Online supplemental materials:

1. **Elaborated methodological information**
2. **Syntax search**
3. **P-curves with disclosure tables**
4. **Interrater reliability**
5. **Tables with study characteristics for all studies**

1. Elaborated Methodological information

Statistical considerations for the meta-analyses of group design studies

We calculated effect sizes using standard deviations for the random trial condition for serial reaction time tasks and the unrepeated sequence condition for the Hebb tasks, as this is equivalent to using the pre-test standard deviations recommended by Morris (2008). We used the standard deviation for the control group only to standardize effect sizes, as this gives a change score that relates directly to the size of the improvement seen, compared to control group performance; this decision will tend to increase the effect sizes obtained slightly compared to using the random (or unrepeated) condition standard deviations for both groups, as standard deviations in the clinical groups tend to be larger than those in the control group. These Cohen's *d* estimates were then entered into CMA using inverse variance weights to calculate effect sizes.

In cases where a single effect size was calculated from condition means for each group across several blocks of a task, a pooled standard deviation for each group condition mean was calculated using the following formula (see Equation 1) in order to take account of the variance between the block means, as well as the variance within them.

$$S_{pGC} = \sqrt{n_{B1}\left(\sigma_{B1}^2 + \delta_{B1}^2\right) + n_{B2}\left(\sigma_{B2}^2 + \delta_{B2}^2\right)... + n_{Bn}^2\left(\sigma_{Bn}^2 + \delta_{Bn}^2\right) \div \left(n_{B1} + n_{B2}... + n_{Bn}\right)} \quad (1)$$

Where S_{pGC} is the pooled standard deviation for a group condition mean, δ is the difference when subtracting the grand mean from the block mean for the group condition, σ denotes the block standard deviation for the group condition and B denotes the task block (1 to *n*).

The correlation between random and sequenced conditions for serial reaction time tasks and unrepeated and repeated sequences for Hebb learning tasks were not reported in any of the papers included in the meta-analyses. The meta-analyses were, therefore, estimated including this correlation at varying levels (0.0, 0.5, 0.7, 0.9) to assess the impact this might have on results. Inclusion of any of these correlations had no appreciable effect on effect size estimates, or the between study variance estimates. Therefore, since actual correlations for each study were unknown, the final meta-analyses were based on a zero correlation between conditions.

It should be noted that this method is different to the method used in previous meta-analyses of serial reaction time tasks discussed earlier (Lum et al., 2013; 2014; Obeid et al., 2016). These previous meta-analyses base their effect size calculations on a method set out in an early meta-analysis of serial reaction time tasks in Parkinson's disease patients by Siegert, Taylor, Weatherall, and Abernethy (2006). At first glance this method looks identical to the one we have used (see Equation 1), but the pooled standard deviation that forms the denominator of the equation in Siegert et al.'s method only uses the standard deviation for each group for the difference between the conditions (see Equation 2), rather than standard deviations for raw scores.

$$d = (M_{Control\ group} - M_{LD\ group}) \div S_p \quad (1)$$

Where S_p is calculated as follows:

$$S_p = \sqrt{\left(S^2_{\textit{Control mean difference}} + S^2_{\textit{LD mean difference}}\right)} \div 2 \quad (2)$$

This method is questionable, since the denominator that represents variance in the effect size equation will be underestimated as a result of using only variance of difference scores (not the variance of component raw scores). Such a numerator will inflate the estimate of effect size obtained (Morris & Deshon, 2002), as has been previously demonstrated (Lund, 1988; Ray & Shadish, 1996).

To underline the impact of using different calculation methods to obtain effect sizes, effect sizes were calculated using both methods for the eight studies in our meta-analysis using serial reaction time tasks that were able to provide information in both formats. First, the results were calculated using the difference score methodology (used by several previous meta-analyses) with the standard deviations for the difference between conditions in the denominator (instead of the standard deviation for the random condition). When using this methodology, the effect size was moderate and significant, g = -0. 55, 95% CI [-0.90, -0.20], showing language disordered groups performing more poorly on serial reaction time tasks compared to age-matched controls. This method showed significant variation in effect sizes Q (7) = 22.18, p = .002, 1^2 = 68.44%, k = 8, Tau^2 = 0.41. However, this method over-estimated the variance due to the small denominator, leading to an under-estimated pooled effect size. The results were then re-calculated using the recommended raw score methodology, with standard deviations for the random condition. This gave a different picture of the data. The effect size for the eight studies was far lower, g = -0.31, 95% CI [-0.57, -0.05]. The heterogeneity estimate was also lower as a result Q (7) = 12.43, p = 0.09, 1^2 = 43.69%, k = 8, Tau^2 = 0.06. This comparison of methods clearly demonstrates the importance of using the optimal raw score method of calculating effect sizes in group design studies using tasks that rely on the difference between experimental conditions as their dependent variable.

Most studies using serial reaction time tasks reported results for accuracy as well as response time, analysis was confined to the latter, as this is the most widely used measure of implicit learning on serial reaction time tasks. The majority of studies using serial reaction time tasks reported insufficient data to calculate an effect size. Therefore, in cases where authors were unable to supply the necessary data but studies included a figure of sufficient quality, with accurately labelled error bars, online digital software (WebPlotDigitizer: Rohatgi, 2017) was used to extract means and standard deviations for both sequenced and random trials for both groups.

Studies using Hebb tasks analysed implicit learning in two different ways. The first method compared the gradient of the regression line for performance on Hebb trials to the gradient for random trials, while other studies chose to compare overall accuracy rates for the repeated and non-repeated sequences. In order to include as many studies investigating Hebb performance and language disorder as possible using a consistent measure, the meta-analysis compared overall accuracy rates, rather than regression-based accuracy measures, converting measures to percentage scores as necessary.

Artificial grammar and statistical learning studies all used a separate testing phase, after the learning trials had been completed, to measure implicit learning in one of two ways. The first type of measure requires participants to judge whether they recognized sequences of items that they had seen during an earlier learning phase (seen items). The second type of measure requires participants to judge whether sequences of items they had not seen before were consistent with the sequential rules followed during the learning phase (transfer items). It should be noted, therefore, that these measures are essentially measures of explicit (declarative) memory for information that may have been learned implicitly. Both types of test either used a two alternate forced choice (2AFC) structure or presented test stimuli that were either correct or incorrect

one at a time (50% of each type). Group scores for statistical learning on the tasks (both correctly-identified recognition and generalization measures, as well as any scores for violations) were entered directly into CMA taking account of the direction of the effect. The mean of these estimates formed the effect size for the comparison. Where only one overall score per group was reported this formed the effect size for the comparison.

For weather prediction tasks, the proportions of correct responses per group were entered directly into CMA per task total or per block. In studies that reported proportions per block, the mean of all block estimates formed the effect size for the comparison.

Several studies in the artificial grammar and statistical learning and weather prediction meta-analyses did not report scores by group, but reported *t*-test values or *F* ratios that enabled an effect size to be calculated using the effect size calculator on the Campbell Collaboration website (http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php). These studies are identified in the tables accompanying each meta-analysis.

2. Syntax search

Table S1. Search terms for the literature search for the meta-analyses

Database	Search Terms
PsychINFO, Medline via Ovid	(OR between all the terms)
	Implicit learning (entered as a subject heading)
	Implicit adj2 learn$*
	Implicit adj2 memory
	Procedur$ adj2 learn$
	Procedur$ adj2 memory
	Probabili$ adj2 learn$
	Probabili$ adj2 memory
	Statistic$ adj2 learn$
	Statistic$ adj2 memory
	Sequence adj2 learn$
	Serial adj2 learn$
	Serial reaction time
	Hebb$ adj2 learn$
	contextual cueing
	Artificial grammar
	finite state grammar
	Weather prediction task
	AND
	(OR between all the terms)
	language disorders (subheading)

	language development disorders (subheading)
	specific language impairment (subheading)
	dyslexia (subheading)
	Language adj2 impair$
	Language adj2 problem$
	Language adj2 disorder$
	Language adj2 deficit$
	Language adj2 difficult$
	Language adj2 abilit$
	Language adj2 fluen$
	Read$ adj2 abilit$
	Read$ adj2 fluen$
	Read$ adj2 impair$
	Read$ adj2 difficult$
	Verbal adj2 impair$
	Verbal adj2 deficit$
	Verbal adj2 abilit$
	Phonolog$ adj2 impair$
	Phonolog$ adj2 deficit$
	Gramma$ adj2 impair$
	Gramma$ adj2 deficit$

3. Examining publishing bias and p-hacking in the literature using the p-curve

P-curves (Simonsohn, Nelson, & Simmons, 2014a; 2014b) were used to investigate the extent of publication bias for the principal measure of implicit learning across all published studies eligible for each of the meta-analyses of group design studies. The p-curve examines the distribution of significant results, with the shape of the curve determining the evidential value of the studies it contains. It does this by calculating the probability of observing a p-value as extreme if the null were true for each significant p-value. It then aggregates these to give a chi square test for skew, such that only right-skewed curves with more low than high p values show evidential value.

There is an important difference between the p-curve and the funnel plot analysis of publication bias in the meta-analysis of artificial grammar and statistical learning studies, however. The p-curve analyses as used here investigate whether there is bias across the literature as a whole, while the funnel plots evaluate whether the effect size in the meta-analysis itself is likely to be inflated as a result of publishing bias. The results from funnel plot and p-curve for the artificial grammar and statistical learning meta-analysis were contradictory. However, although the p-curve is recommended as a more reliable method of determining publication bias than Duvall and Tweedie's (2000) Trim and Fill procedure (Simonsohn et al., 2014b), there is evidence that the p-curve has a high false positive rate for evidential value when heterogeneity within the sample is large (Carter, Schönbrodt, Gervais, & Hilgard, 2019).

The procedural deficit hypothesis at the centre of these meta-analyses, claims that language-disordered groups will display poorer implicit learning on the implicit memory tasks than control groups with normal language. Therefore, a single statistic was coded that related to the principal measure of implicit learning in each of the group design meta-analyses (the two meta-

analyses of correlational designs were not examined for publication bias, owing to the small number of studies they included).

Studies using the serial reaction time task

The principal measure of implicit learning for the serial reaction time task for each of the 52 studies that were eligible for the group design meta-analysis was the statistic that referred to the group difference in RTs between sequenced and random trials (see Table 3.1). For deterministic tasks, the statistic typically related to the difference between the last sequenced block and a subsequent block of random trials. For alternating or probabilistic tasks, this measure was sometimes taken across the whole of the task. Where studies contained two comparisons, a statistic was coded for each one and p-curves were run twice, each time including only the first or the second comparisons from the study, as recommended by Simonsohn et al. (2014a). The results for the two p-curves were equivalent, so only the first one is reported here.

Of the 52 studies, 26 reported significant results for a difference between groups on the principal measures of implicit learning and 26 studies reported null results, underlining the inconsistency of results in the field. Several of these null results came from studies claiming support for the procedural deficit hypothesis, in the light of significant secondary findings, so the full extent of nonsignificant findings on the principal implicit learning measure for the serial reaction time task were not immediately apparent from the literature. For example, Bennett et al. (2008) reported a null result, but claimed support for the procedural deficit hypothesis in light of a positive correlation between implicit learning scores and reading ability. Desmottes et al. (2016a; 2017) reported initial null results, but impaired consolidation of procedural learning in children with developmental language disorder, with poorer performance during a second attempt at the task. Similar results were also reported on an alternating serial reaction

time task in dyslexic children (Hedenius et al., 2013)[1]. Implicit learning impairments in language disorder have also been linked to task-specific differences. Gabriel et al. (2014) reported equivalent learning for groups with regards to response times, but suggested that children with developmental language disorder might be more error prone than typically developing children during an auditory, but not a motor, version of the serial reaction time task. Only seven studies (Bussy et al., 2011; Gabriel et al., 2011; Kelly et al., 2002; Laasonen et al., 2014; Lum & Bleses, 2012; Rüsseler et al., 2006; Vakil et al., 2015) stood firmly behind their null result on the serial reaction time task.

