Distributional Statistical Learning: How and How Well Can It Be Measured?

Bethany Growns
School of Social and Behavioral Sciences, Arizona State University
4701 W Thunderbird Rd, Glendale, Arizona 85069, USA

Erwin J. A. T. Mattijssen
Netherlands Forensic Institute
PO Box 24044, 2490 AA The Hague, The Netherlands
Radboud University Nijmegen, Behavioural Science Institute
PO Box 9104, 6500 HE Nijmegen, The Netherlands

Abstract

Individuals are readily able to extract and encode statistical information from their environment (or *statistical learning*). However, the bulk of the literature has primarily focused on *conditional statistical learning* (i.e. the ability to learn joint and conditional relationships between stimuli), and has largely neglected *distributional statistical learning* (i.e. the ability to learn the frequency and variability of distributions). In this paper, we investigate how and how well distributional learning can be measured by exploring the relationship between and psychometric properties of two measures: discrimination judgements and frequency estimates. Reliable performance was observed in both measures across two different distributional learning tasks (natural and artificial). Discrimination judgements and frequency estimates also significantly correlated with one another in both tasks, and performance on all tasks accounted for the majority of variance across tasks (55%). These results suggest that distributional learning can be measured reliably, and may tap into both the ability to discriminate between relative frequencies and to explicitly estimate them.

Keywords: statistical learning, distributional learning, conditional learning, individual differences, psychometrics

Introduction

One of the ways humans process the overload of sensory information they receive from their environment is to extract patterns and regularities. The ability to do this (known as *statistical learning*) is thought to underlie many basic perceptual and cognitive processes, such as categorisation and language acquisition (Siegelman & Frost, 2015). Individuals are able to extract many different forms of statistical regularities – from joint and conditional relationships between stimuli (e.g. *A* co-occurs with *B* in time or space; or conditional statistical learning; Fiser & Aslin, 2002; Turk-Browne, Jungé, & Scholl, 2005) to the frequency and variability of distributions in the environment (e.g. *C* occurs more often than *D*; or distributional statistical learning; Thiessen & Erickson, 2013; Zacks & Hasher, 2002). Whilst conditional and distributional statistical learning processes are interrelated (Growns, Siegelman, & Martire, under review), the bulk of contemporary statistical learning research has focused only on conditional learning (see Frost, Armstrong, & Christiansen, 2019 for a review). There has been limited focus on distributional learning as a construct – even less is known about *how* and *how well* we can measure the ability to extract distributional regularities from the environment.

Distributional statistical learning research has largely focused on how it facilitates language and object discrimination. Exposure to bimodal distributions of sounds (e.g. sounds from a distribution of 'da' to 'ta') or objects (e.g. faces morphed along a 'continuum') typically facilitates later discrimination of these stimuli, compared to exposure to a unimodal distribution (i.e. where stimuli occur more frequently in the 'middle' of a distribution; Altvater-Mackensen, Jessen, & Grossmann, 2017; Escudero & Williams, 2014; Junge, van Rooijen, & Raijmakers, 2018; Maye, Weiss, & Aslin, 2008).

Whilst some of the factors surrounding distributional learning are beginning to be understood, there has been limited empirical investigation into *how* it is measured. There are multiple ways of examining distributional learning that have not been studied in parallel – from eliciting explicit frequency estimates to judgements in forced-choice discrimination tasks (Hasher & Zacks, 1984). Individuals are generally proficient at discriminating between relative distributional frequencies (Growns & Martire, in press; Growns et al., under review), but are typically poor at precisely providing accurate frequency estimates and judging the base rates of events (Bar-Hillel, 1980; Brenner, Koehler, Liberman, & Tversky, 1996; Lee & Danileiko, 2014; Martire, Growns, & Navarro, 2018). From a theoretical perspective, both discrimination judgements and frequency estimates measure the ability to learn the frequency and variability of distributions in the environment. However, it is not known whether these two measures tap into separate distributional learning abilities, or are part of the same theoretical construct. Are better 'discriminators' also better 'estimators'? No research has investigated the relationship between these different forms of distributional learning.

