpus analyses with no human behavioral data. Furthermore, 927 other articles were excluded, as they contained the relevant search terms but were not statistical learning studies that trained participants on novel linguistic materials (e.g., computer science papers, natural language studies that situate their findings within the statistical learning framework, or papers that analyze the statistical properties of natural languages).

Studies that did not include data on auditory-linguistic statistical learning (e.g., visual, tactile, or auditory non-linguistic statistical learning) were excluded (85 articles in total). Furthermore, only studies that report human behavioral data were analyzed: papers on non-human animals (eight articles) that included no human data were excluded. Similarly, six

neuroimaging papers that contained no behavioral data were excluded (those that did report behavioral data were included in the analysis), as was one article that tested atypical populations but reported no control condition with neuro-typical participants. Despite the use of filters through the Scopus database, additional 51 review papers and meta-analyses were manually excluded from the search. Finally, six articles were not accessible, either through the database or a Google Scholar search.

While our search resulted in a large number of articles, we also acknowledge that some statistical learning studies are likely missing from our analysis. To ensure that our literature search is easily replicable by other researchers, we elected to not manually include studies that fell outside of the search parameters. Nonetheless, as the final sample includes a large number of studies that span many years, we are confident that the current paper still provides a thorough overview of the auditory-linguistic statistical learning literature.

2.2. Study selection

Data extraction was attempted on the remaining 215 articles. However, during this process, it was discovered that a subset of papers was missing data critical to the meta-analysis, such as the raw means, standard deviations, standard errors, and/or the number of participants per condition. The authors of those papers were contacted for the missing data, but did not always respond or no longer possessed the original data. This led to the exclusion of 32 additional articles. One further article was excluded because of a small sample size ($N = 4$), which made the calculation of Cohen's d (the dependent variable of interest) impossible. Finally, 10 additional articles were coded then ultimately removed from the final analysis because they introduced significant collinearities to the data (i.e., two or more of the moderators were the only instances of a particular structure and stimulus type in the entire dataset), which had been resulting in several false positive findings.

As was pre-registered, for studies that reported the results from atypical populations, only the data from typically developing controls were included. For studies that assessed test–retest reliability, only the data from the first session was tabulated. For the papers that contained control conditions where participants were exposed to a speech stream lacking statistical cues, these control studies were excluded, as there was no structure to be learned. For studies that present multiple languages or used multiple tests to assess learning of the same language (e.g., an online test and an offline test; Batterink & Paller, 2017), these were treated as separate studies, but the analyses controlled for multiple comparisons. In total, the search yielded 175 articles, which provided data from 636 studies, consisting of 14,986 unique participants (not including multiple comparisons). The search criteria and process are depicted in the PRISMA flowchart in Fig. 1 (adapted from Moher et al., 2009). The full list of papers included in the meta-analysis can be found on OSF (https://osf.io/7vkdt/).

The papers included in this meta-analysis cover a broad range of tasks, including traditional statistical learning studies that test sensitivity to transitional probabilities or other structures/regularities, to those that test participants' aptitude in mapping statistically learned words to referents (e.g., Hay, Pelucchi, Estes, & Saffran, 2011). The goal of the meta-analysis was to assess the contribution of statistical learning to language acquisition at large, and the

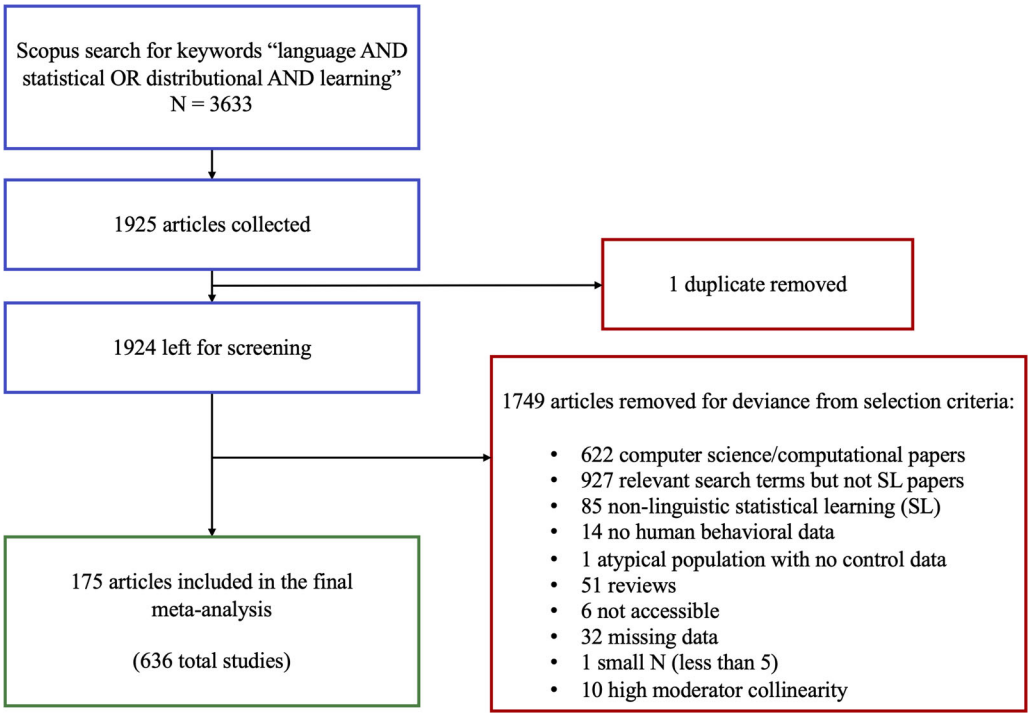

Fig. 1. PRISMA flowchart of the literature search criteria.

inclusion of diverse tasks may provide a more ecologically valid estimate of how statistical learning operates in the natural environment.

2.3. Data extraction and effect-size calculation

The data presented in this meta-analysis were manually coded from the 175 papers in the final literature search. For studies that reported their results graphically, the relevant data were extracted using the program Digitizeit (https://www.digitizeit.de). For studies with missing data, the corresponding author of each paper was contacted, and the missing values requested. In cases where studies reported the standard error (*SE*) rather than the standard deviation (*SD*), the standard deviation was derived by multiplying the standard error by the square root of the number of participants (*N*) in the study ($SD = SE \times \sqrt{N}$).

Cohen's *d* effect sizes were calculated for each study's statistical learning test data using the metafor package (Viechtbauer, 2010) in R version 3.6.1 ($d = Mean_1 - Mean_2/SD_{\text{pooled}}$, where $SD_{\text{pooled}} = \sqrt{SD1^2 + SD2^2/2}$). Based on the methods of Trotter et al. (2020), the effect sizes of alternative forced-choice and grammaticality judgment tasks were computed based on the mean's difference from chance[4] ($d = Mean - Chance/SD_{\text{pooled}}$). For studies that reported RTs, participant RTs were recorded from the final block of the study (where the strongest learning effects are typically observed; e.g., Batterink, 2017), and a difference score was calculated

by subtracting the mean of the target items from the foil items, which was then divided by the pooled *SD*. The same was done for studies that compare RTs to target syllables within different positions of words: RTs to the third/final syllable in the sequence (e.g., the stimulus that is reported to be most predictable if statistical learning has occurred) was subtracted from the averaged RTs to the first and second syllables (which learners are slower to predict; Batterink, 2017; Batterink & Paller, 2017, 2019). A similar method was used for studies that reported looking times (e.g., Saffran et al., 1996), where the looking duration of the foil items in the final test block was subtracted from the looking duration of the target items in the final test block, with a positive number indicating a preference for target items (e.g., a familiarity effect in infant studies). In cases where the authors report a significant novelty preference (e.g., Hay et al., 2011), the absolute values of these data were used,[5] as was done in the meta-analysis by Trotter et al. (2020). As both familiarity and novelty preferences are indicative of infant learning (Hunter & Ames, 1988), this ensured that the current paper did not underreport the extent of learning in this population. Composite scores were computed for studies that tested participants using multiple foil types in the same testing session (e.g., Endress & Mehler, 2009) by calculating the arithmetic mean and pooled standard deviations of the different tests.

It is worth noting that the effect sizes calculated here may differ from those reported in some of the original papers. These divergences do not reflect errors in reporting but instead arise due to differences in how the effect sizes were calculated. For instance, in Lammertink, Boersma, Wijnen, and Rispens (2020), the effect sizes reported in the original text are based on the outcomes of generalized mixed effects models which controlled for a variety of variables in their sample (such as age, experiment version, and condition). By contrast, the current paper calculated effect sizes based on the raw means and standard deviations, with our moderators of interest used as predictors in our meta-analytic models.