Four of the 26 studies with significant results reported statistics in a format that could not be included in the p-curve, failing to report the F-ratio and including only the p value (Menghini et al., 2008; Stoodley et al., 2006; 2008) or omitting the degrees of freedom (Clark & Lum, 2017b). One study reported no between group difference during a first training session, but a significant difference over subsequent sessions in two separate experiments (Desmottes et al., 2017). Only one experiment was included in each analysis, with no significant difference to results. In addition, three studies reported results that approximated the test of interest (significant group differences in the difference in RTs between random and sequenced trials), but with minor variations. The first of these reported significant results for differences in the growth curve of the sequenced phase of the task, without reference to the random phases (Tomblin et al., 2007). Two others reported the group x block difference across all blocks in the task, sequenced and random (Vicari et al., 2003; 2005). As recommended by Simonsohn et al. (2014a) the p-curve analysis was run with and without these three studies, but found equivalent results both times. Therefore, p-curve results are reported for all 22 studies with

[1] A finding of impaired consolidation of implicit learning should be put in context at this point as contradictory results have also been reported. Gabay et al. (2012a) found the opposite, with dyslexic adults performing comparably with controls during later learning stages, while showing impaired learning during initial acquisition.

significant results. Figure 3.1 (top left) shows a right-skewed p-curve which demonstrates evidential value for the 52 studies eligible for the serial reaction time task extreme groups meta-analysis (Z = -3.96, p < .0001). There is also no reliable evidence that the studies' evidential value is inadequate due to low power (power estimate = 49%, 90% CI: [23%, 71%].

Studies using Hebb serial order learning tasks

The p-curve for this set of studies coded the principal measure of learning on the Hebb task, according to each eligible study (see Table 3.2). This included several regression-based measures that indicate improving recall for the Hebb sequence over time, as well as measures that related to an overall group difference in performance across the task. This enabled the inclusion of studies that only reported regression-based inferential statistics. However, the inclusion of both types of measure should be kept in mind when interpreting the result of the p-curve. The low number of studies is also a concern. Eight studies contained sufficient data for the analysis, with three studies providing a significant statistic that represented a different gradient of improvement in implicit learning over the course of the task for the two groups and two studies indicating an overall difference in improvement. Figure 3.1 (bottom left) shows a right-skewed p-curve, demonstrating evidential value (Z = - 4.47, p = .0001) and no reliable evidence that the studies' evidential value is inadequate due to low power (power estimate = 93%, 90% CI [68%, 99%].

Studies using artificial grammar learning and statistical learning tasks

A p-curve analysis was also undertaken to investigate whether the complete body of eligible group design studies using artificial grammar and statistical learning tasks was subject to publication bias. The p-curve focused only on overall group differences, since this is the effect size of interest in the meta-analysis. All 31 studies eligible for the meta-analysis were examined

and a single p-value for each study was coded that related specifically to this group difference (see Table 3.3). For the majority of studies this was an Anova main effect of group.

Three studies were categorized as nonsignificant for the purposes of the p-curve. These reported a nonsignificant main effect of group, but highlighted significant secondary group interactions: Aguilar and Plante (2014) reported differences in scores for correct and incorrect items; Kahta and Schiff (2016) reported similar differences; Nigro et al. (2016) reported differences in scores for transfer to unseen items. One study with significantly different group means was excluded because p-values related only to multiple regression analyses (Mainela-Arnold & Evans, 2014). Another study reported a significant effect, but in the opposite direction, with the dyslexic group performing better than controls (Pothos & Kirk, 2004). This study was, therefore, categorized as a null result for the purposes of the p-curve analysis. Finally, two studies contained significant results on more than one task. Pavlidou and Williams (2010) reported a significant main effect for each of two tasks taken by the same participants. Evans et al. (2009) gave a second task to a subset of the same participants. As recommended by Simonsohn et al. (2014a), a p-curve was run for the values from the first tasks and a second analysis was run that included the values for the second tasks. The results for the two p-curves were equivalent, so only the first one is reported here.

There were 15 significant values for the 31 studies eligible for the meta-analysis that could be entered into the p-curve. Figure 3.1 shows a right-skewed p-curve, demonstrating evidential value (Z = 4.13, p = .0001) and no reliable evidence that the studies' evidential value is inadequate due to low power (power estimate = 62%, 90% CI [32%, 83%]).

Studies using the weather prediction task

A P-curve was also estimated for the six eligible group design studies using the weather prediction task. The principal measure of implicit learning in these studies related to the overall

difference in learning between groups and were typically the main effect of group in a Group x Block Anova (see Table 3.4). No studies reported significant results for the difference in the rate of learning between groups over the task. There were four significant values for this statistic. Figure 3.1 (bottom right) shows a right-skewed p-curve, demonstrating evidential value (Z = - 3.48, p = .0003) and no reliable evidence that the studies' evidential value is inadequate due to low power (power estimate = 87%, 90% CI [46%, 98%].

< Insert Figure S1 here >

Table S2. Disclosure table for 52 group design studies eligible for the meta-analysis using the SRT task.

Study name[1]	**Analysis**	**Quoted test from paper with statistical results**	**Significance***
Bennett, Romano, Howard Jr, & Howard, 2008	Mixed ANOVA: 2 (Group) x 2 (triplet type) x 6 (Epoch). Measure is group difference in RTs between high and low frequency triplets	"Group x triplet type and Group x triplet type x epoch interactions were not significant, *P*'s > .10, indicating that we did not detect group differences in sequence learning." (p. 190)	Null
Bussy et al., 2011	Mixed ANOVA: 2 (Group) x 2 (Sequence) x 6 (Blocks)	"Premierement, L'effet principal du facteur groupe n'est pas significatif ($F(2,40) = 1.43$; $p > 0.10$) [...].La difference de temps de reaction entre le dernier bloc sequential et le dernier bloc aleatoire (le cinqieme bloc) est egalement significative pour CG ($F(2,40) = 32.55$, $p < .001$), pour DP ($F(2,40) = 14.26$, $p < .001$), et pour DS ($F(2,40)= 20.39$, $p < .001$)." (p. 144)	Null
Clark & Lum, 2017a (1)	Mixed ANOVA (FOC): 2 (Group) x 2 (Sequence type: Block 4 (random) vs mean of Blocks 3 & 5 (sequence)).	"However, a significant Group x Block interaction with a medium to large effect size was observed, $F(1,50) = 4.785$, $p = .033$, $\pi^2p = .087$." (p. 154)	Significant

Clark & Lum, 2017a (2)	Mixed ANOVA (SOC): 2 (Group) x 2 (Sequence type: Block 4 (random) vs mean of Blocks 3 & 5 (sequence)).	"Neither the main effect of group [...], nor the interaction between block and group was significant, $F(1,50) = .725$, $p = .399$, $\pi^2 p = .014$." (p. 154)	Null
Clark & Lum, 2017b	*T*-test: Group difference in procedural learning	Table 2 shows that the DLD group performed more poorly than the TD group on most tasks, though this difference only reached statistical significance for the SRTT and reading tasks." Table 2 comparison: $t = -2.48$, $p = .018$, $d = -0.79$	Significant
Conti-Ramsden, Ullman & Lum, 2015	*T*-test: Group difference of Difference Z score between block 4 and 5	"Children with DLD had significantly lower scores on all predictor variables." (p. 6). $t(89) = 3.00$, $p = .003$ (Table 2, p.7)	Significant
Deroost, Zeischka, Coomans, Bouazza, Depessemier, & Soetens, 2010	Mixed ANOVA: 2 (Group) x 2 (task) x 2 (sequence type). RT difference between random (B14) and mean of sequence blocks (B13 & B15). NB: Result includes both tasks (FOC & SOC)	"Critically, no interaction of Group x Sequence learning, nor an interaction of Group x Sequence x Sequence Learning could be observed, both $F < 1$." (p. 566)	Null

Desmottes, Meulemans, & Maillart, 2016a	Mixed ANOVA: 2 (Group) x 4 (Difference score on Epoch 1, 5, 6, 7). Epoch 1 and 5 are start and end of Day 1, Epoch 6 is 24 hrs later and Epoch 7 is 1 week later.	"This analysis showed a marginal effect of Group (F(1,40) = 3.46, p = .066, π^2p = .08), indicating a (slightly) better sequence knowledge in children with TD (M = 0.14, SD = 0.11) than in children with DLD (M = 0.09, SD = 0.09)." (p. 60)	Null
Desmottes, Meulemans, & Maillart, 2016b	Mixed ANOVA: 2 (Group) x 2 (task) x 3 (Block 4-6). RT difference between B5 (random) and mean of B4 & B6 (sequence).	"Interestingly, the interaction between block and group showed that these differences in RT's differed between groups (F(2,92) = 3.22, p = .044) […] Indeed the difference between the random and both surrounding sequence blocks was significant in TD children (F(1,46) = 23.197, p < .001), but not for children with DLD (F(1,46) = 2.525, p = .140)". (p 525)	Significant
Desmottes, Maillart, & Meulemans, 2017 - Experiment 1	Mixed ANOVA: 2 (Group) x 3 (Epoch 1 - 3 difference scores).	"Finally, there was no interaction between group and epoch, F(2,66) = .237, p = .789, π^2p = .007, indicating that a similar improvement in sequence knowledge with practice could be observed in both DLD and TD groups" (p. 8)	Null

Desmottes, Maillart, & Meulemans, 2017 - Experiment 2	Mixed ANOVA: 2 (Group) x 3 (Epoch 3 - 5 difference scores).	[The Anova] "…showed no main effect of group...or epoch…Nevertheless, the interaction between the two variables was statistically significant, $F(2,64) = 5.85$, $p = .004$, $\pi^2p = .155$. This indicated that the evolution of the sequence knowledge differed between the groups over the post-training sessions." (p. 12)	Significant over several sessions
Desmottes, Meulemans, Patinec, & Maillart, 2017	Mixed ANOVA: 2 (Group) x 2 (Epoch 6 & 7) x 2 (Condition: distributed or massed practice).	This analysis found a significant main effect of Group, $F(1,56) = 4.671$, $p = .034$, $\pi^2p = .076$, with better level of sequence knowledge during the retention phase for TD children… (p. 2641)	Significant
Du & Kelly, 2013	Mixed ANOVA: 2 (Group) x 2 (Block 9 (random) vs mean of Blocks 8 & 10 (sequence))	"…no significant effect of group […] and no significant interaction of group x block, $F(1,22) < 1$. These results indicate that both dyslexic and control groups demonstrated significant and comparable learning." (p. 162)	Null
Gabay, Schiff, & Vakil, 2012a (1)	Mixed ANOVA: 2 (Group) x 2 (Block 4 (Sequence) to 5 (Random)). Transfer measure of difference between Block 4 & 5.	"The interaction between these variables did not reach significance, $F(1,22) = 1.648$, MSE = 682, $p > .05$." (p. 284)	Null