There has also been limited empirical investigation into *how well* we can measure distributional learning. Research has identified psychometric problems in conditional statistical learning measures: simpler two-alternate forced-choice (2AFC) measures have poorer reliability and stability than more complex multiple-alternate measures (Siegelman,

Bogaerts, & Frost, 2017). Studies that use 2AFC conditional learning measures generally show psychometric properties that fall below typically recommended values (Arnon, 2019; Siegelman et al., 2017; Streiner, 2003). Low reliability increases a measure's error variance and limits the ability to find individual variability and differences (Siegelman et al., 2017). Yet there has been limited empirical investigation into the reliability of distributional learning measures.

Investigating *how* and *how well* we can measure distributional learning is critical to our broader understanding of statistical learning, as well as the role it may play in other functions. Distributional and conditional statistical learning have been theorised to be underpinned by separate, but inter-related, memory processes (Thiessen & Erickson, 2013; Thiessen, Kronstein, & Hufnagle, 2013). Distributional learning may also play an important role in other cognitive processes. For example, distributional statistical information provides important diagnostic information in visual identification tasks such as forensic visual comparison tasks or disease detection in radiology scans (Bruce & Tsotsos, 2009; Busey, Nikolov, Yu, Emerick, & Vanderkolk, 2016; Growns & Martire, in press). Yet a limited understanding of how distributional learning is best measured hinders the ability to empirically explore its predictive validity of other cognitive functions.

In this paper, we investigate how and how well distributional learning can be measured. We examine how well distributional learning can be measured by examining the reliability of discrimination judgement and frequency estimate measures. We examine how distributional learning can be measured by exploring whether these measures are part of a unified distributional learning ability, or separate sub-processes of the same theoretical construct. We would expect to see significant associations between the measures if the tasks measure the same ability, but no association if they measure different abilities. In this study, we investigate distributional statistical learning of two types of stimuli (natural and artificial) to examine whether such associations might generalise.

Method

Design

Participants completed two statistical learning tasks in a set order to minimise error variance (Mollon, Bosten, Peterzell, & Webster, 2017): a natural task containing real-world stimuli; and then an artificial task with generated stimuli. Participants first completed an exposure phase and then a test phase for each task. The study pre-registration, data, analysis scripts and supplementary materials can be found at https://osf.io/2ux9q/.

Participants

Participants were 110 undergraduates from a large university in south-western United States who received course credit for their participation. The participants were 22.75 years of age (*SD* = 8.15, *min* = 18, *max* = 65) and the majority reported they were female (80.91%; 17.27% = male; 1.82% = transgender or gender-fluid). Participants were required to have normal or corrected-to-normal vision in order to participate.

Exclusion Criteria We excluded 37 participants who did not meet a 2/4 attention-check[1] correct inclusion threshold (compared to the pre-registered 3/4 threshold that only 40% (*n* = 58) of the whole sample met). We report analyses of the sample with this threshold to better represent the collected sample, but the analyses of both samples did not meaningfully differ (see supplementary materials).

Materials

Participants completed the experiment on an online survey platform, Qualtrics (2005). They were instructed to adjust their browser zoom so images could be fully seen and to only take breaks at appropriate points when prompted.

Artificial Task and Dependent Measures

Exposure Phase Participants viewed 60 artificial patterns manipulated to contain features that occurred with different frequencies (as in Growns & Martire, in press): ten features with a frequency of 0.1; one feature with a frequency of 0.3; one feature with a frequency of 0.7; and one feature with a frequency of 1.0. As features with a frequency of 0.1, 0.3 and 0.7 always co-occurred with the 1.0 feature, these frequencies resulted in three joint probabilities: 0.1; 0.3; and 0.7. Three features appeared in each exemplar on opposing pattern 'arms' (see Figure 1).

Participants viewed exemplars in a pseudo-randomised order to minimise error variance where one trial order was randomly generated when coding and all participants completed the trials in this order.