The moderators recorded for each study provided data about the population of the sample (i.e., age), characteristics of the training stimuli, and characteristics of testing. This included information about the properties of the trained language, including stimulus type (whether the input manipulated phoneme, syllable, or word-level dependencies), structure type (whether participants were trained on adjacent dependencies, non-adjacent dependencies, or multiple-regularity languages), and the strength of the transitional probabilities (whether they occurred with 100% regularity or otherwise). In addition, we recorded the number of exposures to each training item that participants received, the manner in which they were tested (e.g., using processing-based vs. reflection-based tasks), the number of test trials (for studies that report this information), the presence of multiple cues, and whether the experiment utilizes stimuli that leverages (or conflicts with) participants' prior knowledge. Lastly, we logged whether the study utilized a group-level approach or self-identified as employing an individual differences approach.

3. Results

All of the analyses presented in this section were pre-registered[6] through aspredicted.org (https://aspredicted.org/rr49u.pdf). In addition, all data and code are available through the Open Science Framework (https://osf.io/7vkdt/).

Table 1
Participant characteristics

Age Group	*N* Studies	*N* Papers	*N* Participants	Mean Age (Years)	Age Range
Adults	429	105	11,733	25.14	18.52–78.60
Children	60	24	2,487	7.83	3.30–14.30
Infants	147	46	3,248	12.33 (months)	5.1–31.2 (months)
Total	*636*	*175*	17,468		

3.1. Descriptive statistics

The first analyses provide a comprehensive overview of the entire dataset. They also lend insights into some of the trends that have emerged in the auditory-linguistic statistical learning literature over the last 25 years. These summaries are descriptive in nature and report on the characteristics of the participants, training materials/methods, and tests.

3.1.1. Participant characteristics

The sample contains data from a total of 17,468 participants (with multiple comparisons). This covers a wide array of ages, ranging from 5 months to 78.60 years of age. Participants fall into three broad categories: infants, children, and adults. One paper (Hsu, Tomblin, & Christiansen, 2014) identified their participants (14-year-olds) as adolescents. However, since another paper in the meta-analysis classified 13-year-old participants as children (Mayo & Eigsti, 2012), we included the 14-year-olds from Hsu et al. (2014) in the child sample rather than the adult sample. Additionally, one paper with toddlers was grouped with the infants (∼31-month-olds in Scott & Fisher, 2012), as the test procedure used was the same as those employed by infant studies (central fixation paradigm). Summary statistics of the participant characteristics are reported in Table 1.

Most of the studies in this sample were conducted with native English speakers (354 of 636 total studies). Other prominent native language groups include Spanish (33 studies), French (33 studies), Hebrew (20 studies), and Dutch (16 studies). The lowest frequency language groups included Cantonese and Japanese (one study each), Danish, Korean, and Norwegian (two studies each), Catalan, Thai, and Khalka Mongolian (three studies each), German (six studies), and Mandarin (eight studies). The remaining studies either consist of participants of mixed language backgrounds (five studies) or did not report the participants' native language (103 studies).

3.1.2. Language properties and training methods

The next set of summaries describes the characteristics of the training stimuli in the current dataset, including the types of structures presented, the kinds of dependencies that the language manipulated, the number of training items that participants were tasked with learning, and the amount of exposure to each item (for studies that report this information).

Most of the studies in the sample manipulate dependencies at the syllable level (348 studies), while 162 manipulate dependencies between words. An additional 89 present

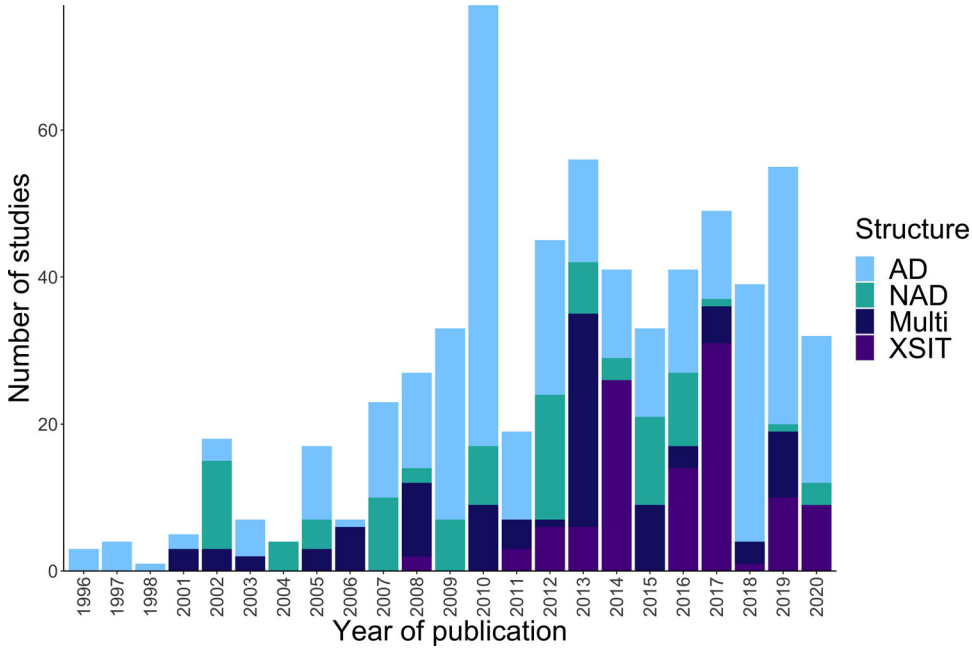

Fig. 2. The number of studies by structure type and publication year (AD = adjacent dependency, NAD = non-adjacent dependency, Multi = multiple regularity languages, XSIT = cross-situational learning).

participants with isolated words (as is the case with many cross-situational learning tasks), 19 manipulate dependencies between phonemes, 2 present isolated phonemes, while 16 present novel words embedded in sentences (5 in foreign language sentences, 11 in native-language sentences).

The bulk of the studies train participants on dependencies between adjacent syllables (308 studies in total, including studies that present isolated words that were not cross-situational learning experiments and studies that present words embedded in native language sentences). An additional five train participants on dependencies between adjacent words, while six train on adjacent phoneme dependencies. Furthermore, 101 studies train participants on dependencies between non-adjacent elements (13 phoneme level, 42 word level,[7] 46 syllable level). A further 99 present languages with multiple regularities (5 syllable level, 2 phoneme level, 87 word-level). Finally, 108 consist of cross-situational learning studies. The distribution of these studies over time is depicted in Fig. 2.

Surprisingly, the survey of the literature revealed considerable paucity in the types of transitional probabilities that participants are trained on. Of the 405 studies that report the transitional probability strength of their items, 335 utilize transitional probabilities of 1. Indeed, many of the studies in our sample largely follow the general method of the landmark Saffran et al. (1996) study: 142 studies consist of stimuli that utilize triplets of consonant-vowel syllables with adjacent transitional probabilities of 1. Raincloud plots depicting the number of studies, distributions, and boxplots of the effect sizes for each transitional probability type can be found in Fig. 3.

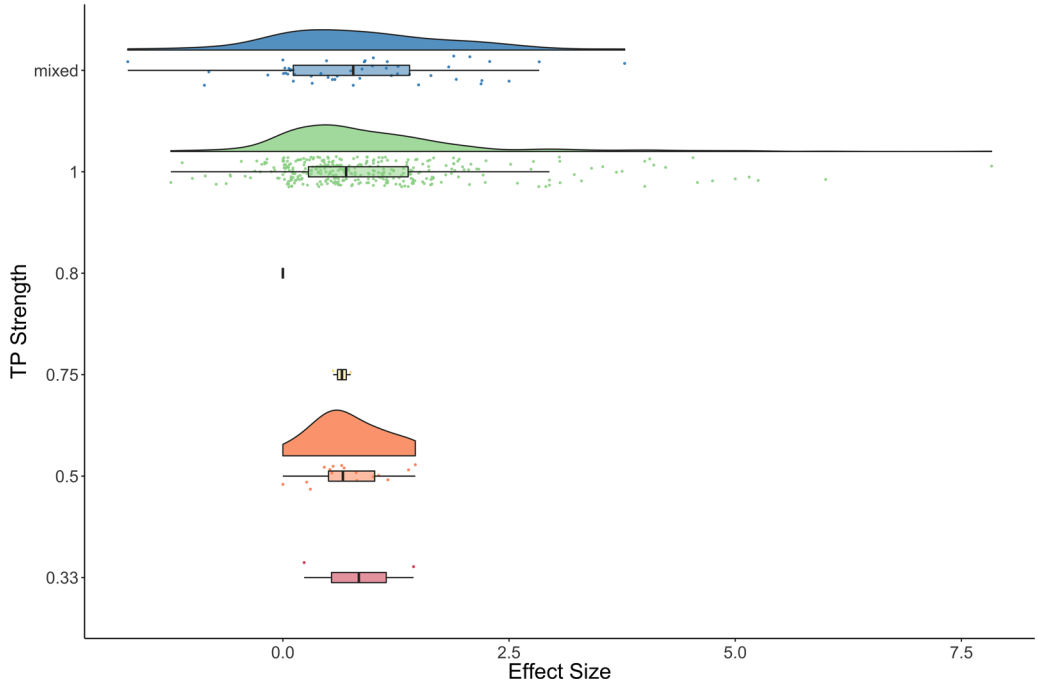

Fig. 3. The distributions, boxplots, and number of studies for each transitional probability strength. Most of the studies in this sample present items comprising transitional probabilities of 1.