Gabay, Schiff, & Vakil, 2012a (2)	Mixed ANOVA: 2 (Group) x 2 (Block 6 (Sequence) to 5 (Random)). Recovery measure of difference between Block 5 & 6.	"The interaction between those variables was also significant, $F(1,22) = 7.458$, MSE = 680, $p < .05$, $\pi^2p = .25$. This pattern indicates that the DD group needs a longer time in order to recover from learning of a different sequence than does the control group." (p. 284)	Significant
Gabay, Schiff, & Vakil, 2012b	1st ratio is Mixed ANOVA: 2 (Group) x 2 (sequence transfer - Block 3 to 4) x 2 (task: motor vs letters); 2nd ratio is the same but task specific. The 2nd ratio (Letters SST) is entered into p-curve.	"The group by transfer interaction was marginally significant, $F(1,26) = 3.53$, $p = .07$ [….] In order to analyse this interaction, separate 2 (transfer) x 2 (group) Anovas were computed for each sequence type. For the motor sequence, the group by transfer interaction was far from significance $F<1$, suggesting that both groups learned the specific motor sequence [...] For the letter names sequence, the group by transfer interaction was significant, $F(1,26) = 7.89$, $p < .01$." (p. 2438)	Significant for letter names sequence only
Gabriel, Maillart, Stefaniak, Lejeune, Demottes, & Meulemans, 2013	Mixed ANOVA: 2 (Group) x 2 (Block 6S vs 7R). Difference in RTs between last sequenced and random block.	"However, the Group by Block interaction was not significant $F(1,40) = 2.87$, MSE = 1642, $p = .09$, $\pi^2p = .06$, [...] suggesting that the magnitude of the RT difference between blocks 6 and 7 does not differ significantly between groups." (p. 268)	Null

Gabriel, Maillart, Guillaume, Stefaniak & Meulemans, 2011	Mixed ANOVA: 2 (Group) x 2 (Block 12 (probable sequence) vs. Block 13 (improbable sequence)).	"…This analysis showed that RTs were similar in both groups […]and that block 12 was processed faster than Block 13 […] for both groups (nonsignificant interaction, F(1,28) = 2.61, MSE = 5254, p - .11, π^2p = .085). Thus learning appears to be similar in both groups." (p. 340)	Null
Gabriel, Meulemans, Parisse, & Maillart, 2015 (1)	Auditory modality Mixed ANOVA: 2 (Group) x 2 (Block) Difference in RTs between B6 sequenced and B7 random blocks.	"We first performed and ANOVA in the auditory modality […] The results showed no group effect…, a block effect..., and no interaction effect, $F(1,26) = 1.05$, $p = .31$, $\pi^2p = .039$." (p. 14)	Null
Gabriel, Meulemans, Parisse, & Maillart, 2015 (2)	Visual modality Mixed ANOVA: 2 (Group) x 2 (Block) Difference in RTs between B6 sequenced and B7 random blocks.	"We then performed the same analysis in the viusal modality and found comparable results: no group effect… a Block effect… and no interaction effect, $F(1,26) = 0.46$, $p = .503$, $\pi^2p = .017$)..." (p. 14)	Null
Gabriel, Stefaniak, Maillart, Schmitz, & Meulemans, 2012 (1)	Mixed ANOVA: 2 (Group) x 2 (Block) Difference in RTs between B6 sequenced and B7 random blocks.	"However, the interaction was not significant, π^2p (1,28) = .0005, MSE = 12172, $p = .98$, $\pi^2p < .001$, suggesting that both groups demonstrated a significant increase in their RTs from Block 6 to Block 7." (p. 334)	Null

Gabriel, Stefaniak, Maillart, Schmitz, & Meulemans, 2012 (2)	Mixed ANOVA: 2 (Group) x 2 (Block) Difference in RTs between B6 sequenced and B7 random blocks.	"...the Block x Group interaction was nonsignificant, $F(1,28) = 2.59$. MSE = 12172, $p = .11$, $\pi^2 p < .08$." (p. 335)	Null
He & Tong, 2017	Paired-sampled *t* tests of (mean of random Blocks 1 & 10 and sequenced Block 9) for each group separately	"…children with dyslexia exhibited a significant learning effect, $t(26) = 5.236$, $p < .001$. Similar significant learning effects were also observed in the age-matched controls, $t(27) = 8.625$, $p < .001$, and the reading level-matched controls, $t(27) = 9.025$, $p < .001$." (p. 1087)	Null
Hedenius, Persson, Tremblay, Adi-Japha, Verissimo, Dye, Alm, Jennische, Tomblin, and Ullman, 2011	ANCOVA (controlling for NVIQ): 2 (Group) x 5 (Epoch difference score). Group difference between high and low frequency triplets by epoch.	"…, though this was qualified by a significant Group x Epoch interaction, also with a medium to large effect size ($F(1,45) = 6.56$, $p = .014$, $\pi^2 p = .127$)." (p. 10)	Significant
Hedenius, Persson, Alm, Ullman, Howard,	Mixed ANOVA: 2 (Group) x 2 (trial-type interaction) x 3 (learning stage). Group	"Of particular interest here, the two groups did not differ with respect to sequence learning effects on RT (group x trial type interaction: $F(1, 27)$	Null

Howard, & Jennische, 2013	difference between high and low frequency triplets by epoch.	< 1; group x trial type x learning stage interaction: $F(2,54) = 1.51$, $p = .230$, $\pi^2 p = .053$." (p. 3928)	
Henderson & Warmington, 2017	Mixed ANOVA: 2 (Group) x 2 (sequence type) x 5 (Block). RT difference between sequenced and random trials across task.	"There were no significant interactions: […] Condition x Block x Group $F < 1$." (p. 204) (NB: This is for Day 1 only, but results are also null for consolidation sessions too).	Null
Howard, Howard, Japikse, & Eden, 2006	Mixed ANOVA: 2 (Group) x 2 (Sequence)	"Although both groups show sequence learning, the dyslexics show significantly less learning than controls on both measures. This is supported by significant Trial Type x Group interactions for [...] speed $F(1,21) = 4.61$, MSE = 226.58." (p. 1135)	Significant
Hsu & Bishop, 2014	Group differences compared using growth curve analysis (as in Tomblin et al, 2007)	"…we examined changes in the RTs when the task proceeded from the pattern phase to the subsequent random phase […] There was a significant effect of group ($F(2,41.76) = 9.51$, $p < .0001$), with a greater reboundin RTs in the age-matched group than the other two groups" (p. 359)	Significant

Jiménez-Fernández, Vaquero, Jiménez, & Defior, 2011	Mixed ANOVA: 2 (Group) x 2 (Sequence type)	The Group x Type of Block interaction also reached significance ($F(1,26) = 13.49$, $p = .002$). (p 96)	Significant
Kelly, Griffiths, & Frith, 2002	Mixed ANOVA: 2 (Group) x 2 (Sequence type)	"$F<1$. The lack of significance for these interactions suggests that the amount of learning shown by the two groups is not significantly different from each other…" (p. 49)	Null
Laasonen, Vare, Oksanen-Hennah, Leppamaki, Tani, Harno, Hokkanen, Pothos, & Cleeremans, 2014	ANCOVA (controlling for IQ): 3 (Group: control, dyslexia, ADHD) x 2 (sequence type). Difference in RTs between last random block 12 and mean of sequence blocks 11 & 13.	"The group x block type interaction did not reach significance, $F(2,82) = .308$, $p = .736$, $\pi^2p = .007$, observed power = 0.097." (p. 18)	Null
Lee & Tomblin, 2015	Mixed ANOVA: 2 (Group) x 2 (sequence type). RT difference between interleaved Random and Sequence blocks.	"However, the interaction effect was not significant, $F(1,46 = .39$, $p = .54$, $\pi^2p = .01$." (p. 224)	Null

Lee, Mueller, & Tomblin, 2016	*T*-test: Group difference for learning score. RT difference between Random and Sequence blocks.	Independent samples t-test showed that the learning effect was not significantly different between the two groups in our study, $t(39) = .13$, $p = .90$. (p. 1105)	Null
Lukacs & Kemeny, 2014	Univariate ANOVA (Group) on transformed difference scores to take account of participant variability. Difference between sequenced block 11 and random block 12.	"Next, the difference between the mean of z-transformed Block 11 (the last sequence block) RTs were extracted from the mean of the z-transformed Block 12 (random block) RTs. This difference reflecting the size of sequence learning was compared by group, revealing a significant group main effect, $F(1,113) = 5.888$, $p < .05$, $\pi^2p = .050$, with bigger learning effect in the control than in the SLI group." (p. 478)	Significant
Lum & Bleses, 2012	Analysis of normalized RT difference between sequence (B4) and random block (B5) was conducted separately for each group and difference in effect sizes compared for significance.	"The first analysis revealed that the TD group had significantly slower RTs in Block 5 compared to Block 4 ($F(1,19) = 42.194$, $p < .001$, $\pi^2p = .690$). The second analysis indicated that the SLI group also had significantly slower RTs in Block 5 compared to Block 4 ($F(1,12) = 6.354$, $p = .027$, $\pi^2p = .389$). While both groups were found to have slower RTs in Block 5, it is interesting to note that th eff3ect size for the RD group is larger in comparison to the SLI group. However, the	Null

		difference in effect sizes was not found to be statistically significant (z = 1.15, p = .25).” (p 54)	
Lum, Conti-Ramsden, Page, & Ullman, 2012	One way ANOVA on normalised RT difference between sequence (B4) and random block (B5).	“One-way repeated-measures ANOVA revealed a significant effect of group [$F(1,102) = 5.17$, $p = .026$, $\pi^2 p = .58$], with an approximately medium effect size, indicating a larger RT difference between blocks 4 and 5 for the TD children than the children with SLI.” (p. 1148)	Significant
Lum, Gelgic, & Conti-Ramsden, 2010	*T*-test: Group difference for normalized RT difference between sequence (B4) and random block (B5), controlling for motor speed.	“Analysis of these standardised residuals indicated the magnitude of difference between the fourth and fifth Blocks was significantly larger for the TD than the SLI group ($t(27) = 2.545$, $p = .017$, $r^2 = .193$).” (p. 104)	Significant
Mayor-Dubois, Zesiger, Van der Linden, & Roulet-Perez, 2014	Mixed ANOVA: 2 (Groups) x 2 (Sequence type) x 5 (Block)	“The groups (SLI versus C) differed in their performance in the Blocks, Groups x blocks,... but not in the sequence, Groups x sequence, $F(1,80) = .614$, ns. No triple interaction, Blocks x sequences x Group, $F(4,77) = .369$, ns), indicating an absence of statistical differences in motor learning between both groups.” (p. 18)	Null

Menghini, Finzi, Benassi, Bolzani, Facoetti, Giovagnoli, Ruffino, & Vicari, 2010	MANCOVA (with Age as covariate): Group as between subjects factor and cognitive task measures as DVs, including *z* score difference in RTs between last sequenced (B6) and random block (B7).	"Finally, in the GLM procedure, no significant difference was found in the SRTT between children with DD and NR childre, considering the difference between RTs of the last pseudo-random block (R2) and the last sequenced block (S4) as an index of viusal-motor sequence learning (in DD mean z-score +/- *SD*: SRTT: -.17 +/- 1.09)." (p. 867)	Null
Menghini, Hagberg, Caltagirone, Petrosini, & Vicari, 2006	Mixed ANOVA: 2 (Group) x 2 (Sequence type) x 2 (Block). Difference in RTs between B6 sequenced and B7 random blocks.	"The block effect […] and the group by block interaction ($F(1,26) = 6.5$, $p < .05$) were significant, while the group effect […] did not reach significance." (p. 4)	Significant
Menghini, Hagberg, Petrosini, Bozzali, Macaluso, Caltagirone, & Vicari, 2008	One way ANOVA comparing group difference in RTs between last sequenced (Block 6) and random block (Block 7).	…the group of 10 subjects with DD selected for the current study were impaired in IL, showing no SRTT changes between S5 and R2 (DD means; one-way ANOVA: $p > .1$). In contrast, the subgroup of NRs showed an IL effect (NR means; one way Anova: $p > .05$. (p. 216) (NB: No F-ratio given)	Significant