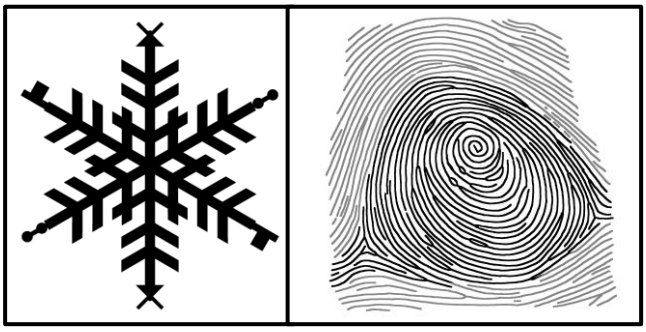

Figure 1. Artificial (left) and natural (right) stimuli used

Test Phase Participants completed two statistical learning measures in a set order: discrimination judgements; and then frequency estimates. Discrimination judgements were 21 recognition trials and 12 completion trials presented in a

[1] E.g. 'Please select 'A' out of the options available below.

pseudo-randomised order (see Figure 2; adapted from Siegelman et al., 2017).

On recognition trials, participants viewed two, three or four pairs of features and were asked to choose which pair was more familiar. On completion trials, participants viewed one 'target' feature and two or three additional features. One of these additional features was the correct answer and the others were foils. Participants were asked to "choose the feature that best completes the pair". Correct answers were based on the true joint probabilities from the exposure phase. Accuracy was measured by the total number of correct trials and group-level chance performance was calculated by aggregating the different probabilities of responses for each trial (40.15% accuracy or 13.25 trials).

Participants were instructed to ignore the orientation and location of features when making their selections. The spatial location of each feature in a pair and the spatial location of all the pairs on the screen was also pseudo-randomised per trial.

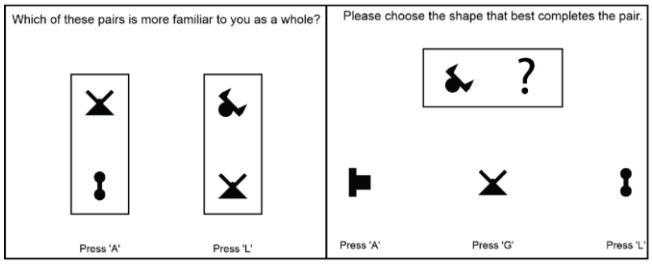

Figure 2. Recognition (left) and completion (right) discrimination judgement trials

Frequency estimates were the estimated proportion of time participants saw each of 13 features in all the images that they saw during the exposure phase. Participants were asked 'what proportion of the time did this [shape/fingerprint] occur in the images that you saw?,' and they provided their answers in a textbox restricted to a scale of 0-100%. Participants provided frequency estimates for each feature in a pseudo-randomised order. It was not ensured that the total of the estimates added up to 100% nor were the estimates normalised. Accuracy in this task (henceforth: *estimation accuracy*) was measured by calculating absolute error for each participant by subtracting the true feature frequency from the absolute estimated feature frequency for each feature, then averaging across estimates. Lower absolute error indicates better estimation accuracy.

Natural Task and Dependent Measures

Exposure Phase Stimuli Participants viewed 430 fingerprint patterns in a pseudo-randomised order manipulated to appear with their 'ground-truth' frequencies in the general population ranging from 0.002–0.305 (as in de Jongh, Lubach, Lie Kwie, & Alberink, 2019).[2]

Test Phase Participants completed discrimination judgements, then frequency estimates. Discrimination judgements were 64 recognition trials (similar to left panel of Figure 2) presented in a pseudo-randomised order. Participants viewed two, three or four fingerprints and decided which was more familiar to them. Correct answers were based on the true frequencies from the exposure phase. Accuracy was the total correct trials and group-level chance performance was 38.67% accuracy or 24.75 trials.

Frequency estimates were the estimated proportion of time participants saw each of 35 fingerprints in all the fingerprints that they saw during the exposure phase. Estimation accuracy was calculated as for the artificial frequency estimates.