Lastly, we tabulated the number of training items and the number of exposures to each item that participants receive during training. The majority of the studies train participants on a relatively small number of items (4–10 items), and most commonly administer a relatively small number of exposures (1–10 exposures to each item). An overview of the training details can be found in Table 2.

3.1.3. Testing methods

Of the entire sample, 182 studies employ processing-based measures to test learning (of which 147 are infant studies), while 454 employ reflection-based methods. Of the latter, 393 utilize forced-choice tasks (and of these, 317 are 2AFC tasks). The second most common paradigm includes looking time-based measures. Most of these were infant studies that self-reported using central fixation preference (59 studies) or headturn preference (88 studies). Central fixation paradigms present visual stimuli on a monitor (or side-by-side monitors; Pons & Toro, 2010; Scott & Fisher, 2012) while auditory stimuli play from speakers located in the same vicinity of the screen, requiring infants to only orient toward a central source of information. Headturn preference procedures place infants in an experiment room with one central light, and two side lights. A trial is initiated by the blinking of the central light, after which one of the side lights is activated. As soon as the infant orients toward the blinking side light, auditory stimuli are repeated from a speaker in that same location for as long as the

Table 2
Training methods

Moderator		Number of Studies
Number of training items		
	1	10
	2	42
	3	19
	4	165
	5–10	194
	11–24	110
	25+	96
Number of exposures		
	1–10	174
	11–20	38
	21–40	72
	41–100	104
	101+	120
	Mixed	43
	Zipfian	9

infant maintains fixation to that side light, or until the maximum trial duration elapses. This procedure requires infants to turn their heads toward different sources of information and has been cited as being potentially more cognitively demanding than central fixation procedures (Cristia, Seidl, Singh, & Houston, 2016). The sample also includes eye-tracking methods that measure looking time in children (Kavakci & Dollaghan, 2019) and looking accuracy in adults (Wonnacott, Newport, & Tanenhaus, 2008). Fig. 4 depicts the number of studies that utilize each task by publication year.

In addition, a total of 65 studies self-report taking an individual differences approach to studying statistical learning. The remaining 571 did not self-identify as such, and for the purpose of this meta-analysis these were classified as taking a group-level approach.

Finally, 59 studies in the meta-analysis comprise conditions designed to disrupt learning: for example, studies that divided participant attention (Batterink & Paller, 2019), increased cognitive load (Palmer, Hutson, White, & Mattys, 2019), or presented statistical cues that conflicted with lexical stress (Fernandes, Ventura, & Kolinsky, 2007). These studies did not significantly decrement learning in the whole sample ($F(1, 634) = 2.34$, $p = .13$). However, since such conditions were designed to disrupt learning, and exhibited lower means overall, they were nonetheless excluded from the main moderator analyses to ensure that no additional noise was introduced to the analyses (except for in Section 3.6, where the effects of conflicting multiple cues and prior knowledge on statistical learning were analyzed).

3.2. Overall effect of statistical learning

First, we analyzed the pooled effect size of the entire dataset to determine whether significant evidence of statistical learning was present in the sample as a whole (following the

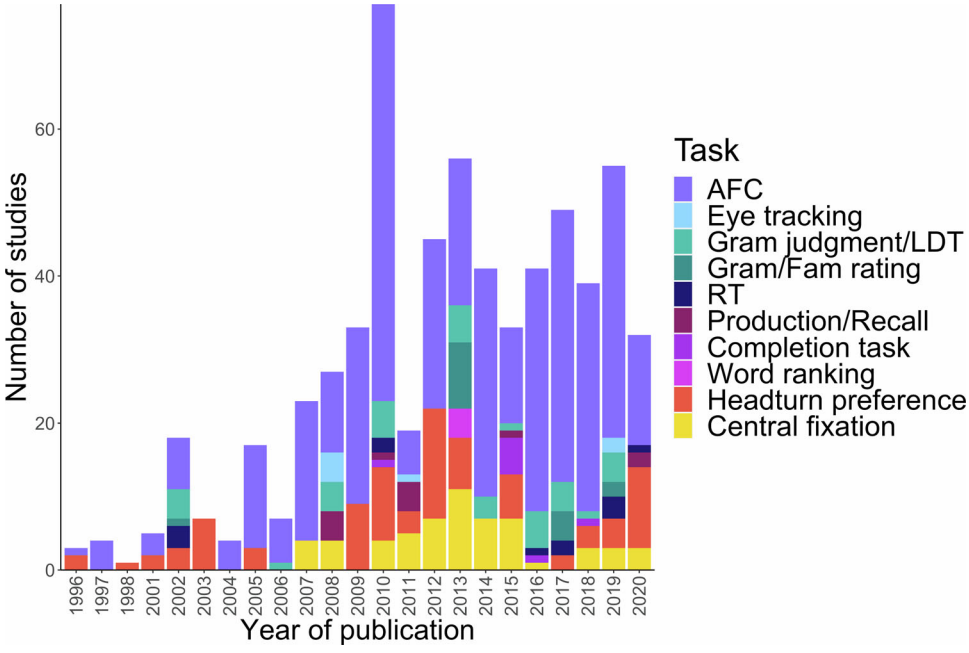

Fig. 4. The number of studies using specific tasks by year (Gram = grammaticality, LDT = lexical decision task, Fam = familiarity, RT = reaction time).

guidelines of Trotter et al., 2020). To this end, we conducted a random-effects model using the R package metafor. This model operates under the assumption that the data of each paper were collected from separate populations and thus assumes a distribution of true effect sizes rather than a single true effect size (Borenstein, Hedges, Higgins, & Rothstein, 2010). It also assumes that the effect sizes of the studies in the meta-analysis may deviate from the true effect due to sampling error as well as differences in population. As in Trotter et al. (2020), all of the models in this meta-analysis utilize the restricted maximum likelihood estimate, which accounts for multiple sources of error variation.

The results demonstrate that on the whole, significant statistical learning occurred in the sample. The mean effect size as measured by Cohen's d was 1.22 (95% Confidence interval (CI) = 1; 1.44, $p < .0001$) when controlling for paper and multiple comparisons. Significant learning was also observed within each age group (adults: $d = 1.50$, 95% CI = 1.21; 1.80, $p < .0001$; children: $d = 0.99$, 95% CI = 0.58; 1.40, $p < .0001$; infants: $d = 0.63$, 95% CI = 0.32; 0.92, $p < .0001$), while controlling for paper and multiple comparisons. The distribution of effect sizes for each study by age group can be found in Fig. 5.

In addition, a Cochran's Q test, which calculates the difference between the observed effect sizes and the estimate of the random-effects model, revealed significant heterogeneity in the entire sample ($Q(576) = 41{,}474{,}013.32$, $p < .0001$). This suggests that the studies display greatly varying degrees of learning (prediction interval: $d = -2.01$; 4.58), which may in part be explained by the age of the participants and our moderators of interest.