Perlant & Largy, 2011	Experiment 2 only Mixed ANOVA: 2 (Group) x 2 (Sequence type) x 5 (block) x 2 (item: linguistic and non-linguistic) Difference between interleaved sequenced and random trials over blocks. Separate analyses for each group also conducted.	"In typical readers […] analysis also shows the significance of condition x block interaction, principal indicator of sequence learning, $F(4,76) = 4.03$, $p < .001$ [...] In children with dyslexia […] The analysis also reveals the presence of significant condition x block interaction, principal indicator of sequence learning ($F(4,96) = 4.49$, $p < .01$)." (p. 309) (NB: No Group interactions for main ANOVA were reported, indicating a null result. Both groups separately show a significant learning effect. However, the three way interaction result in each of these is different and this is then claimed as a difference between groups.)	Null
Przekoracka-Krawczyk, Brenk-Krakowska, Nawrot, Rusiak, & Nasrecki, 2017	ANOVA: 2 Group on the difference score (EF_{IML}) between RTs on random Block 12 x mean of Block 11 & 13.	2…the mean EF_{IML} was significantly lower in the DG than in the CG, and was confirmed by the significant effect of the group ($F_{1,55} = 6.78$, p = 0.012, $\chi^2 = 0.11$)." (p. 6476)	Significant
Rüsseler, Gerth, & Munte, 2006	Mixed ANOVA: 2 (Group) x 2 (sequence). Difference in RTs between Block 10 (random) and mean of Blocks 9 and 11 (sequence).	"A post-hoc *F* test indicates that the amount of learning did not differ reliably between the two groups (GROUP by BLOCK: $F(1,22) = 2.8$, $p < .1085$)." (p. 817)	Null

Sengottuvel & Rao, 2013	ANOVA (structure unclear): Group x ISL sequence learning score (mean of final 30 trials of random - mean of final 30 trials of sequence)	"Children with SLI performed significantly poorer compared to TD children on sequence learning skill (see Table 3)." $F(1,40) = 29.61$, $p < .001$ (p. 3323)	Significant
Sengottuvel & Rao, 2014	ANOVA (structure unclear): Group x Difference between sequenced and random RTs.	"Even though, the SLavg1 of SLI was not significantly lower than TD, ISL value of the SLI group (ie: RLavg - SLavg1) was significantly lower than that of the TD group, thereby suggesting obvious slow RTs for the SLI group even in initial learning trials (see Table 2)." $F(1,54) = 10.72$, $p < .001$ (p. 58)	Significant
Sengottuvel, Rao, & Bishop, 2016	ANCOVA (controlling for NVIQ and age): Group x ISL sequence learning score: (Mean untransformed difference btw random and sequence blocks).	"This showed that children with SLI were significantly poorer than TD children, F(1,52) = 5.76, p = .02." (p. 10)	Significant
Staels & van Den Broeck, 2017	Latent growth curve modelling used to compare the group difference in the increase from sequenced (B9) to random block (B10).	"For both groups the increase in RTs in Block 10 looks similar (beta of the group effect on the corresponding growth factor with the dyslexic group coded as one was 12.41, p = .316, 95% CI [-39.2, 63.5]) thus, the	Null

		amount of implicit learning does not seem to differ between groups…" (p. 376)	
Stoodley, Harrison, & Stein, 2006	Mixed ANOVA: 2 (Group) x 2 (Sequence) Difference between RTs on random and sequence blocks.	"A repeated measures ANOVA showed a significant group by condition interaction during the random and repeated sequence blocks (p = .03)." (p. 796) (NB: No F-ratio given).	Significant
Stoodley, Ray, Jack, & Stein, 2008	Mann-Whitney test comparing dyslexic and control group. Percent decrease in RTs during the sequence condition compared to 1st random condition	"In the repeated measures analysis, there was a significant effect of block type […] and a significant block by group interaction (p = .001)." (p. 178) (NB: No F-ratio given)	Significant
Tomblin, Mainela-Arnold & Zhang, 2007	Growth curve analyses: Group difference on the 2 types of sequence (group differences in intercept for Pattern and for Random trials conducted separately). The growth curve analysis measure highlighted as the measure of interest in the paper is for pattern trials, so this is the F ratio we selected.	Pattern Phases: […] This model showed that the SLI group was significantly slower than the NL group at the third trial block which represents the intercept [group difference in intercept = -39.94 (SD = 14.49), F(1,602) = 7.59, p = .018]. (p. 281)	Significant

Vakil, Lowe, & Goldfus, 2015	Mixed ANOVA: (Block 7). 2 (Group) x 2 (Sequence). Difference between last sequenced (B6) and random block (B7)	In this case as well, an interaction effect was not found between the group and the influence of training, $F(1,50) = .432$, $p > .05$, as no significant difference was identified between individuals with or without DD in the increase in RT to the random sequence. (p. 475)	Null
Vicari, Finzi, Menghini, Marotta, Baldi, & Petrosini, 2005	Mixed ANOVA: 2 (Group) x 6 (block). NB: This interaction F ratio does not specifically reference implicit learning, so much as group differences over the whole task.	"…the group x block interaction ($F(5,150) = 2.8$, $p = .02$) were significant, demonstrating a different patterns of RT changes in the two groups across blocks. Critically, for the aims of this study, the two groups RTs differed significantly (Tukey's test) passing from the fifth to the sixth block [...] controls ($p = .0002$) [...] dyslexic children ($p = 1$)." (p. 1394)	Significant
Vicari, Marotta, Menghini, Molinari, & Petrosini, 2003	Mixed ANOVA: 2 (Group) x 6 (block), so the interaction F ratio does not specifically reference implicit learning, so much as group differences over the whole task. Control group differed significantly on difference between 5th	The group x block interaction was also significant $F(5,170) = 5.95$, $p < .0001$, thus demosntrating a different pattern of RT changes in the two groups across blocks…Critically, for the aims of the study, the RTs of the two groups strongly differed passing from the fifth to the sixth block. (p. 110)	Significant

	and 6th block ($p < .001$), but the dyslexics did not (ns).		
Yang & Hong-Yan, 2011	Left and right hand Mixed ANOVAs separately: 2 (group) x 5 (block), so the interaction F ratio does not specifically reference implicit learning, so much as group differences over the whole task. Left hand: control group differed significantly on difference between 3rd and 4th block ($p < .05$), but the dyslexics did not (ns). Right hand: both groups showed significant differences ($p < .05$)	Left hand: "The interaction between block and group was not significant, $F(4,49) = 1.16$, $p = .34$." (p. 4). Right Hand: "The interaction between block and group was not significant, $F(4,49) = .21$, $p = .93$." (p. 5)	Null
Yang, Hong-Yan, Zhi-Ying, & Shao, 2013	Mixed ANOVA: 2 (group) x 5 (block), so the interaction F ratio does not specifically reference implicit learning, so much as group differences over the whole task. The group	"… the interaction of group and block were not significant, […] $F(1,14) = 1.222$, $p = 0.345$, ES = 0.259 [...]The mean learning rate of RT of dyslexic group ([Block 5 - Block 4] / [Block 4 + Block 5]...) was 0.06	Null

	difference between sequenced block 4 and random block 5 is quantified with a *t* test statistic, however, and this is used in the p-curve.	and control group was 0,095. But, the difference of learning rate did not reach statistic significance [$t(18) = -1.188$, $p = 0.25$]" (p. 303)	
Zwart, Vissers, Kessels, & Maes, 2018	Probabilistic learning: Mixed ANOVA: 3 (group: ASD, SLI, TD) x Trial type (Standard, Deviant) x 3 (Block). Deterministic learning: Mixed ANOVA: 3 (group: ASD, SLI, TD) x 3 (Block).	Probabilistic learning: "No main Group effect was found, $P = 0.084$, suggesting similar response speed across groups. No other significant interaction effects were found, suggesting no group difference in probabilistic learning." Deterministic learning: "No Group x Block interaction was found, $P = 0.26$, suggesting similar learning effects across groups." (p. 1055)	Null

[1] = Number in brackets after study name refers to whether the measure is the first or second mention of a principal measure of implicit learning. Only results of 1st p-curve is reported, since results were equivalent; *Significance of principal indicator of implicit learning.

Table S3. Disclosure table for the 9 group design studies eligible for the meta-analysis using Hebb serial order learning tasks

Study name	Analysis	Quoted test from paper with statistical results	Significance
Archibald & Joanisse, 2013[1]	ANCOVA: 2 (Group) x 2 (task modality) x 2 (sequence type) x 2 (Task half), with WM and NVIQ as covariates	"The results revealed two significant interactions with group: the interaction between modality and group […] .all remaining effects and interactions involving group were not significant […] Importantly this interaction was not differentiated by list types, indicating a general auditory retention difficulty rather than a specific deficit in carryover learning on the Hebb lists." (p. 274)	Null
Bogaerts et al., 2015 (Expt 1)[2]	Mixed ANOVA: 2 (Group) x 3 (Task) x 2 (Sequence type)	"Crucially, we found a significant interaction between Sequence type and Group, $F(1,46) = 4.73$, $p < .05$, $\pi^2p = .09$. Planned comparisons indicate a HRL effect in both groups, however, HRL was significantly stronger for controls." (p. 111)	Significant for development of implicit learning over task
Bogaerts et al., 2015 (Expt 2)[2]	Mixed ANOVA: 2 (Group) x 3 (Task) x 2 (Sequence type)	"… a significant interaction was found between Sequence type and Group, $F(1,34) = 5.52$, $\pi^2p = 0.14$, $p < .05$." (p. 115)	Significant for development of implicit learning over task

Bogaerts et al., 2016	Mixed logit models (Jaeger, 2008): Fixed vs = Group, Sequence type, task, block, NVIQ as control variable	"A group difference in the disadvantage of the poor readers would surface as a threeway interaction, Type x Presentation x Group, with a negative coefficient…A simple slopes analysis...suggesting that Hebb learning is present in both groups but to a lesser extent for the poor readers, $chi^2(2) = 56.04$, $p < .001$." (p. 146)	Significant for development of implicit learning over task
Gould & Glencross, 1990[1]	Verbal task Mixed ANOVA: 2 (Group) x 2 (sequence type) x 2 (early vs late trials)	"Table 2 shows that Normal Readers were more accurate on the repeated sequences in both the Early and Late Trials whereas the Disabled Readers did not show greater accuracy until the Late Trials." Table 2: Group x sequence interaction effect = ns; Group x sequence x trials: $F(1,18) = 8.6$, $p < .009$ (p. 275)	Null for consistent measure
	Visuospatial task Mixed ANOVA: 2 (Group) x 2 (sequence type) x 2 (early vs late trials)	"Table 3 shows that the pattern of results was very similar for both groups." (p. 275)	Null

Henderson & Warmington, 2017[1]	Mixed ANOVA: 2 (Group) x 2 (sequence type) x 2 (1st half vs 2nd half)	"…a marginally significant List x Half x Group interaction ($F(1,57) = 3.99$, $p = .051$, $\pi^2p = .07$." NB: Group x sequence type = ns (p. 202)	Null for consistent measure
Hsu & Bishop, 2014[2]	3 (Group) ANCOVA, with Random gradient as covariate	"There was a significant effect of group, $F(2,76) = 3.68$, $p = .03$, $\pi^2p = .09$. Pair-wise comparisons indicated that the age-matched group showed a steeper learning rate of word sequences than the SLI and the grammar-matched group." (pp. 357, 358)	Significant for development of implicit learning over task
Majerus et al., 2009[1]	Mixed ANOVA: 3 (Group) x 2 (Sequence type)	"This analysis revealed no significant group effect, $F(2,33) = 1.14$, ns [....] and no interaction effect, : $F(2,33) < 1$, ns." (p. 714)	Null
Staels, & Van der Broek, 2015 (Expt 1) [2]	Mixed ANOVA: 2 (Group) x 3 (Task) x 2 (Sequence type)	"Unlike Szmalec et al. (2011), however, the crucial Group x Sequence type interaction effect was not significant, $F(1,57) = .128$, $p = .722$, $\pi^2p = .002$, indicating a similar Hebb effect for the control and the dyslexic group. Planned comparisons [...] confirmed the absence of a differential Hebb effect for [all 3 tasks]." (p. 6)	Null

Staels, & Van der Broek, 2015 (Expt 2) [2]	Mixed ANOVA: 2 (Group) x 3 (Task) x 2 (Sequence type)	The crucial Group x Sequence type interaction effect was also not significant, $F(1,55) = .087$, $p = .769$, $\pi^2p = .002$, indicating a similar Hebb effect for the control and dyslexic group. (p. 13)	Null
Szmalec et al., 2011[2]	Mixed ANOVA: 2 (Group) x 3 (Task) x 2 (Sequence type)	"The crucial interaction effect between Group and Sequence Type was significant, $F(1,30) = 23.22$, $p < .001$, $\pi^2p = .44$, indicating a stronger Hebb effect for the control group. Further planned comparisons [...] demonstrate that the persons with dyslexia showed reduced Hebb learning for all stimulus and presentation modalities." (p. 12)	Significant for development of implicit learning over task

[1] = Mean proportion of correct responses for Hebb vs Random; [2] = Repeated regression line compared to random one

Table S4. Disclosure table for the 23 group design studies eligible for the meta-analysis using artificial grammar and statistical learning tasks.