Procedure

Participants completed the natural exposure phase where they viewed fingerprint images (exposure duration (ED) of 1.5-sec and interstimulus interval (ISI) of .25-sec), and then provided both natural statistical learning measures. After a short self-determined break, participants then completed the artificial exposure phase (ED = 3-sec and ISI = 1-sec)[3] and then provided both artificial statistical learning measures. Upon completion of the experiment, participants viewed a debriefing screen that thanked them for their participation and informed them about the aims of the study.

Results

Pre-Registered Results

Descriptive Statistics and Psychometric Properties Natural (M = 42.71, $t_{(109)}$ = 42.46, p < .001, 95% CI [41.03, 44.39]) and artificial (M = 21.47, $t_{(109)}$ = 16.57, p < .001, 95% CI [20.49, 22.46]) discrimination judgements were significantly better than chance performance (24.75 and 13.25 respectively; see Figure 3). Absolute error was relatively low in each condition and within the range of similar previous experiments (Growns & Martire, in press).

Artificial discrimination judgements (Cronbach's α = .80, 95% CI [.74, .85], split-half Spearman-Brown's r = .67) and estimation accuracy (α = .81, 95% CI [.76, .86]; r = .74) displayed psychometric properties close to or above recommended psychometric values (> 0.8; Streiner, 2003). Natural discrimination judgements (α = .87, 95% CI [.83, .90]; r = .77) and estimation accuracy (α = .96, 95% CI [.95, .97]; r = .94) displayed psychometric properties above recommended values.

[2] Note that to balance experiment length feasibility and ecological validity, we rounded up eight frequencies to the nearest trial (n = 1) as some 'ground-truth' frequencies were 1/1000+, and also dropped one fingerprint from the original paper as there was no image available.

[3] Note that ED and ISI differed between tasks to maintain feasibility of exposure phase length in due to the different number of stimuli in needed to maintain the distributional information in both tasks.

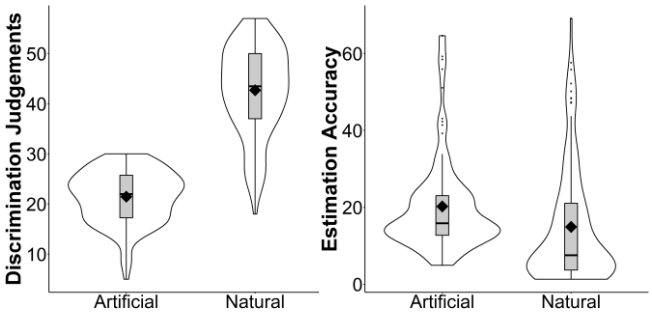

Figure 3. Total discrimination judgements correct (left) and frequency estimation accuracy (right)

Correlational Analyses Correlational analyses were conducted using the *cor.test* and *correlationBF* functions from the core *stats* and *BayesFactor* packages in R (Morey, Rouder & Jamil, 2018). Artificial discrimination accuracy significantly correlated with artificial estimation accuracy (r = -.22, p = .019, BF = 3.01) as did natural discrimination accuracy with natural estimation accuracy (r = -.47, p < .001, BF = 85176.08; see Figure 4). Note that lower absolute error indicates better frequency estimation accuracy so negative correlations between discrimination judgements and estimation accuracy indicate a positive relationship between the statistical learning measures.

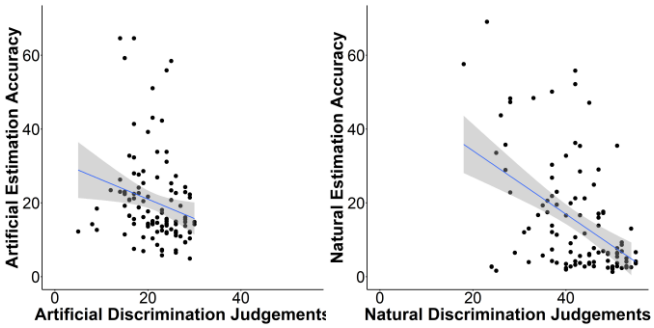

Figure 4. Correlations between artificial (left) and natural (right) discrimination judgements and estimation accuracy