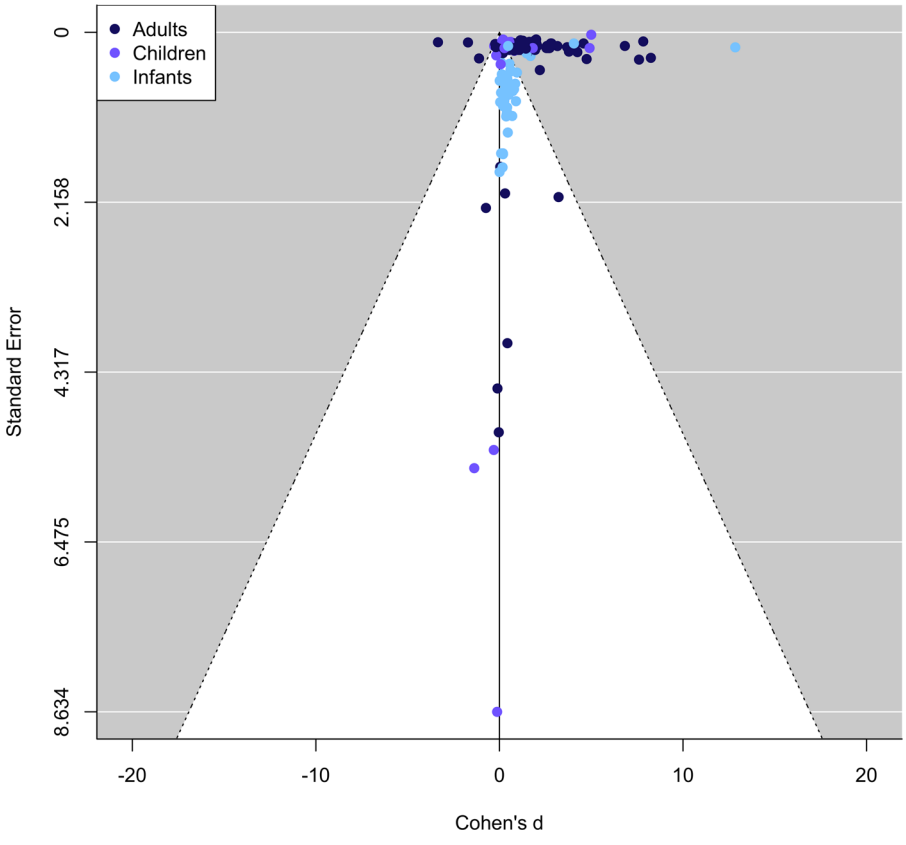

Fig. 5. Funnel plot depicting the relationship between effect size (Cohen's *d*) and standard error. The plot exhibits notable asymmetry, with many studies clustering on the higher end of the null (represented by the vertical black line) and reporting standard errors near 0.

3.3. Publication bias

Next, the data were analyzed for the presence of publication bias. Historically, published articles have tended to report studies that exhibit larger effect sizes, whereas studies with small effect sizes or null results are often left unpublished. This "file-drawer effect" may lead to an overabundance of published studies with larger effect sizes, whereas studies with small effect sizes may be underrepresented in the literature (Rothstein, Sutton, & Borenstein, 2005). This can lead to the true effect of statistical learning in the overall population being overestimated.

Although a substantial number of the studies in this meta-analysis do in fact report null results (102 studies; 123 when counting studies designed to disrupt learning), the data nevertheless demonstrate significant publication bias. The results of an Egger's test, which quantifies the asymmetry of the funnel plot in Fig. 5, reveal significant asymmetry in the model ($t(576) = -3.64$, 95% CI $= -69.93; -21.01$, $p = .0003$). Similar results hold within age groups as well (adults: $t(387) = 3.19$, 95% CI $= 23.78; 99.73$, $p = .002$;

children: $t(56) = -7.78$, 95% CI $= -154.67; -92.40$, $p < .0001$; infants: $t(131) = -2.06$, 95% CI $= -10.59; -0.27$, $p = .04$). The results from this meta-analysis must thus be interpreted in light of the publication bias evident in the sample.

3.4. Moderator analyses

We next analyzed the impact of our pre-registered moderators on study effect size: stimulus type (syllables, words, etc.), structure type (adjacent dependencies, non-adjacent dependencies, multiple-regularity languages), the number of training items, and the tasks used at test. Additional models were run on the studies that included information on the number of repetitions of items during training, the strength of the transitional probabilities or other statistical dependencies in the input language, and the number of test trials, excluding the papers that failed to report this information in text. Like in Trotter et al. (2020), we conducted these models separately for each age group to gain more reliable estimates of the moderators' contribution to the effect sizes in each population. Following their methods, we also ran separate models for each moderator to gain estimates of their unique contribution to effect size. We did not pre-register any interaction terms. Furthermore, exploratory analyses revealed that two post hoc interactions (the number of training items by the number of exposures to each item and structure type by the number of exposures) did not reach significance for any of the three age groups, and are hence not reported here.

We elected to analyze most of the structure types (studies that present adjacent dependencies, non-adjacent dependencies, and multiple-regularity languages) together, as they did not possess significant collinearities with the other moderators: all displayed variance inflation factors (VIF) below the suggested cutoff of 10 (Craney & Surles, 2002), save for in the child sample, where collinearities for two moderator analyses were controlled for (see note 8). This also allowed us to assess one of our main questions of interest: whether different language properties significantly impact effect size. Furthermore, as the data were already subsetted by age group, it was not always possible to analyze each structure type alone due to small sample sizes (e.g., there were only six non-adjacent dependency studies in the child sample, only five multiple-regularity language studies in the infant sample, etc.). We also elected to analyze the cross-situational learning studies separately from the other structures, as the task demands (e.g., mapping words to competing referents) are considerably different than studies that present linguistic materials in isolation. Furthermore, the cross-situational learning studies did display high collinearity with stimulus type (most of the studies in the dataset that present isolated words were cross-situational learning studies), which made it impossible to tease apart the unique contributions of structure from stimulus type (i.e., whether a significant finding might have been an effect of cross-situational learning itself or an effect of presenting isolated words).

For our analyses, we ran three-level multianalytic models, which account for variance at three levels: the study level, the participant level (nested within each study level), and variance between studies (Assink & Wibbelink, 2016). The use of these models was motivated by the fact that they provide more accurate estimates of the true effect size of the population by accounting for any potential interdependencies between studies, including

Table 3
Moderator contributions to effect size in adult statistical learning

Moderator category	Moderator	*F*	DFs	*p*-Value	Bayes factor
Language properties					
	Stimuli (syllables, words, etc.)	1.50	(6, 301)	.18	23.03
	Structure (adjacent, non-adjacent, multiple)	1.10	(2, 305)	.34	9.77
	Transitional probability strength	0.61	(4, 224)	.66	0
Training methods	Number of training items	0.19	(1, 306)	.66	4.08
	Number of exposures (to each item)	2.64	(1, 262)	.11	0
Testing methods	Test type (Processing vs. reflection-based)	2.42	(1, 306)	.12	0.99
	Task (AFC, RT, recall, etc.)	540.67	(7, 380)	<.0001	–
	Number of test trials	3.69	(1, 353)	.06	0
	Approach (Group-level or IDs)	2.26	(1, 306)	.13	0.96

Note. AFC, alternative forced choice; DFs, degrees of freedom; RT, reaction time.

multiple comparisons. The models were run with the moderator of interest as a fixed effect and with paper and study as a nested random effect.

3.4.1. Analysis of adult statistical learning

In terms of the language properties, structure, stimuli, and transitional probability strength did not significantly impact effect size in adults ($N = 308$ studies). The F statistics of each model, which provide an estimate of how much between-study heterogeneity each moderator accounts for, and their associated p-values can be found in Table 3.

To further evaluate the null effects of the language properties, we ran Bayes factor (BF) analyses on structure, stimulus, and transitional probability strength. This allowed us to assess whether these non-significant results lend credence to the theory that statistical learning is equally robust across different language properties in this sample (i.e., is in favor of the null hypothesis), or whether the data are simply insensitive to potential differences in these areas. A BF of greater than 3 is taken as evidence in favor of the null hypothesis, whereas a BF of less than 3 is taken as evidence for data insensitivity (Dienes, 2014). In the case of adult participants, stimulus and structure revealed BF greater than 3 (BF = 23.03 and 9.77, respectively). By contrast, transitional probability strength revealed a BF of less than 3 (BF = 0), suggesting that the current dataset may be insensitive to differences in transitional probability strength.

In terms of testing methods, the type of test (reflection-based vs. processing-based) did not influence effect size (and revealed data insensitivity; BF = 0.99), although trending in the expected direction, with processing-based measures revealing larger effect sizes. The specific choice of task did significantly impact effect size. Specifically, production/recall tasks, grammaticality/familiarity rating, RT, and eye-tracking tasks all led to significantly higher effect sizes than forced-choice tasks (both $p < .0001$). However, the number of test trials had no impact on effect size in adults nor did the study's approach (i.e., taking a group-level or individual differences approach), and both revealed data insensitivity (BF = 0 and 0.96, respectively).