Study name	Model	Quoted test from paper with statistical results	Significance of main effect
Aguilar & Plante, 2014 (Expt 1)	Mixed ANOVA: 2 (Group) x 4 (item type: correct seen, correct generalization, co-occurence violation; linear order violation)	"The main effect for Group was not significant, $F(1,22) = .43$, $p = .5186$, $\pi^2p = .02$, nor was the Group x Item Type interaction." (p. 1398)	Null.
Aguilar & Plante, 2014 (Expt 2)	Mixed ANOVA: 2 (Group) x 4 (item type: correct seen, correct generalization, co-occurence violation; linear order violation)	"The main effect of group was not significant, $F(1,54) = 2.49$, $p = .12$, $\pi^2p = .04$. […] This was qualified by a significant Group x Item Type interaction, Wilk's $F(1,162) = 69.03$, $p = .0116$, $\pi^2p = .07$. […] this reflected a general pattern for the NL group to accept more correct items than the LLD group, whereas the LLD group tended to accept more incorrect items than their NL counterpart." (p. 1400)	Null for main effect of group, significant for Group x item type interaction.

Bahl, Plante, & Gerken, 2009 (Expt 1)[1]	Mixed ANOVA: 2 (Group) x 2 (language A vs B) x 2 (item type (correct vs incorrect) x 2 (generalization type - pattern or principle) x 2 (item type - correct & incorrect)	"The ANOVA revealed a significant main effect of group, $F(1,25) = 9.16$, $p < .005$, $\pi^2p = .276$, with hLLD group accepting more items overall than the NL group." (p. 317)	Significant. Insufficient data for meta-analysis.
Bahl, Plante, & Gerken, 2009 (Expt 2)	Mixed ANOVA: 2 (Group) x 2 (item type (correct vs incorrect) x 2 (generalization type: pattern or principle) x 2 (item type - correct & incorrect)	"There was no significant main effect for group, $F(1,24) = 1.39$, $p < .25$, or generalization type." (p. 319)	Null. Insufficient data for meta-analysis.
Du, 2013 (Expt 2a) [1]	Mixed ANOVA: 2 (Group) x 2 (Condition: new grammatical (GN) & nongrammatical (NG))	"A significant effect of Group was also found: $F(1,22) = 11.00$, $p = .003$, $\pi^2 = .33$." (p. 116)	Significant
Du, 2013 (Expt 2b)	Mixed ANOVA: 2 (Group) x 2 (Condition: new grammatical (GN) & nongrammatical (NG))	"…no significant effect was found for Group, $F(1,22) = .59$, $p = .45$, $\pi^2 = .03$…" (p. 122)	Null

Evans, Saffran, & Robe-Torres, 2009 (Expt 1) [1]	ANCOVA: 2 (Group) with Age & NVIQ as covariates	"An analysis of covariance with age and nonverbal IQ as covariates revealed that the SLI group's ability to attend to transitional probabilities in the speech stream was significantly poorer than the NL group's, $F(1,109) = 5.6$, $p < .01$, $\pi^2p = .05$." (p. 7)	Significant
Evans, Saffran, & Robe-Torres, 2009 (Expt 2)	Mixed ANCOVA: 2 (Group) x 2 (Task variant - Speech or Tone) with Age and NVIQ as covariates	"A repeated measures ANCOVA with age and nonverbal IQ as covariates revealed a main effect for group, $F(1,26) = 7.4$, $p = .003$, $\pi^2p = .37$, across the speech and tone conditions, with overall performance for the children with SLI being poorer than that of their typical language peers. (p 9)	Significant
Gabay, Theissen & Holt, 2015[1]	Mixed ANOVA: 2 (Group) x 2(SL task variant)	There was a main effect of group, $F(1,30) = 10.366$, $p = .003$, $\pi^2p = .256$), indicating that the DD group performed significantly less accurately (M = 69%) than the control group (M = 85%). (p. 939)	Significant

Grunow, Spaulding, Gomez, & Plante, 2006	Mixed ANOVA: 2 (Group) x 2 (set size) x 2 (grammaticality) x 2 (item type)	"We predicted that the hL/LD group would perform poorly relative to the ND group overall. However, this between group difference in the ANOVA did not reach statistical significance (F = 0.47, df = 1,40), p = 0.4967)." (p. 164)	Null
Haebig, Saffran, & Weismer, 2017[1]	Mixed-effect logistic regression model for 3 groups (ASD, DLD and TD)	TD vs DLD contrast only: "The TD and ASD groups performed significantly better on the segmentation task than the SLI group (TD vs. SLI group: Estimate = -0.41; SE = .16; z = -2.58…" (p. 1255)	Significant
Hall, Owen Van Horne, McGregor, & Farmer, 2017	Linear mixed effects model (DV = scale rating; main effects = item type, group and their interaction)	"In answer to our primary question of whether group performance differed, we found no main effect of group, p = .19…" (p. 3275)	Null
Hall, Owen Van Horne, McGregor, & Farmer, 2018	Linear mixed effects model (DV = scale rating; main effects = item type, age, group and item order)	Results reported in Table 4: Diagnostic group (reference category = TD); β = -0.08; SE = 0.17; p = .65. (p. 701)	Null

Hsu, Tomblin, & Christiansen, 2014[1]	Mixed ANOVA: 2 (Group) x 3 (variability condition) x 2 (grammaticality)	"There was a significant main effect of grammaticality […] and Grammaticality x Language Group interaction, $F(1,114) = 6.34$, $p = 0.01$, $\pi^2p = .05$." (p. 4)	Significant
Iao, Ng, Wong, & Lee, 2018[1]	Mixed ANOVA: 2 (Group) x 2 (grammaticality) x 2 (item type)	"…a significant grammaticality x group interaction, $F(1,30) = 4.15$, $p = .05$, $\pi^2p = .12$". (p. 10 ScholarOne manuscript)	Significant
Inacio, Faisca, Forkstam, Araujo, Bramao, Reis, & Petersson, 2018	Mixed ANOVA: 2 (Group) x 2 (grammaticality) x 2 (chunk strength)	"Importantly, there was no main effect of group [$F(2,57) = 0.10$, $p = 0.903$; $\pi^2p = 0.004$]." (p. 8)	Null
Kahta & Schiff, 2016	Mixed ANOVA: 2 (Group) x 2 (grammaticality)	"No significant main effect was found for group (F 1<). However, there was a significant interaction for grammaticality x group, $F(1, 27) = 11.86$, $p = .002$, $\pi^2p = .3$." (p. 241)	Null for main effect of group. Significant for Group x Grammaticality interaction.
Katan, Kahta, Sasson, & Schiff, 2017 (Expt 1)	Mixed ANOVA: 2 (Group) x 2 (grammaticality)	"The main effect of group, $F(2,60) = 0.43$, $p = 0.65$, $\pi^2 = 0.01$, was not significant." (p. 169)	Null

Katan, Kahta, Sasson, & Schiff, 2017 (Expt 2)	Mixed ANOVA: 2 (Group) x 2 (grammaticality)	"The main effect of group, $F(2,63) = 1.40$, $p = 0.25$, $\pi^2 = 0.04$, was not significant." (p. 172)	Null
Laasonen, Vare et al., 2014	Mixed ANCOVA: 3 (Group) x 2 (answer type: Accuracy vs Similarity)	"A 3 x 2 mixed ANCOVA with Group as a between subjects factor, answer type as a within subjects factor and proportion of correct responses as the dependent variable resulted in a nonsignificant main effect of group ($F(2,84) = 2.416$, $p = .095$, $\pi^2p = .054$..." (p. 22)	Null.
Lukacs & Kemeny, 2014[1]	Univariate ANOVA (Group on performance difference score)	"The control group outperformed the clinical group, as revealed by a significant main effect of group, $F(1,113) = 6.645$, $p < .05$, $\pi^2p = 0.056$." (p. 478)	Significant
Iao, Ng, Wong, & Lee, 2017	Mixed ANOVA: 2 (Group) x 2 (grammaticality) x 2 (Item type)	"A 2 x 2 x 2 three-way mixed analysis of variance […] resulted in a main effect of grammaticality […] and a main effect of item type […] There were no other main effects…" (p. 697)	Null

Mainela-Arnold & Evans, 2014	Analyses relate to whether SL ability predicts performance on lexical gating and definition tasks: Multiple regression with age, NVIQ, SL, Group, Group x SL interaction	From table: predicting lexical phonology: Group x statistical learning interaction: ß = -.08, R^2 = .27, R^2 change = .01, *F* change = .52; predicting lexical-semantics: Group x statistical learning interaction: ß = .36, R^2 = .46, R^2 change = .00, *F* change = .15	N/A
Mayor-Dubois, Zesiger et al., 2014[1]	*T*-test for Group difference	"Significant difference in scores between the SLI and the Control groups, $t(77) = 3.137$, $p < .01$. The performance of the SLI group did not differ from chance level […], contrary to the Control Group who obtained scores above the chance level..." (p. 18)	Significant
Nigro, Jiménez-Fernández et al., 2016 (Expt 1)	*T*-test by Group against chance	"...participants from the TD group performed above chance level in all three cases […] $t(20) = 3.85$, $p = .001$, $r = .65$ […] Participants with DD also performed above chance level in the overall task […] $t(20) = 3.20$, $p = .005$, $r = .58$." (p. 208)	Null for overall difference, but significant difference with transfer to unseen items.