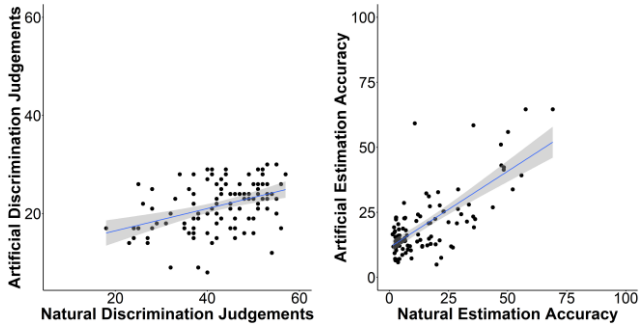

Figure 5. Correlations between discrimination judgements (left) and estimation accuracy (right) in both tasks.

Artificial and natural discrimination accuracy also significantly correlated with one another (r = .36, p < .001, BF = 233.06), as did artificial and natural estimation accuracy (r = .73, p < .001, BF = 9.68e+15; see Figure 5).

We calculated Bayes Factors to examine the likelihood of the data under the null hypothesis (i.e. the absence of correlations) compared to an alternative hypothesis (i.e. the presence of correlations; Wetzels et al., 2011). There was strong support for a negative correlation between natural discrimination and estimation accuracy (BF = 85176.08), and positive correlations between artificial and natural discrimination (BF = 233.06) and artificial and natural frequency accuracy (BF = 9.68e+15), although only moderate support for a negative correlation between artificial discrimination and estimation accuracy (BF = 3.01).

Exploratory Results

Principle Components Analysis (PCA) We further explored the shared and unshared variance across all four measures with a Principle Components Analysis using the *prcomp* function from the core *stats* package in R (see Table 1 for loadings of all four components and proportion of variance explained by each). The first factor explained the majority of variance in performance (55%) across all four measures. Three of the measures (artificial estimates and natural judgements and estimates) similarly loaded onto this factor which suggests that the factor represents the shared component of variance across those three tasks, whilst artificial judgement accuracy loaded less strongly onto the first factor.

Importantly, the second and third factors also explained a substantial amount of the observed variance (24% and 14% respectively). The task loadings suggest that these additional components reflect task-specific variance related specifically to each task. The second component is strongly related to performance in artificial discrimination judgements, and the third component is strongly related to performance in natural discrimination judgements. The fourth component explains less of the variance (6.4%) but differentiates performance in artificial and natural frequency estimates. Overall, these results suggest that performance in each measure reflects a mixture of shared and non-shared variance.

Table 1: Loadings matrix and variance explained in PCA

	Component			
	1	2	3	4
Artificial				
Judgements	0.32	-0.83	-0.44	0.14
Estimates	-0.57	-0.32	0.39	0.65
Natural				
Judgements	0.50	-0.26	0.81	-0.18
Estimates	-0.57	-0.38	0.07	-0.72
Variance Explained	55.31%	24.03%	14.26%	6.41%

Discussion

This paper provided the first empirical investigation into *how* and *how well* distributional learning can be measured. We investigated whether the ability to discriminate between relative frequencies (*discrimination judgements*) and the ability to explicitly estimate frequencies (*frequency estimates*) tapped into one unified distributional statistical learning ability, or were separate sub-processes of the same theoretical construct. We also simultaneously investigated the psychometric properties of these measures.

Participants learned distributional information in both natural and artificial tasks – discrimination judgements were significantly above chance and frequency estimation accuracy was relatively low. We also identified a stable relationship between discrimination judgements and frequency estimates in both tasks, and a large portion of the variance of all measures was accounted for by one factor (although artificial judgements loaded onto this factor less strongly). This suggests that on one level, the ability to discriminate between relative frequencies and estimate explicit frequencies may be part of a unified ability to extract distributional information from the environment.