Table 4
Moderator contributions to effect size in child statistical learning

Moderator category	Moderator	F	DFs	p-Value	Bayes factor
Participant	Age	0.67	(1, 32)	.42	0.0001
Language properties	Stimuli (syllables, words, etc.)	2.90	(2, 36)	.07	0.13
	Structure (adjacent, non-adjacent, multiple)	0.36	(2, 36)	.70	1.06
	Transitional probability strength	0.23	(2, 24)	.80	0
Training methods	Number of training items	0.04	(1, 37)	.85	1.62
	Number of exposures (to each item)	0.77	(1, 34)	.39	0.02
Testing methods	Test type (Processing vs. reflection-based)	33.20	(1, 37)	<.0001	—
	Task (AFC, RT, recall, etc.)	7.25	(5, 33)	<.0001	—
	Number of test trials	2.93	(1, 31)	.10	0
	Approach (Group-level or IDs)	0.38	(1, 37)	.54	1.27

Note. AFC, alternative forced choice; DFs, degrees of freedom; RT, reaction time.

3.4.2. Analysis of child statistical learning

The same set of analyses conducted on the adult data was next run on the child data ($N =$ 39 studies). We also included child age as an additional moderator to determine whether age significantly influenced effect size in this sample.

Overall, none of the language properties or training methods impacted effect size.[8] By contrast, two of the testing moderators were significant for the child dataset. The type of test (reflection based vs. processing based) significantly influenced effect sizes in children, with reflection-based tasks leading to significantly smaller effect sizes than processing-based tasks ($t(38) = -5.76$, $p < .0001$; mean effect size, reflection based = 0.86; mean effect size, processing based= 1.66). Task was also significant, with production and recall tasks leading to larger effect sizes than forced-choice tasks ($t(38) = 5.83$, $p < .0001$). The number of test trials was not significant nor was there an effect on whether the study took a group-level or individual differences approach (Table 4). The BFs for all null moderators were less than 3, suggesting that the current dataset may be insensitive to variation in these predictors.

3.4.3. Analysis of infant statistical learning

Once more, the same set of analyses was run on the infant data ($N = 127$ studies). The looking-time-based measures used in infant studies were not counted as forced-choice paradigms, since forced-choice tasks in the adult and child literature constitute reflection-based tasks: tasks that require participants to reflect over presented stimuli and make an explicit decision about which stimulus was present in the training set. Looking-time-based measures in infants reflect more implicit processing, with a child's attention being drawn to a learned or novel stimulus. They were thus categorized as processing-based task since they do not require translating implicitly acquired information into a conscious response (this is also the criteria used in the meta-analysis on artificial grammar learning conducted by Trotter et al., 2020). Since infant behavior was exclusively measured using processing-based paradigms, test type was omitted from the analyses.

Table 5
Moderator contributions to effect size in infant statistical learning

Moderator Category	Moderator	*F*	DFs	*p*-Value	Bayes factor
Participant	Age	0.18	(1, 123)	.68	0.15
Language properties					
	Stimuli (syllables, words, etc.)	1.06	(5, 121)	.39	7.21
	Structure (adjacent, non-adjacent, multiple)	2.39	(2, 118)	.10	0
	Transitional probability strength	1.65	(2, 124)	.20	1.33
Training methods	Referent mapping	0.22	(1, 125)	.64	2.07
	Number of training items	0.08	(1, 113)	.77	0
	Number of exposures (to each item)	0.79	(1, 81)	.38	0
Testing methods					
	Task (headturn preference/central fixation)	1.15	(1, 125)	.29	1.35
	Number of test trials	0.19	(1, 118)	.66	0.0001
	Approach (group level or IDs)	0.05	(1, 125)	.82	2.84

None of the language or training properties significantly influenced infant statistical learning. Stimulus had a BF of above 3, suggesting that infant statistical learning is robust across stimulus types, but structure, transitional probability strength, the number of training items, and number of exposures to each item indicated data insensitivity. None of the testing moderators impacted effect size in this population, and all revealed data insensitivity.

A supplementary moderator analysis was run to test whether the presence of referents significantly impacted effect sizes in infants. Unlike with the adults and older children, where all of the studies that trained participants to map referents were cross-situational learning tasks, most of the infant studies that involved referents were not. This made it possible to evaluate the unique contribution of referent mapping on effect size in a way that was not possible with the older children and adults in our sample. A total of 16 infant studies (not including cross-situational learning studies) involved the mapping of labels to referents. However, the presence of referents did not significantly alter effect sizes in this sample and revealed data insensitivity (Table 5).

3.5. Cross-situational learning

We performed a subgroup analysis that solely appraised cross-situational learning studies (103 studies in total; 80 adults, 18 children, and 5 infants). This decision was determined by two main factors. First, the task constraints of cross-situational learning experiments are significantly different from purely auditory statistical learning experiments (i.e., mapping linguistic items to competing referents vs. listening to linguistic material in isolation). Second, this paradigm was highly collinear with stimulus type: most studies that presented isolated words in this meta-analysis were cross-situational learning studies. While we observed significantly higher effect sizes for studies of this nature compared to the other studies in our sample (cross-situational: $M = 1.60$, $SD = 2.09$; non-cross-situational: $M = 1.13$, $SD = 1.55$, $t(127.47) = 2.16$, $p = .03$, $d = 0.28$), it was impossible to tease apart whether this difference

Table 6
Moderator contributions to effect size in cross-situational learning

Moderator Category	Moderator	*F*	DFs	*p*-Value	Bayes factor
Participant	Age group (adults, children, infants)	0.89	(2, 100)	.41	1.19
Language properties					
	Stimuli (isolated words, multiple words, etc.)	0.85	(2, 100)	.43	1.14
Training methods	Number of training items	0.92	(1, 101)	.34	1.03
	Number of exposures (to each item)	0.30	(1, 81)	.59	0
Testing methods	Task (AFC, word ranking, etc.)	0.18	(2, 96)	.84	0
	Number of test trials	0.02	(1, 85)	.90	0
	Approach (group level or IDs)	2.81	(1, 101)	.10	0.35

Note. AFC, alternative forced choice; DFs, degrees of freedom.

was due to the presentation of isolated words, the presence of referents, the cross-situational nature of stimulus presentation, or a combination of these factors.

Test type (reflection or processing based) was not included as a moderator, as there was a near-perfect collinearity with age group (i.e., all of the studies that utilize processing-based methods were infant studies, with the exception of one). Similarly, transitional probability strength was also omitted, as all the studies presented items with fixed transitional probabilities.

Of all the potential moderators, none significantly influenced cross-situational learning effect sizes. The BFs of all of the null moderators were below 3, suggesting that the data may be insensitive to variation in cross-situational learning among the targeted dimensions (Table 6). When the analyses were redone on the adult sample only (excluding infants and children), the same pattern of data insensitivity emerged, suggesting that it was not the inclusion of the infants and child data in these models that led to these effects.

3.6. The impact of multiple cues and prior knowledge on learning

The presence of multiple cues has long been hypothesized to significantly impact language acquisition (e.g., Bates & MacWhinney, 1989; Gleitman & Wanner, 1982; Morgan, 1996). For this reason, we were interested in assessing the impact of multiple cues on statistical learning effect sizes.

In total, 179 studies in the meta-analysis reported manipulating multiple cues (153 congruent cues, 26 conflicting). These cues either took the form of auditory cues, visual cues, social cues (classified as such by the authors of the original papers), or conflicting cues. Conflicting cues are mismatched with the statistical information of the presented language. Examples include the asynchronous presentation of potential referents (Lavi-Rotbain & Arnon, 2018), or where lexical stress cues that denote word boundaries in one's native language contradict the learned language's statistical information (Thiessen & Erickson, 2013). A full list of the cues present in the meta-analysis can be found in Table 7.

Table 7
Multiple cue types

Cue domain	Cue Type	Number of Studies
Auditory ($N = 140$)	Auditory nonlinguistic	8
	Combined auditory cues	6
	Conflicting cues	15
	Duration	1
	Lexical/sublexical	17
	Morphological	6
	Pauses	20
	Phonological	14
	Phonotactic	3
	Pitch	16
	Prosodic	3
	Speaker (voice)	9
	Stress cues	18
	Tone	4
Visual ($N = 30$)	Conflicting cues	10
	Speaker (video or image)	3
	Visual cue	17
Social ($N = 9$)	Eye gaze	8
	Conflicting cues	1
Total		*179*

Overall, there was a significant effect of congruent multiple cues on effect size (excluding conflicting cues; $F(1, 634) = 319.47$, $p < .0001$), with the presence of multiple cues leading to an increase in effect size. There was also an effect of modality ($F(3, 632) = 109.07$, $p < .0001$), with congruent auditory and visual cues on the whole leading to significantly larger effect sizes.

To disentangle the results of modality, an analysis of multiple cue type was performed on the full dataset of 636 studies (including studies designed to disrupt learning, which allowed us to gauge the effect of conflicting cues on effect size). The effect of cue type was significant, with auditory pause cues, in particular, leading to larger effect sizes ($F(16, 607) = 34.86$, $p < .0001$). Within this model, conflicting cues did not lead to a significant decrement in effect size.