Nigro, Jiménez-Fernández et al., 2016 (Expt 2)	*T*-test by Group against chance	"Results from single-sample t-tests showed that participants from the TD group again performed above chance level in all three cases (overall […]: $t(20) = 4.06$, $p = .001$, $r = .67$). […] Participants with DD also performed above chance level in the overall task (…$t(20) = 3.07$, $p = .006$, $r = .57$)." (p. 211)	Null for overall difference, but significant difference with transfer to unseen items.
Pavlidou, Kelly, & Williams, 2010[1]	Mixed ANOVA: 2 (Group) x 2 (Grammaticality) x 2 (Chunk strength)	"The between subjects ANOVA revealed a main effect of Participant type ($F(1,30) = 4.521$, $p < .05$, p-value reported two-tailed): the two types of children were performing significantly different…" (p. 152)	Significant
Pavlidou & Williams, 2010[1]	Both models: Mixed ANOVA: 2 (group) x 2 (grammaticality) x 2 (chunk strength)	Non transfer task: "Between subjects ANOVA revealed an effect of group ($F(1,30) = 14.46$, $p = .001$): The typical group outperformed the dyslexic group." (p. 3292) Transfer task: "Between subjects tests showed a group effect ($F(1,30) = 4.63$, $p < .05$). The two groups of children were	Both significant

		performing significantly different during the testing phase…" (p. 3294)	
Pavlidou & Williams, 2014	Mixed ANOVA: 2 (Group) x 2 (Grammaticality) x 2 (Chunk strength)	"A main effect of reader Group was obtained (F(1,30) = 14.46, p = .0001), with higher number correct for typically developing children […] than dyslexic children..." (p. 1462) Transfer task: "A main effect of reader Group was obtained (F(1,30) = 4.63, p < .05), such that grammaticality-decisions for the test items were more accurate for TD […] than DD children..." (p. 1465)	Both significant (Same experimental data as Pavlidou and Williams (2010), so not included)
Pavlidou, Williams, & Kelly, 2009[1]	Mixed ANOVA: 2 (Group) x 2 (Grammaticality) x 2 (Chunk strength)	"The ANOVA revealed a main effect of group (F(1,30) = 8.18, p < .01)." (p. 63)	Significant
Plante, Bahl, Vance, & Gerken, 2010 (Expt 1)	Mixed ANOVA: 2 (Group) x 2 (generalization type - pattern or	"No other main effect or interaction effect was significant."	Null. Not in meta-analysis.

	principle) x 2 (item type - correct & incorrect)	Significant effect were not considered relevant to implicit learning by authors (see p. 402)	
Plante, Bahl, Vance, & Gerken, 2010 (Expt 2)	Mixed ANOVA: 2 (Group) x 2 (generalization type - pattern or principle) x 2 (item type - correct & incorrect)	"No other effect was significant […] the variance that contributed to the the three-way interaction occurred only because incorrect items were accepted more frequently than correct items under certain conditions. (p. 403)	Null. Not in meta-analysis.
Plante, Gomez, & Gerken, 2002[1]	*T*-Test for Group difference	"In contrast, the NLD average [...] was both above chance levels and significantly greater than the mean of the L/LD group ($t(30) = 2.75$, $p = .01$)." (p. 458)	Significant
Pothos & Kirk, 2004	Mixed ANOVA: 2 (Group) x 2 (Task variant)	"There was a main effect for the factor Dyslexia ($F(1,210) = 4.39$, $p = .04$), showing that dyslexic participants performed better than non-dylexic ones…" (p. 71)	Effect in opposite direction
Rüsseler, Gerth, & Munte, 2006[1]	3 (Group) ANOVA on grammaticality judgements	"…both the normal and the dyslexic readers' classification scores exceeded that of the random comparison group [...] main effect GROUP: $F(2,33) = 23.94$, $p < .0001$…" (p. 819)	Significant

Samara & Caravolas, 2017 (Expt. 1)	*T*-Test for Group difference (Chunk strength sensitivity)	"Discrimination ability between skilled and dyslexic readers (Table 2) was not statistically different, $t(48) = 0.23$, $p = .817$, $d = 0.07$." (p. 83)	Null
Samara & Caravolas, 2017 (Expt. 2)	*T*-Test for Group difference (Chunk strength sensitivity)	"Dyslexic readers' discrimination ability (Table 4) was not significantly different from that of skilled readers, $t(50 = 0.63$, $p = .531$, $d = 0.18$." (p. 85)	Null
Schiff, Sasson, Star, & Kahta, 2017[1]	Mixed ANOVA: 2 (Group) x 2 (Learning condition)	"A significant main effect of group was found, $F(1,42) = 4.96$, $p < .05$, $\pi^2p = .11$…" (p. 340)	Significant
Sigurdardottir, Danielsdottir, Gudmundsdottir, Hjartarson, Thorarinsdottir, & Kristjansson, 2017[1]	*T*-Test for Group difference	"Dyslexic readers correctly identified significantly fewer base pairs during the statistical learning test than typical readers (Table 1; Fig.1; independent samples t-test, $t(72) = 2.449$, $p = 0.017$, $d = 0.569$)…" (p. 4)	Significant

[1] = Included in p-curve

Table S5. Disclosure table for the 6 group design studies eligible for the meta-analysis using weather prediction tasks

Study name	**Analysis**	**Quoted test from paper with statistical results**	**Significance**
Gabay, Vakil, Schiff & Holt, 2015	Mixed ANOVA: 2 (Group) x 2 (Task: FB vs PA)	"The main effect of Group was significant, $F(1, 28) = 7.51$, $p = .011$, $\pi^2p = .204$, indicating that test-phase accuracy of the dyslexia group […] was poorer than that of the control group." (p. 6)	Significant for overall group difference for 2 tasks.
Kemeny & Lucaks, 2010	Mixed ANOVA: 3 (Group) x 3 (Block)	"There was a significant main effect of group ($F(2,46) = 15.584$, $p < 0.001$, $\pi^2p = .409$) showing that there is a significant difference between the groups with adults giving the most correct answers, followed by typically developing children, and children with LI giving the least […] The group block interaction did not appear to be significant ($F(4,46) = .882$, $p = .478$, $\pi^2p = .409$)…" (p. 18)	Significant for overall group difference
Lee & Tomblin, 2015	Mixed ANOVA: 2 (Group) x 5 (Block)	"Figure 1 (d) shows the results of a significant main effect of Group, $F(1,46) = 6.72$, $p = .01$, $\pi^2p = .13$ […] The interaction effect was not significant, $F(4,184) = .75$, $p = .56$, $\pi^2p = .02$." (pp. 225, 226)	Significant for overall group difference.
Lee, Mueller, & Tomblin, 2016	Mixed ANOVA: 2 (Group) x 5 (Block)	"Results showed a significant Group effect, $F(1,39) = 11.54$, $p = .0021$ […] The interaction effect was not significant, $F(4,156) = .85$, $p = .50$." (p. 1106)	Significant for overall group difference

Lukacs & Kemeny, 2014	Mixed ANOVA: 2 (Group) x 4 (Block)	"The Huyhh-Feldt corrected ANOVA revealed that neither the main effect of block (p = .196) nor the main effect of group (p = .814) was significant. The Block x Group interaction approached, but did not reach significance, F(2.502, 285.197) = 2.302, p = .089." (p. 478) NB: Main effect of Group supplied by authors: F(1,114) = .56, p = .814.	Null.
Mayor-Dubois, Zesiger, Van der Linden, & Roulet-Perez, 2014	Mixed ANOVA: 2 (Group) x 4 (Block)	"...but no interaction between Blocks and Groups, F(3,85) = 1.072, ns, indicating a similar improvement of cognitive learning in both groups." (p. 19) NB: Effect of group not reported.	Null.

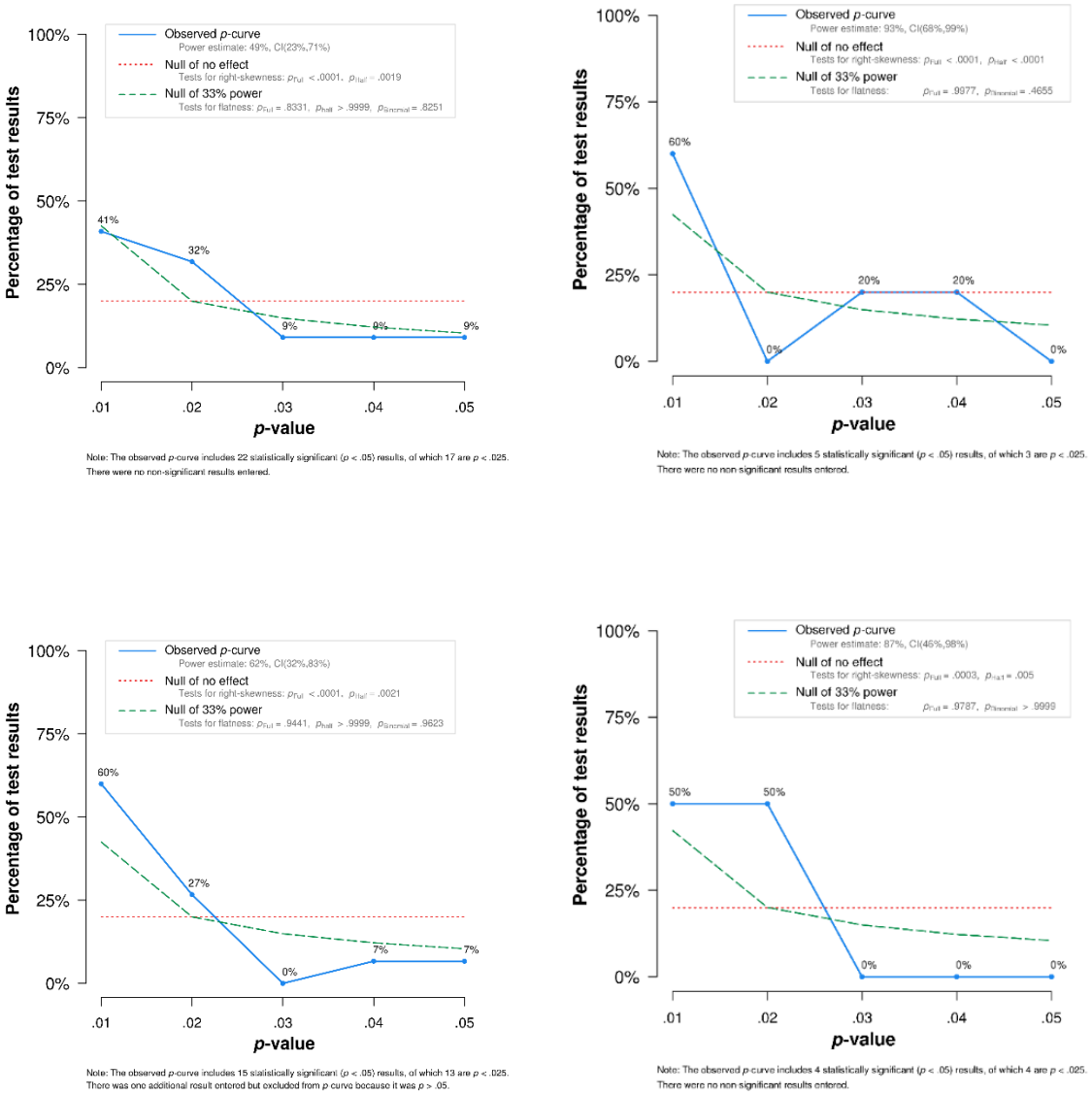

Figure S1. P-curve examining publishing bias in extreme groups studies using the serial reaction time task (top left), Hebb serial order learning tasks (top right), artificial grammar learning or statistical learning tasks (bottom left), and weather prediction tasks (bottom right) to investigate the procedural deficit hypothesis.

4. Interrater reliability

For effect sizes on the primary outcomes for the serial reaction time task (group design and correlational studies) the Pearson correlation between the raters was 0.97, agreement rate 77%, for moderators the correlation was 0.87, agreement rate 84%. For the Hebb task the correlation was 0.86, agreement rate 78% and for moderators of the Hebb task the correlation was 0.98 and agreement rate 76%. For artificial grammar and statistical learning tasks (group design and correlational studies) the correlation was 0.73, agreement rate 70% and for moderators of this task the correlation was 0.97 and agreement rate 78%. Finally, for the weather prediction task the correlation was 0.75 and agreement rate 68%, and for moderators of this the correlation was 0.80 and agreement rate 91%.