Natural and discrimination judgement accuracy also significantly correlated with one another, as did estimation accuracy in both tasks. This generalisation across stimuli provides more evidence that we may be tapping into a broader distributional learning construct. It suggests that not only are better 'discriminators' better 'estimators,' but that better discriminators' and estimators' abilities generalise across different types of stimuli.

Importantly, our results also demonstrate that individual ability on each measure also accounts for a substantial amount of the variance. Only moderate correlations were observed between most measures, and performance on the artificial and natural discrimination judgement measures discriminated performance on all other tasks on two factors, whilst the remaining factor discriminated performance between artificial and natural frequency estimates. Our results overall suggest that distributional statistical learning is comprised of both the ability to discriminate relative frequency and estimate explicit ones, but there is also individual skill in both abilities that may be stimulus-specific. This is consistent with research suggesting that conditional statistical learning may be stimulus-specific to some degree (Conway & Christiansen, 2006; Vouloumanos, Brosseau-Liard, Balaban, & Hager, 2012).

Although distributional learning may similarly be stimulus-specific to some degree, it is also possible that our results were constrained by the reliability of their measures. Both correlational and principle components analyses are constrained by the reliability and validity of the measures used in analyses (Raykov, Marcoulides, & Li, 2017; Siegelman et al., 2017). Artificial discrimination judgements displayed the lowest reliability (α = .80, r = .67) compared to the other measures (α = .81-.96, r = .74-.94). It was also the measure to load the least strongly onto the first factor in the PCA, and the lowest correlation was observed between artificial discrimination judgements and estimates (r = -.22). It is possible that the decreased complexity of the artificial judgement task due to smaller and less complex trials resulted in its lower reliability – similar to how increased complexity increases the reliability of conditional learning measures (Siegelman et al., 2017). Whilst this research highlights the importance of increased complexity and difficulty of statistical learning measures, the lower reliability of the artificial discrimination judgements measure likely impacted the observed correlations and impacted the results of the PCA.

Theoretical Implications

Distributional and condition statistical learning have been theorised to be underpinned by separate, but inter-related, memory processes (Thiessen & Erickson, 2013; Thiessen et al., 2013). It has been suggested that conditional learning is underpinned by extraction processes where discrete units (e.g. words) are stored in memory, whilst distributional learning is underpinned by integration processes where a central tendency and variability surrounding this is stored in memory (Thiessen & Erickson, 2013; Thiessen et al., 2013). Similarly, the ability to discriminate between frequencies and explicitly estimate the same frequencies may be underpinned by separate memory processes.

Human memory is typically theorised to contain separate, but related, *recognition* and *recall* systems (Haist, Shimamura, & Squire, 1992). Importantly, discrimination judgements reflect the ability to *recognise* differences in frequencies, whilst frequency estimates reflect the ability to explicitly *recall* these abilities. Just as conditional and distributional learning are theorised to be facilitated by separate but interrelated memory processes, this may also be the case for the ability to explicitly estimate and discriminate between relative frequencies. Future research should investigate the memory processes involved with statistical learning more broadly.

Limitations and Future Directions

Although we identified significant associations between all measures and the PCA identified a common factor across them, this may not necessarily mean that we are only tapping into a statistical learning ability. It is possible that other mechanisms may underpin this relationship – such as participant motivation or attention. Some individuals (particularly university students who participate for course credit) may be more motivated or pay more attention than others across all tasks. This could affect the relationship seen between the distributional learning measures in this study, and the common factor identified in the PCA.

Nevertheless, it is important to note that stable relationships are not always identified between measures of statistical learning – such as auditory and visual statistical learning (Siegelman & Frost, 2017). If unrelated abilities (such as motivation or attention) produced relationships between statistical learning measures, you might expect all statistical learning measures to be significantly associated.

Yet this is not the case. It is therefore plausible that we are tapping into one distributional learning ability. However, other unrelated mechanisms may also play a role and future research would do well to explore how such possible mechanisms may impact statistical learning.