In addition to the presence of multiple cues, we also analyzed the data for the effect of prior knowledge (22 studies in total). Studies that investigate the impact of prior knowledge manipulate properties of the artificial language, such that they are either congruent or incongruent with the learners' native language statistics (e.g., the phonotactics of the trained language either resemble or violate the phonotactics of the learner's native language; Finn & Hudson Kam, 2008; 2015). We found no effect of either congruent or incongruent prior knowledge on effect size, with a BF greater than 3 ($F(2, 633) = 1.34$, $p = .26$, $BF = 10.01$), suggesting that the manipulation of prior knowledge congruency had no meaningful impact on learning in this dataset.

4. Discussion

Over the last 25 years, the study of statistical learning has greatly enriched our understanding of language acquisition and cognition as a whole. In the current paper, we tested the potential strengths and limitations of auditory-linguistic statistical learning across development, language properties, and methodologies. The goal of these analyses was to test common theoretical assumptions about statistical learning, including the generality of this phenomenon across structures and stimuli, the impact of input qualities, and the influence of multiple cues. Our pre-registered hypotheses predicted that statistical learning would be influenced by many of these moderators. However, while the presence of multiple cues came out as a significant predictor, as did task in adults and children, overall, the analysis revealed considerable data insensitivity and limited variability in many of the null moderators, such as in transitional probability strength. In addition, the presence of publication bias in the current sample means that all results must be interpreted with this consideration in mind.

Statistical learning is a key construct in cognitive science across domains and modalities, especially in the study of language. However, while typically discussed and investigated as a general-purpose learning mechanism that unites many diverse behaviors, research suggests that statistical learning varies considerably across different input features, even within modalities and domains (R. Frost et al., 2015). This variability bears to question whether such divergent forms of learning do in fact deploy the same computations or whether the differences between tasks necessitate distinct computations (Thiessen, 2017). In terms of the theoretical implications of our results, this meta-analysis suggests that the different studies of auditory-linguistic statistical learning in this sample can achieve largely similar degrees of learning despite differences in stimuli in infants and adults. Adult statistical learning is also resilient across linguistic structures, suggesting robust learning of different types of dependency patterns—whether this is the outcome of a developmental trend will require more research given the data insensitivity of the infant and child populations.

From a theoretical viewpoint, it is important to note that our meta-analysis does not resolve the issue of whether statistical learning consists of one or more mechanisms—it only shows that statistical learning is robust across different auditory-linguistic stimuli and structures in certain populations. It also does not speak to how statistical learning operates in different modalities and domains. Thus, although statistical learning is likely to recruit a host of cognitive abilities, including attentional (Toro, Sinnett, & Soto-Faraco, 2005), memory (Thiessen, Kronstein, & Hufnagle, 2013), and modality-specific processes (Conway & Christiansen, 2005), its power across linguistic structures highlights how such learning might support an array of complex language-related abilities in adults. These results provide strong support for the role of statistical learning in multiple aspects of language acquisition, in accord with a long line of theories advocating its fundamentality.

Perhaps the most surprising results were those concerning the lack of impact of transitional probability strength on statistical learning across all three populations. Differences in item recognition based on transitional probability strength is one of the core assumptions of statistical learning, dating back to the original Saffran et al. (1996) study: Infants were able to reliably discern words from the input (with transitional probabilities of 1) from

partwords that straddle item boundaries, which occur in the input but exhibit weaker transitional probabilities. This null effect in the meta-analysis likely derives from the fact that across the entire sample, the bulk of the studies trained participants on linguistic stimuli comprising transitional probabilities of 1 (see Fig. 3). The overrepresentation of such transitional probabilities has critical implications for the field of statistical learning. First, many hypotheses surrounding statistical learning may be predicated upon conditions where syllable co-occurrence information is perfectly uniform. This may also tie into why many studies fail to find reliable individual differences between statistical learning of artificial languages and language learning in the real world, where natural languages are infinitely more varied in terms of their statistical properties (Isbilen et al., 2022). It is thus imperative for future research to expand the diversity of the transitional probabilities that participants are trained on. This may lead to stronger theories about how statistical learning operates across different inputs, both in the lab and in the natural world.

Furthermore, effect sizes remained largely unchanged by the number of training items in the language (with BFs for this moderator being above 3 for adults), and the number of exposures to these items (BFs indicated data insensitivity for this moderator across all populations). These results contradict research suggesting that such factors critically influence the strength of statistical learning (Frank et al., 2010) and have far-reaching implications for children's natural language acquisition (Hart & Risley, 2003). It is possible that adult learning is more resilient to larger learning loads than younger populations, although the null effect of the number of exposures across the three populations is puzzling, given the legacy of psycholinguistic research underscoring the importance of frequency in language learning and processing (for reviews, see, e.g., Ambridge, Kidd, Rowland, & Theakston, 2015; Diessel, 2007; Ellis, 2002; Gries & Divjak, 2012). Although there was greater variability in these moderators than for transitional probability strength, these results may still reflect the tendency of most studies to utilize a narrow set of simplified languages. The limited diversity of the language and training properties may tie into the publication bias[9] found in the sample: using low transitional probability strengths, many training items, and/or a small number of exposures to each item are likely to significantly impair learning, which may lead to null results that are often left unpublished. The relatively brief length of statistical learning studies may also contribute, as learners are limited in what they are able to learn in a short period of time, and studies conducted with children and infants are required to be brief to fit the attention spans of these populations. A fruitful area for future research may include diversifying the input languages or investigating ways to bolster learning of more difficult to acquire material.

In adults and children, task was a significant predictor of statistical learning effect size, suggesting that the methods used at test have a profound bearing on the effects observed in the literature. Processing-based measures led to consistently higher effect sizes than reflection-based measures in children (and trended in that direction with adults). Forced-choice tasks in particular resulted in markedly weaker effect sizes in both populations. This finding is especially noteworthy given that these tasks dominate the statistical learning literature, and since an increasing number of studies have brought to light the poor reliability, sensitivity, and internal consistency of forced-choice paradigms (Arnon, 2020; Isbilen et al., 2020).

Production and recall-based tasks appear to be the most efficacious in testing adult and child learning (although grammaticality/familiarity rating tasks were robust in adults as well), in line with studies showing that recall measures are more reliable and consistent measures of statistical learning in both populations (Isbilen et al., 2020; Kidd et al., 2020c). These results thus further question the efficacy of reflection-based forced-choice tasks in measuring what is widely understood to be an implicit form of learning. Reflection-based tasks require participants to ruminate over tacit knowledge and then decide how to respond. The scores of these tasks may therefore capture some degree of individual differences in decision processes, in addition to measuring the behavior of interest (Christiansen, 2019). This feature of reflection-based tasks may wash out subtle variation in learning, potentially leading researchers to either over- or underestimate the true effect of this phenomenon in the population (Siegelman et al., 2017). Indeed, the classic two-alternative forced-choice task has been shown to only capture knowledge that the participant is consciously aware of, even when RTs and neural activity indicate learning of material that is not available to conscious awareness (Batterink, Reber, Neville, & Paller, 2015). The results from this meta-analysis further endorse the notion that processing-based tasks provide a more powerful index of learning in children, which may in part stem from the fact that they circumvent many of the confounds associated with reflection-based tasks. The adoption of more sensitive and reliable measures may serve to enhance future theories on statistical learning. The shortcomings of these tasks may also in part explain the prevalent data insensitivity of the other moderators observed in the child sample.

Like the child population, the cross-situational learning sample revealed substantial data insensitivity despite its relatively large sample size (103 studies). However, on the whole, these studies did appear to yield larger effect sizes than non-cross-situational studies, though it is difficult to determine precisely which factors drove the finding. It may in part be explained by the fact that many (but not all) of the cross-situational learning studies in the sample presented two or more words separated by pauses, which could potentially be easier to acquire than words embedded in continuous speech. Indeed, the meta-analysis confirmed that pauses inserted at word boundaries facilitate the learning of structure from continuous speech (e.g., Endress & Mehler, 2009; Finn & Hudson Kam, 2008), suggesting that hearing words in relative isolation may bootstrap acquisition. It has also been shown in corpus analyses of child-directed speech that the frequency with which caretakers use a given word in isolation positively predicts children's comprehension and production of that word (Brent & Siskind, 2001). Isolated words may therefore work to scaffold early vocabulary development, perhaps in part by solving what is often cited as one of the primary obstacles to word acquisition: the challenge of identifying word boundaries in continuous speech (see e.g., Monaghan & Christiansen, 2010, for a computational account). It is also possible that the inclusion of visual scenes may facilitate the task and make it more interesting for participants (perhaps because it more closely resembles language learning in natural settings; e.g., Bergelson & Swingley, 2012; Tincoff & Jusczyk, 1999, 2012).