5. Tables of characteristics for each study

Table S6

Characteristics of the 52 group design studies eligible for the meta-analysis using the SRT task

Study	Effect size	Sample Size*	Age	Diagnosis	Task	Sequence Complexity	Sequence Length	Sequence Repetitions [f]	Declarative tasks incl.	Additional information
Bennett, Romano, Howard Jr, & Howard, 2008	*g* = 0.14	16; 18	Adult	DD	Alt.	SOC	3	-	No	Single measure (high vs low frequency triplets)
Bussy, Krifi-Papoz, Vieville, Frenay, Curie, Rouselle, Rougeot, Des Portes, & Herbillon, 2011		24; 18	Child	DD	Det.	SOC [d]	10	36	No	
Clark & Lum, 2017a[c]		25; 25	Child	DLD	Det.	FOC/SOC	10	18	No	FOC & SOC tasks
Clark & Lum, 2017b		20; 20	Child	DLD	Det.	FOC	10	18	No	

Conti-Ramsden, Ullman & Lum, 2015		45; 46	Adult	DLD	Det.	Not stated	10	36	CMS & span tasks	
Deroost, Zeischka, Coomans, Bouazza, Depessemier, & Soetens, 2010	*g* = -0.05	28; 28	Child	DD	Det.	FOC/SOC	12	108	Digit span	FOC & SOC tasks
Desmottes, Meulemans, & Maillart, 2016a[b]	*g* = -0.21	21; 21	Child	DLD	Alt.	SOC	10	50	No	3 measures: immediate, 24 hrs later, 1 week later
Desmottes, Meulemans, & Maillart, 2016b		24; 24	Child	DLD	Det. (SST)	-	6	40	No	Motor & verbal tasks
Desmottes, Maillart & Meulemans, 2017		18; 17 & 17; 17	Child	DLD	Alt.	SOC	10	50	No	2 comparisons
Desmottes, Meulemans, Patinec, & Maillart, 2017		30;30 & 30;30	Child	DLD	Alt.	SOC [d]	10	70	No	2 comparisons (distributed & massed learning, each incl. 4 sessions)
Du & Kelly, 2013		12; 12	Adult	DD	Det.	FOC/SOC [e]	12	64	No	

Gabay, Schiff, & Vakil, 2012	*g* = -0.64	14; 14	Adult	DD	Det. (SST)	SOC	8	60	No	Motor & verbal tasks
Gabay, Schiff, & Vakil, 2012a	*g* = -0.56	12; 12	Adult	DD	Det.	SOC	12	36	No	Across 2 sessions. Transfer and recovery measures
Gabriel, Maillart, Guillaume, Stefaniak, & Meulemans, 2011		16; 16	Child	DLD	Prob.	SOC	8	96	No	
Gabriel, Maillart, Stefaniak, Lejeune, Desmottes, & Meulemans, 2013	*g* = -0.36	21; 25	Child	DLD	Det.	SOC	12	48	No	
Gabriel, Meulemans, Parisse, & Maillart, 2015[b]	*g* = 0.28	14; 14	Child	DLD	Det.	SOC	8	48	No	Visual & auditory tasks, only visual task coded
Gabriel, Stefaniak, Maillart, Schmitz, & Meulemans, 2012	*g* = -0.23	15; 15	Child	DLD	Det.	SOC	8	48	No	2 tasks: keyboard and touchscreen versions

He & Tong, 2017[b]	*g* = -.23	27; 28	Child	DD	Det.	SOC	8	180	No	Also included a reading age-matched sample of 28 children; only data for full task included here
Hedenius, Persson, Alm, Ullman, Howard, Howard, & Jennische, 2013[b]	*g* = -0.02	12; 17	Child	DD	Alt.	FOC	8	250	No	3 measures across 2 sessions
Hedenius, Persson, Tremblay, Adi-Japha, Verissimo, Dye, Alm, Jennische, Tomblin, and Ullman, 2011		21; 27	Child	DLD	Alt.	FOC	8	250	No	Across 2 sessions
Henderson & Warmington, 2017	*g* = 0.02	30; 29	Adult	DD	Alt.	FOC	8	45	Span	Across 3 sessions
Howard, Howard, Japikse, & Eden, 2006		23,23	Adult	DD	Alt		8	400	No	
Hsu & Bishop, 2014	*g* = -0.42	48; 20	Child	DLD	Det.	FOC	10	20	No	Only 2nd half measure included

Jiménez-Fernández, Vaquero, Jiménez, & Defior, 2011[b]	*g* = -0.94	14; 14	Child	DD	Det.	SOC	6	74	No [g]	
Kelly, Griffiths, & Frith, 2002		14; 14	Adult	DD	Det.	SOC [d]	9	64	No	
Laasonen, Vare, Oksanen-Hennah, Leppamaki, Tani, Harno, Hokkanen, Pothos, & Cleeremans, 2014[c]		36; 35	Adult	DD	AG.	-	-	-	No	Sequence follows AGL-type grammar. Data appear normalized with Z-score transformation
Lee & Tomblin, 2015	*g* = -0.11	23; 25	Adult	DLD	Det.	SOC	12	18	No	Alternating sequence & random blocks
Lee, Mueller, & Tomblin, 2016		22; 19	Adult	DLD	Det.	SOC	12	18	No	
Lukacs & Kemeny, 2014	*g* = -0.12	28; 87	Child	DLD	Det.	SOC	12	55	No	
Lum & Bleses, 2012[c]		13; 20	Child	DLD	Det.	SOC	10	24	CMS & span	Normalised with Z-score transformation

Lum, Conti-Ramsden, Page, & Ullman, 2012[c]		51; 51	Child	DLD	Det.	-	10	36	CMS & span	Normalised with Z-score transformation
Lum, Gelgic, & Conti-Ramsden, 2010[b]	*g* = -0.68	15;15	Child	DLD	Det.	SOC	10	36	CMS, PAL & span	Log transformed RTs
Mayor-Dubois, Zesiger, Van der Linden, & Roulet-Perez, 2014		18; 65	Child	DLD	Det.	SOC	10	20	No	
Menghini, Finzi, Benassi, Bolzani, Facoetti, Giovagnoli, Ruffino, & Vicari, 2010		60; 65	Child	DD	Det.	FOC	9	30	No	
Menghini, Hagberg, Caltagirone, Petrosini, & Vicari, 2006	*g* = -0.45	14;14	Adult	DD	Det.	SOC [d]	9	30	No	
Menghini, Hagberg, Petrosini, Bozzali, Macaluso, Caltagirone, & Vicari, 2008	*g* = -0.31	10; 10	Adult	DD	Det.	SOC	9	30	No	

Perlant & Largy, 2011		25; 20	Child	DD	Det.	-	6	25	No	
Przekoracka-Krawczyk, Brenk-Krakowska, Nawrot, Rusiak, & Nasrecki, 2017[b]	*g* = -.57	29;30	Adult		Det.	SOC	12	110	No	Task administered to groups viewing monocularly or binocularly (only latter included here)
Rüsseler, Gerth, & Munte, 2006[b]	*g* = 0.48	12; 12	Adult	DD	Det.	SOC	12	80	No	
Sengottuvel & Rao, 2013	*g* = -1.26	17; 23	Child	DLD	Det.	SOC	10	20	No	Random measure taken early in task. SLavg1, 2 & 3 included.
Sengottuvel & Rao, 2014	*g* = -0.62	22; 34	Child	DLD	Det.	SOC	10	20	No	Random measure taken early in task. SLavg1, 2 & 3 included.
Sengottuvel, Rao, & Bishop, 2016	*g* =- 0.18	30; 30	Child	DLD	Det.	FOC	12	40	DecLearn, PAL	Only 2nd half measure

Staels & Van den Broek, 2017	*g* = 0.03	30; 38	Child	DD	Det.	SOC	6	74	No[g]	Replication of task in Jiménez-Fernández et al. (2011)
Stoodley, Harrison, & Stein, 2006	*g* = -0.57	19; 21	Adult	DD	Det.	SOC [d]	10	10	Span	Only 2nd half measure
Stoodley, Ray, Jack, & Stein, 2008	*g* =- 0.12	45; 44	Child	DD	Det.	SOC [d]	6	14	No	Only 2nd half measure
Tomblin, Mainela-Arnold & Zhang, 2007		38; 47	Adol	DLD	Det.	SOC	10	20	No	DLD group: 15 yrs olds with kindergarten diagnosis of DLD
Vakil, Lowe, & Goldfus, 2015		23; 30	Child	DD	Det.	SOC [d]	12	54	No	Older children (age 11 to 13)
Vicari, Finzi, Menghini, Marotta, Baldi, & Petrosini, 2005		16; 16	Child	DD	Det.	SOC [d]	5	60	No	

Vicari, Marotta, Menghini, Molinari, & Petrosini, 2003[b]	*g* = -1.38	18; 18	Child	DD	Det.	FOC	9	24	No	
Yang, Bi, Long, & Tao, 2013	*g* = -0.41	9; 12	Child	DD	Det.	SOC [d]	8	18	No	
Yang & Hong-Yan, 2011	*g* = -0.19	27; 27	Child	DD	Det.	SOC	6	20	No	2 tasks (each hand)
Zwart, Vissers, Kessels, & Maes, 2018		13,17	Child	DLD	Prob. & Det.	SOC	8	20 & 20	No	Prob. & Det. sequence conditions within one combined task

Comparisons in bold are included in final meta-analysis; [*] = Sample size, disordered group first; [a] = Additional information supplied by authors; [b] = Figure of sufficient quality available to enable digitized data extraction that includes labelled error bars (WebPlotDigitizer: Rohatgi, 2017); [c] = Data normalized with z-score transformation, removing between subjects variance; [d] = structure categorized differently from Lum et al's (2013) meta-analysis, with 10 item sequences labelled as SOC; [e] = Both conditional properties within one sequence; [f] = repetitions prior to calculation of implicit learning (deterministic tasks) or included in measure of implicit learning (alternating and probabilistic tasks); [g] = implicit and explicit versions of SRT task included in study; DD = Dyslexia; DLD = Developmental Language Disorder; Det. = Deterministic SRT sequence structure; Alt. = Alternating SRT sequence structure; SST = Serial Search Task; Prob. = Probabilistic SRT sequence structure; AG. = Artificial Grammar SRT sequence structure; FOC = first order conditional; SOC = second order conditional; CMS = verbal declarative tasks from the Children's Memory Scale (Cohen, 1967); PAL = Paired Associate Learning; DecLearn = a verbal recognition measure of declarative memory (see Hedenius et al. 2013 for details).

Table S7

Characteristics of the 6 studies eligible for the meta-analysis investigating correlational studies using the SRT task.

Study	Effect size	Sample Size	Age	Diagnosis	Task(s)	Sequence length	Sequence Repetitions
Kidd, 2012	*r* = 0.18	100	Child	Unselected	Det.	10	24
Kidd & Kirjavainen, 2011	*r* = -0.04	120	Child	Unselected [a]	Det.	10	24
Lum & Kidd, 2012	*r* = 0.05	58	Child	Unselected [b]	Det.	10	24
Schmalz, Moll, Mulatti, & Schult-Korne, 2019	*r* = 0.00	65	Adult	Unselected [a]	Det.	16	10
Waber, Marcus, Forbes, Bellinger, Weiler, Sorensen, & Curran, 2003		422	Child	Incl. DD	Det.	6	50
West, Vadillo, Shanks, & Hulme, 2018	*r* = 0.10	98	Child	Unselected	Prob.	12	c.45

Comparisons in bold are included in final meta-analysis; [a] = monolingual only; [b] = children receiving support for language or learning-related problems excluded

Table S8

Characteristics of the 10 group design studies eligible for the meta-analysis using the Hebb serial order learning task.