It is also important to note that the distributional information and task parameters were not identical across all experiments and tasks in this paper. For example, the distributional information available to learn was different in the artificial and natural tasks. In the artificial tasks, distributional information varied within and between exemplars, whereas it only varied between exemplars in the natural task. Further, the task parameters also varied in terms of exposure duration and interstimulus intervals. Although this choice was intentional,[3] an ideal experimental design would use tasks with identical parameters to provide better control. Nevertheless, it is important to note that a significant relationship between both distributional learning measures was identified despite these differences. Future research should aim to further investigate the relationship between distributional learning measures utilising tasks with similar task parameters.

Conclusion

Overall, we provided the first evidence that distributional statistical learning is comprised of both the ability to discriminate relative frequencies and explicitly estimate them, as well as individual ability in both of these. We demonstrated that increasing the complexity and reliability of distributional learning measures increases their reliability. This will improve the ability of future research to explore the theoretical underpinnings of statistical learning in a broader context and investigate its role in other cognitive processes.

References

Altvater-Mackensen, N., Jessen, S., & Grossmann, T. (2017). Brain responses reveal that infants' face discrimination is guided by statistical learning from distributional information. *Developmental Science, 20*(2), 1-8.

Arnon, I. (2019). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, 1-14.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211-233.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65*(3), 212-219.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3), 5-5.

Busey, T., Nikolov, D., Yu, C., Emerick, B., & Vanderkolk, J. (2016). Characterizing human expertise using computational metrics of feature diagnosticity in a pattern matching task. *Cognitive Science, 41*, 1717-1759.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science, 17*(10), 905-912.

de Jongh, A., Lubach, A. R., Lie Kwie, S. L., & Alberink, I. (2019). Measuring the rarity of fingerprints patterns in the Dutch population using an extended classification set. *Journal of Forensic Sciences, 64*, 108-119.

Escudero, P., & Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition, 133*(2), 408-413.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*(3), 458-467.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin Journal, 145*(12), 1128-1153.

Growns, B., & Martire, K. A. (in press). Forensic feature-comparison expertise: statistical learning facilitates visual comparison performance. *Journal of Experimental Psychology: Applied*, 1-18, https://doi.org/10.31234/osf.io/pzfjb.

Growns, B., Siegelman, N., & Martire, K. A. (under review). The multi-faceted nature of visual statistical learning: individual differences in learning conditional and distributional regularities across time and space. *Psychological Bulletin & Review*.

Haist, F., Shimamura, A. P., & Squire, L. R. (1992). On the relationship between recall and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 691.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*(12), 1372-1388.

Junge, C., van Rooijen, R., & Raijmakers, M. (2018). Distributional information shapes infants' categorization of objects. *Infancy, 23*(6), 917-926.

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making, 9*(3), 259-273.

Martire, K. A., Growns, B., & Navarro, D. J. (2018). What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review, 25*(6), 2346–2355.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*(1), 122-134.

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science:

What can be learned and what is good experimental practice? *Vision Research, 141*, 4-15.

Qualtrics. (2005). Qualtrics (Version September, 2018). Provo, Utah, USA: Qualtrics.

Raykov, T., Marcoulides, G. A., & Li, T. (2017). On the fallibility of principal components in research. *Educational and Psychological Measurement, 77*(1), 165-178.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49*(2), 418-432.

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

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*(3), 217-222.

Thiessen, E. D., & Erickson, L. C. (2013). Beyond word segmentation: A two-process account of statistical learning. *Journal of Current Directions in Psychological Science, 22*(3), 239-243.

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

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552-564.

Vouloumanos, A., Brosseau-Liard, P. E., Balaban, E., & Hager, A. D. (2012). Are the products of statistical learning abstract or stimulus-specific? *Frontiers in Psychology, 3*(70), 1-11.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Journal of Perspectives on Psychological Science, 6*(3), 291-298.

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency Processing and Cognition* (pp. 21-36). New York, NY, US: Oxford University Press.