Our final set of analyses revealed that the presence of multiple cues significantly increased statistical learning effect sizes. In line with a long tradition of research emphasizing its merits in assisting infants with natural language acquisition (e.g., Bates & MacWhinney, 1989; Gleitman & Wanner, 1982; Morgan, 1996), multiple cue integration also reliably bootstraps

language acquisition in the laboratory. This was particularly the case for auditory and visual cues. Pause cues (e.g., brief silences between items) were particularly helpful for learning. Conflicting cues that distract from the statistics of the input language did not decrease learning, similar to the findings of the meta-analysis by Black and Bergman (2017) on infant statistical learning. Given that spurious cues unrelated to the patterns to be learned are likely to be ubiquitous, from a theoretical perspective this result thus points to the kind of robustness needed for multiple cue-based statistical learning to play a foundational role in language acquisition. While trending in the expected direction, prior knowledge did not influence statistical learning effect sizes, regardless of whether the trained language was congruent or incongruent with aspects of the participants' native language. Further work may thus be needed in order to isolate the effects of prior learning on the acquisition of novel linguistic materials.

The results of this meta-analysis bring to light several key considerations that may benefit future research. On the methodological front, our results provide important insights into several features of experimental design that may help researchers develop more reliable statistical learning studies. In terms of training procedures, the inclusion of multiple cues may contribute to the robustness of learning. Such cues can even be used in conjunction with cross-situational learning tasks (e.g., Frinsel, Trecca, & Christiansen, 2020; Monaghan et al., 2019; Walker, Monaghan, Schoetensack, & Rebuschat, 2020), offering a more ecological approach to studying language acquisition than presenting isolated linguistic input. Furthermore, processing-based measures appear to provide more robust assessments of learning in children (and trend in that direction with adults), suggesting that a shift away from reflection-based measures, and forced-choice tasks specifically, may improve our understanding of individual differences in statistical learning across development. In particular, testing using production (e.g., Hopman & MacDonald, 2018; Perek & Goldberg, 2015; Wonnacott, 2011; Wonnacott et al., 2008), and recall (e.g., Botvinick & Bylsma, 2005; Conway et al., 2010; Isbilen et al., 2020; Kidd et al., 2020; Majerus, van der Linden, Mulder, Meulemans, & Peters, 2004) may lead to demonstrably stronger results (although grammaticality/familiarity rating tasks are also strong options for adults). The widespread employment of such methods may thus provide a clearer picture of learning in adults and children and in turn critically inform the development of statistical learning theories.

Theoretically speaking, our meta-analysis has provided substantial support for theories of language that argue for a key role of statistical learning in the acquisition of different kinds of linguistic structures (e.g., Saffran, 2003). More research is needed, however, to flesh out this theoretical conclusion—especially given the publication bias we observed in our results. We suggest that a fruitful way forward to further elaborate the role of statistical learning in theories of language acquisition is to investigate how individual differences in statistical learning map onto variation across individuals in their language skills. Following Bogaerts et al. (2022), we argue that this might be best done by an approach that utilizes corpus analyses to determine the distributional patterns of specific aspects of linguistic structure and then conduct statistical learning experiments that closely target similar patterns in artificial miniature languages. This dovetails with recent empirical work illustrating how aligning the kinds of structures targeted between artificial and natural language learning tasks provide stronger links between the two phenomena, and better informs how they are related (Isbilen et al.,

2022). Such studies would be expected to provide stronger evidence for the theoretical link between statistical learning and native language function (and dysfunction) and may also help inform studies of second language acquisition.

In conclusion, we performed a large-scale meta-analysis of research on auditory-linguistic statistical learning spanning the last 25 years. The results reveal that adult statistical learning is largely multipurpose across linguistic stimuli and structures, although further work is required to determine whether this extends to developmental populations. We found that processing-based tasks led to significantly larger statistical learning effect sizes in children, with similar findings in adults, and are significantly boosted by additional cues. However, we also uncovered several methodological limitations that critically impact the effect size of statistical learning and its potential theoretical implications. We, therefore, hope that these results may help improve future experimentation and the scientific community's comprehension of statistical learning as a whole.

Acknowledgments

This research was in part supported by the National Science Foundation Graduate Research Fellowship Program (#DGE-1650441) and the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health (#F32HD104542) awarded to ESI. The authors thank Stephen Parry from the Cornell University Statistical Consulting Unit for his help with the analyses presented in this paper as well as Collin Edwards, Padraic Monaghan, Felix Thoemmes, Riccardo Fusaroli, and Christopher Cox. We also thank Noam Siegelman, Tony Trotter, Pierre Perruchet, and Richard Aslin for their helpful comments on earlier drafts as well as the many authors who contributed data for the meta-analysis.

Open Research Badges


This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/7vkdt/.

Notes

1 While auditory-linguistic statistical learning is likely most pertinent to spoken language acquisition, it is important to acknowledge that visual statistical learning also plays a central role in language, particularly for sign language acquisition and the development of reading-related skills (although see Qi, Sanchez, Georgan, Gabrieli, & Arciuli, 2019, for a discussion of the importance of auditory statistical learning for reading).

2 Whereas studies investigating adjacent and non-adjacent dependencies typically only measure learning of a single kind of regularity (e.g., adjacent dependencies: Saffran et al., 1996; non-adjacent dependencies: Endress & Mehler, 2009), other studies involve multiple regularities implemented by grammars. The latter type of studies includes work in the artificial grammar learning literature (e.g., Conway, Bauernschmidt, Huang, &

Pisoni, 2010) as well as statistical learning studies involving artificial languages manipulating syntactic patterning (e.g., Saffran, 2002), and studies that present adjacent and non-adjacent regularities together (e.g., Vuong et al., 2016).

3 The keywords "language" AND "statistical learning" were pre-registered. However, upon editorial review, a broadening of the search terms was requested, and the meta-analysis was redone to include the term "distributional learning." This is why the final search terms deviate from the pre-registration. We did not include the search terms "auditory statistical learning," as this line of work also focuses on the acquisition of tone and non-linguistic sound sequences. We also did not search for "artificial grammar learning," as many such studies focus on the acquisition of visual items (e.g., Reber, 1967) and would have limited the search to a certain kind of structure. Furthermore, as this line of work predates Saffran et al. (1996), the term captured an abundance of work not modeled on this paradigm (although see the meta-analysis on artificial grammar learning by Trotter et al., 2020). The term implicit learning was excluded for similar reasons, while "artificial language learning" missed many key studies. Furthermore, we did not include the terms "word segmentation" or "cross-situational learning," as it was the goal of the meta-analysis to capture as wide an array of structures and methodologies as possible.

4 It is possible that this equation may overestimate the effect of learning, if participants learn at test.

5 We analyzed the mean effect sizes of the novelty versus familiarity preference data and found no significant differences. Resigning the looking time means thus has not influenced the analyses.

6 In our pre-registration, we had proposed using foil type as a moderator. However, due to near-perfect collinearities between foil and task, these analyses were dropped from the analysis. We had also proposed a set of exploratory analyses that evaluated the pattern of findings on the correlations between individual differences in statistical learning and other aspects of cognition. However, the information on this was sparse in the dataset. Null correlations were often not reported, and the tasks used were highly varied, which precluded the deduction of any meaningful conclusions for any of the age groups. They were thus not included in the final meta-analysis. Similarly, the data for the pre-registered moderator of natural language stimuli was also sparse ($N = 13$ each across the three age groups). We found no significant effect of natural language stimuli ($F(1, 634) = 1.94$, $p = .16$, $BF = 1.68$), with the sample revealing data insensitivity.

7 Some non-adjacent dependency studies utilize monosyllabic words. These were logged as words rather than syllables, based on what the original authors of the paper classified their stimuli as.

8 Multiple regularity languages exhibited a variance inflation factor above the suggested cutoff of 10 (VIF = 18.52) in the child dataset, as such studies exclusively presented words. Further, studies that presented transitional probabilities of 1 were frequently (but not always) studies that presented adjacent dependencies and had a variance inflation factor just above the cutoff (VIF = 10.23). However, the same null results and data insensitivity emerged when these factors were controlled for in the models (stimuli:

$F(1, 23) = 3.31$, $p = .08$, BF = 0; transitional probability strength: $F(1, 15) = 0.58$, $p = .46$, $BF = 0$).