Study	Effect size	Sample Size*	Age	Diagnosis	Modality	List length	Total trials (Hebb trials)	Additional Information
Archibald & Joanisse, 2013	*g* = -0.14	23; 27	Child	DLD	Verbal (visual & auditory)	Variable (supraspan)	84 (42)	3 sessions
Bogaerts, Szmalec, Hachmann, Page & Duyck, 2015	Expt. 1 *g* = -0.26; Expt. 2 *g* = -0.57	Expt 1: 25; 23 Expt 2: 18; 18	Adult	DD	Verbal-visual	9 items	Expt 1: 9 (3)** Expt 2: 18 (6)**	2 comparisons
Bogaerts, Szmalec, De Maeyer, Page & Duyck, 2016	*g* = -0.52	23; 23	Child	DD	Verbal-visual; visuospatial	6 and 7 items respectively	16 (8)	
Gould & Glencross, 1990		18; 18	Child	DD	Verbal–visual; Visuospatial	Variable (supraspan)	32 (10)	2 tasks
Henderson & Warmington, 2017 [a]	*g* = -0.13	29; 30	Adult	DD	Verbal-auditory	6 items	26 (8)**	Main testing session only

Hsu & Bishop, 2014	*g* = -0.87	28; 20	Child	DLD	Verbal-visual	Variable (supraspan)	13 (5)	
Majerus, Leclercq, Grossmann, Billard, Touzin, Van der Linden, & Poncelet, 2009	*g* = 0.34	12; 12	Child	DLD	Verbal-auditory	Variable (supraspan)	24 (8)	Expt. 2 only
Staels, Van der Broek, 2015[b]	Expt. 1 *g* = -0.33; Expt. 2 *g* = -0.05	26; 32	Expt 1: Adult Expt 2: Child	DD	Verbal-visual; Verbal-auditory; Visuospatial	9 and 7 items respectively	30 (10)	2 comparisons, each has 3 tasks
Szmalec, Loncke, Page, & Duyck, 2011[b]	*g* = -0.72	16; 16	Adult	DD	Verbal- visual; Verbal-auditory; Visuospatial	9 items	30 (10)	3 tasks

Comparisons in bold are included in final meta-analysis; * = Sample size disordered group first; ** = Length of task taken by all participants, as supplied by authors; [a] = proportional scores converted to percentages; [b] = mean raw scores converted to percentages

Table S9

Characteristics of the 31 group design studies eligible for the meta-analysis using artificial grammar and statistical learning tasks.

Study	Effect size	Sample Sizes*	Age	Diagnosis	Task	Domain	Modality	Additional Information
Aguilar & Plante, 2014	Expt. 1 *g* = -0.60; Expt. 2 *g* = -0.56	12; 12 & 28; 28	Adult	DLD	SL	Verbal	Visual	2 comparisons
Bahl, Plante, & Gerken, 2009		15; 15 & 13; 13	Adult	DLD	SL	Non-verbal	Auditory	2 comparisons
Du, 2013	Expt. 2a *g* = -0.69; Expt. 2b *g* = -0.14	12;12 & 12;12	Adult	DD	AGL	Non-verbal	Visual	2 comparisons
Evans, Saffran, & Robe-Torres, 2009	Expt. 1 *g* = -0.48; Expt. 2 *g* = -0.98	35; 78 & 15; 15	Child	DLD	SL	Expt. 1: Verbal; Expt. 2: Verbal & non-verbal	Auditory	2 comparisons
Gabay, Theissen & Holt, 2015	*g* = -0.81	16; 16	Adult	DD	SL	Verbal & non-verbal	Auditory	Single task
Grunow, Spaulding, Gomez, & Plante, 2006 [b]	*g* = -0.20	22; 22	Adult	DLD	SL	Verbal	Auditory	

Haebig, Saffran, & Weismer, 2017	*g* = -0.68	23; 26	Child	DLD	SL	Verbal	Auditory	
Hall, Owen Van Horne, McGregor, & Farmer, 2017		17;17	Adults	DLD	SL	Verbal	Auditory	
Hall, Owen Van Horne, McGregor, & Farmer, 2018		16; 26 children; 17;17 adults	Both	DLD	SL	Verbal	Auditory	Single task, but means for adults and children entered separately
Hsu, Tomblin, & Christiansen, 2014	LV g = -0.20; MV *g* = 0.04; HV *g* = -0.09	20; 20 (in each comparison)	Child	DLD	SL	Verbal	Auditory	3 comparisons
Iao, Ng, Wong, & Lee, 2017	*g* = -0.56	16; 16	Child	DLD	SL	Verbal	Auditory	
Iao, Ng, Wong, & Lee, 2018		16;16	Children	DLD	SL	Verbal	Auditory	

Inacio, Faisca, Forkstam, Araujo, Bramao, Reis, & Petersson, 2018		20;20	Children	DD	AGL	Non-verbal	Visual	
Kahta & Schiff, 2016	*g* = -1.25	14; 15	Adult	DD	AGL	Verbal	Visual	
Katan, Kahta, Sasson, & Schiff, 2017		Expt 1: 19; 26 Expt 2: 18; 24	Children	DD	AGL	Non-verbal	Visual	2 comparisons
Laasonen, Vare, Oksanen-Hennah, Leppamaki, Tani, Harno, Hokkanen, Pothos, & Cleeremans, 2014 [a]	*g* = -0.43	36; 35	Adult	DD	AGL	Non-verbal	Visual	
Lukacs & Kemeny, 2014	*g* = -0.56	28; 87	Child	DLD	AGL	Verbal	Auditory	
Mainela-Arnold & Evans, 2014	*g* = -1.13	20; 20	Child	DLD	SL	Verbal	Auditory	
Mayor-Dubois, Zesiger, Van der Linden, & Roulet-Perez, 2014 [b]	*g* = -0.83	18; 65	Child	DLD	SL	Verbal	Auditory	
Nigro, Jiménez-Fernández, Simpson, & Defior, 2016	Expt. 1 *g* = -0.20; Expt. 2 *g* = -0.44	21; 21 & 21; 21	Child	DD	AGL	Expt. 1: Non-verbal; Expt. 2: Verbal	Visual	2 comparisons

Pavlidou, Kelly, & Williams, 2010	*g* = -0.68	16; 16	Child	DD	AGL	Non-verbal	Visual	
Pavlidou & Williams, 2010	*g* = -1.34	16; 16	Child	DD	AGL	Non-verbal	Visual	
Pavlidou & Williams, 2014 [c]		16; 16	Child	DD	AGL	Non-verbal	Visual	
Pavlidou, Williams, & Kelly, 2009	*g* = -0.86	16; 16	Child	DD	AGL	Non-verbal	Visual	
Plante, Bahl, Vance, & Gerken, 2010		29; 29 & 16; 16	Child	DLD	AGL	Non-verbal	Auditory	2 comparisons
Plante, Gomez, & Gerken, 2002	*g* = -0.93	16; 16	Adult	DD / DLD	SL	Verbal	Auditory	
Pothos & Kirk, 2004 [b]	*g* = 0.66	77; 146	Adult	DD	AGL	Non-verbal	Visual	
Rüsseler, Gerth, & Munte, 2006 [b]	*g* = -0.24	12; 12	Adult	DD	AGL	Verbal	Visual	
Samara & Caravolas, 2017 [b]	Expt. 1 *g* = -0.07; Expt. 2 *g* = -0.18	19; 31 & 21; 31	Adult	DD	AGL	Expt. 1: Verbal; Expt. 2: Non-verbal	Visual	2 comparisons
Schiff, Sasson, Star, & Kahta, 2017	*g* = -0.57	21;25	Adults	DD	AGL	unspecified	unspecified	
Sigurdardottir, Danielsdottir, Gudmundsdottir, Hjartarson, Thorarinsdottir, & Kristjansson, 2017	*g* = -0.56	37;37	Adults	DD	SL	Visual	Non-verbal	

Comparisons in bold are included in final meta-analysis; *Sample size, disordered group first; [a] = Effect size reported in Schmalz et al. (2016); [b] = effect size estimate calculated from reported *t*-test statistic or *F* ratio; [c] = Duplicate data; LV = low variability group; MV = medium variability group; HV = high variability group

Table S10

Characteristics of the 5 correlational studies eligible for the meta-analysis using artificial grammar or statistical learning task.

Study	Effect Size	Sample Size	Age	Task	Domain	Modality	Additional Information
Arciuli & Simpson, 2012	*r* = 0.33; *r* = 0.34	Expt 1: 42 Expt 2: 37	Expt 1: Child Expt 2: Adult	SL	Non-verbal	Visual	2 comparisons
Kidd & Arciuli, 2016	*r* = 0.30	68	Child	SL	Non-verbal	Visual	
Misyak & Christiansen, 2012	*r* = 0.28	30	Adult	SL	Verbal	Auditory	2 tasks (adjacent & non-adjacent dependencies)
Qi, Araujo, Georgan, Gabrieli, & Arciuli, 2019	r = 0.37 (or r = 0.05 children only)	72	Both (36 of each)	SL	Both	Visual & Auditory	Correlation with sentence reading for whole sample and word / nonword reading for subset of 36 children
Schmalz, Moll, Mulatti, & Schult-Korne, 2019	*r* = 0.13	40	Adult	AGL	Verbal	Visual	Stimuli were familiar keyboard symbols, so task is ostensibly verbal

All comparisons are included in final meta-analysis.

Table S11

Characteristics of the 9 group design studies eligible for the meta-analysis using the weather prediction task.

Study name	Effect size	Sample Size*	Age	Diagnosis	Task variant	Trial Total	Combinations	Stimuli	Probabilities (of sun)
Gabay, Vakil, Schiff & Holt, 2015 [b]	*g* = -1.27	15; 15	Adult	DD	Holl et al. (2012)	150	14	Geometric	89%; 78%; 22%; 11%
Kemeny & Lucaks, 2010	*g* = -0.92	16; 16 [d]	Child	DLD	Not stated	150	Not stated	Geometric	90%; 70%; 30%; 10%
Lee & Tomblin, 2015 [c]	*g* = -0.40	23; 25	Adult	DLD	Knowlton et al. (1994)	50	14	Not stated	75%; 57%; 43%; 25%
Lee, Mueller, & Tomblin, 2016 [a]		22; 19	Adult	DLD	Knowlton et al. (1994)	50	14	Not stated	75%; 57%; 43%; 25%
Lukacs & Kemeny, 2014	*g* = -0.02	29; 87	Child	DLD	Knowlton et al. (1994)	200	13	Geometric	85.7%; 70%; 30%; 14.3%
Mayor-Dubois, Zesiger, Van der Linden, & Roulet-Perez, 2014	*g* = -0.82	18; 65	Child	DLD	Shohamy et al. (2004)	200	14	Mr Potato Head	20%; 40%; 60%; 80%

Comparisons in bold are included in final meta-analysis; * = Sample size, disordered group first; [a] = insufficient data for inclusion in meta-analysis; [b] = included feedback and paired associate versions of the task; [c] = 3 groups took this task (16 DLD & 16 TD children & 16 normal adults - only age-matched groups are coded); [d] = effect size estimate calculated from reported *t*-test statistic.

Table S12

Characteristics of the 4 group design studies eligible for meta-analysis using the contextual cueing task.

Study name	Sample Size*	Age	Diagnosis	Task variant	Result
Bennett, Romano, Howard Jr, & Howard, 2008	16; 18	Adult	DD	Chun & Jiang, 1998	Null
Howard, Howard, Japikse, & Eden, 2006	11; 12	Adult	DD	Chun & Jiang, 1998	Null
Jiménez-Fernández, Vaquero, Jiménez, & Defior, 2011 (Expt. 3)	24; 26	Child	DD	Jiménez & Vázquez, 2008	Null
Staels & Van den Broeck, 2017	30; 38	Child	DD	Merrill et al., 2013	Null

* = Sample size, disordered group first