9 It is important to note that the analyses cannot definitively specify whether it was the homogeneity of the language and training properties that caused the publication bias in this sample. It is possible that researchers have only tested a limited array of stimuli, even in unpublished work.

References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–273.

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. https://doi.org/10.3758/s13428-019-01205-5

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: a step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*, 154–174.

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*, 1382–1407.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, *20*, 395–418.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The* crosslinguistic study of sentence processing (pp. 3–76). New York, NY: Cambridge University Press.

Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science*, *28*, 921–928.

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

Batterink, L. J., & Paller, K. A. (2019). Statistical learning of speech regularities can occur outside the focus of attention. *Cortex*, *115*, 56–71.

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*, 65–78.

Benitez, V. L., Yurovsky, D., & Smith, L. B. (2016). Competition between multiple words for a referent in cross-situational word learning. *Journal of Memory and Language*, *90*, 31–48.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*, 3253–3258.

Black, A., & Bergmann, C. (2017). *Quantifying infants' statistical word segmentation: A meta-analysis*. In *39th Annual Meeting of the Cognitive Science Society* (pp. 124–129). Austin, TX: Cognitive Science Society.

Bogaerts, L., Siegelman, N., Christiansen, M. H., & Frost, R. (2022). Is there such a thing as a 'good statistical learner'? *Trends in Cognitive Sciences*, *261*, 25–37.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111.

Botvinick, M., & Bylsma, L. M. (2005). Regularization in short-term memory for serial order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 351–358.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33–B44.

Christiansen, M. H. (2019). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*, 468–481. https://doi.org/10.1111/tops.12332

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*, 356–371.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 24.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*, 391–403.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy*, *21*, 648–667.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*, 108–127.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*, 351–367.

Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. A. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra*, *2*, 14. doi: https://doi.org/10.1525/collabra.41

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*, 429–448.

Fernandes, T., Ventura, P., & Kolinsky, R. (2007). Statistical information and coarticulation as cues to word boundaries: A matter of signal quality. *Perception & Psychophysics*, *696*, 856–864.

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*, 477–499.

Finn, A. S., & Hudson Kam, C. L. (2015). Why segmentation matters: Experience-driven segmentation errors impair "morpheme" learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1560.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499–504.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, *134*, 521–537.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2010). Mere exposure alters category learning of novel objects. *Frontiers in Psychology*, *1*, 40.

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation. *Experimental Psychology*, *62*, 346–351.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *1172*, 107–125.

Frinsel, F. F., Trecca, F., & Christiansen, M. H. (2020). *The picture guessing game: The role of feedback in active artificial language learning*. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 1–7). Austin, TX: Cognitive Science Society.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible directions. *Psychological Bulletin*, *145*, 1128–1153.

Frost, R. L. A., Jessop, A., Durrant, S., Peter, M. S., Bidgood, A., Pine, J. M., Rowland, C. F., & Monaghan, P. (2020). Non-adjacent dependency learning in infancy, and its link to language development. *Cognitive Psychology*, *120*, 101291.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *193*, 117–125.

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

Getz, H., Ding, N., Newport, E. L., & Poeppel, D. (2018). Cortical tracking of constituent structure in language acquisition. *Cognition*, *181*, 135–140.

Gleitman, L. R., & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3–48). Cambridge, England: Cambridge University Press.

Gómez, D. M., Bion, R. A., & Mehler, J. (2011). The word segmentation process as revealed by click detection. *Language and Cognitive Processes*, *26*, 212–223.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*, 183–206.

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

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.

Graf Estes, K. (2012). Infants generalize representations of statistically segmented words. *Frontiers in Psychology*, *3*, 447.

Gries, S. T., & Divjak, D. (Eds.). (2012). *Frequency effects in language learning and processing*. Berlin: Gruyter Mouton.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*(1), 4–9.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *783*, B53–B64.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*, 93–106.

Hopman, E. W., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science*, *29*, 961–971.

Hsu, H. J., Tomblin, J. B., & Christiansen, M. H. (2014). Impaired statistical learning of non-adjacent dependencies in adolescents with specific language impairment. *Frontiers in Psychology*, *5*, 175.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, *5*, 69–95.

Isbilen, E. S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2022). Statistically-based chunking of nonadjacent dependencies. *Journal of Experimental Psychology: General*, Advance online publication. https://doi.org/10.1037/xge0001207

Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition*, *225*, 105123. https://doi.org/10.1016/j.cognition.2022.105123

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*, *44*, e12848.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Karuza, E. A., Farmer, T. A., Fine, A. B., Smith, F. X., & Jaeger, T. F. (2014). On-line measures of prediction in a self-paced statistical learning task. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 725–730). Austin, TX: Cognitive Science Society.

Kavakci, M., & Dollaghan, C. (2019). A new method for studying statistical learning in young children. *Journal of Speech, Language, and Hearing Research*, *62*, 2483–2490.

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.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*, 1134–1140.

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*, *60*, 3474–3486.

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2020). Children with developmental language disorder have an auditory verbal statistical learning deficit: Evidence from an online measure. *Language Learning*, *70*, 137–178.

Lavi-Rotbain, O., & Arnon, I. (2018). Developmental differences between children and adults in the use of visual cues for segmentation. *Cognitive Science*, *42*, 606–620.

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology*, *94*, 67–84.

Majerus, S., van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language*, *51*, 297–306.

Mayo, J., & Eigsti, I. M. (2012). Brief report: A comparison of statistical learning in school-aged children with high functioning autism and typically developing peers. *Journal of Autism and Developmental Disorders*, *42*, 2476–2485.

Mermier, J., Quadrelli, E., Turati, C., & Bulf, H. (2022). Sequential learning of emotional faces is statistical at 12 months of age. *Infancy*, *273*, 479–491.

Mirman, D., Graf Estes, K., & Magnuson, J. S. (2010). Computational modeling of statistical learning: Effects of transitional probability versus frequency and links to word learning. *Infancy*, *15*, 471–486.

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

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 302–331.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*, 264–269.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*, 545–564.

Monaghan, P., Schoetensack, C., & Rebuschat, P. (2019). A single paradigm for implicit and statistical learning. *Topics in Cognitive Science*, *11*, 536–554.

Morgan, J. L. (1996). Prosody and the roots of parsing. *Language and Cognitive Processes*, *11*, 69–106.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, *19*, 498–550.

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2010). Learnability of embedded syntactic structures depends on prosodic cues. *Cognitive Science*, *34*, 338–349.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162.

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, *492*, 85–117.

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*, 139–146.

Perek, F., & Goldberg, A. E. (2015). Generalizing beyond the input: The functions of the constructions matter. *Journal of Memory and Language*, *84*, 108–127.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238.

Pons, F., & Toro, J. M. (2010). Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition*, *116*, 361–367.

Poulin-Charronnat, B., Perruchet, P., Tillmann, B., & Peereman, R. (2017). Familiar units prevail over statistical cues in word segmentation. *Psychological Research*, *81*, 990–1003.

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

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*, e12593.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: Wiley.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition*, *81*, 149–169.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*, 172–196.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *124*, 110–114.

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*, 27–52.

Scopus. (2019). Retrieved from https://www.scopus.com.

Scott, R. M., & Fisher, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, *122*, 163–180.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160059.

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

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

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

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Thiessen, E. 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*, *372*(1711), 20160056.

Thiessen, E. D., & Erickson, L. C. (2013). Discovering words in fluent speech: The contribution of two kinds of statistical information. *Frontiers in Psychology*, *3*, 590.

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*, 792–814.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology*, *39*, 706–716.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*, 172–175

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*, 432–444.

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

Trecca, F., McCauley, S. M., Andersen, S. R., Bleses, D., Basbøll, H., Højen, A., Madsen, T. O., Bjønness, I. S. R., & Christiansen, M. H. (2019). Segmentation of highly vocalic speech via statistical learning: Insights from a cross-linguistic study of Danish, Norwegian, and English. *Language Learning*, *69*, 143–176.

Trotter, A. S., Monaghan, P., Beckers, G. J., & Christiansen, M. H. (2020). Exploring variation between artificial grammar learning experiments: Outlining a meta-analysis approach. *Topics in Cognitive Science*, *12*, 875–893.

Van den Bos, E., Christiansen, M. H., & Misyak, J. B. (2012). Statistical learning of probabilistic nonadjacent dependencies by multiple-cue integration. *Journal of Memory and Language*, *67*, 507–520.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Vuong, L. C., Meyer, A. S., & Christiansen, M. H. (2016). Concurrent statistical learning of adjacent and nonadjacent dependencies. *Language Learning*, *66*, 8–30.

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, *70*, 221–254.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*, 1–14.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165–209.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*, 244–262.

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*, 1–22.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*, 891–921.