

**Trends in Testing Effect Research:
From Lab to Classroom, but Not Yet for All Learners
Supplementary Materials**

Supplementary Note 1: Methodological flowchart and literature search

Supplementary Figure 1 shows the flow of the literature review process, from initial database searches through term extraction, filtering, and final clustering of relevant studies. Two separate literature searches were performed using the Web of Science (WoS) research database [1]. Both searches included quantitative, qualitative, mixed-method, and review articles published in peer-reviewed journals. Book chapters and conference proceedings were also considered, while PhD theses and studies lacking an English title or abstract were excluded. No publication date restrictions were applied.

Search A was performed to gather all existing literature on retrieval practice, and used the search terms listed in Supplementary Table 1. This first search yielded 1,774 results. The second search was intended to *also* gather all literature on specific learning disabilities, and used the set of search terms listed in the second row of Supplementary Table 1. The latter search yielded 23,850 results.

Supplementary Note 2: Journal categories

For the 20 journals that produced overall most output in the testing effects research field (sources that published at least 15 publications covering the testing effect), we categorized the main aim of the journal as 'basic', 'applied' or 'mixed' based on its editorial statement. Categories for these journals can be found in Supplementary Table 2.

Supplementary Note 3: Thematic analysis

The network- and bibliographic analyses reported in this paper were conducted using VOSviewer 1.6.20 [2]. The thematic keyword analysis was based on the testing effects literature only (Search A, see Supplementary Table 1). VOSviewer uses natural language processing algorithms to extract unique noun phrases from the title and abstract text of all documents. Structured abstract labels and copyright statements were removed before extracting these noun phrases. Next, a sentence detection algorithm splits the text into individual sentences. Subsequently, a part-of-speech tagging algorithm assigns grammatical roles to each word. Noun phrases are defined as individual sequences of words ending in a noun, with the longest possible phrases being prioritized. Lastly, noun phrase unification standardizes phrases by removing special characters, accents, and converting plural forms into singular equivalents. This process resulted in a total of 26,279 noun phrases. We used a binary counting method, indicating that only the presence or absence of a term matters: the number of times a term is mentioned within an article is not tracked.

To create a thematic map suitable for visualization and interpretation, it is necessary to select a subset of noun phrases. Using default selection criteria [2], we chose noun phrases that appeared at least 10 times, resulting in a subset of 755 phrases. To enhance the quality of the thematic map, VOSviewer filters out general terms by using relevance scores. To determine the relevance score, for each term, a distribution of (second-order) co-occurrences over all terms is determined. Noun phrases have a low relevance score if their co-occurrences with other noun phrases follow a more or less random pattern, while they receive a higher relevance score if they co-occur mainly with a limited set of other noun phrases. General noun phrases, such as 'paper' or 'new method', tend to exhibit uniform co-occurrence distributions, while specific noun phrases, like 'text mining' or 'natural language processing', display biased distributions linked to certain other phrases. High-relevance noun phrases are typically clustered together, with each cluster representing a distinct topic. Here, we used the default exclusion of the 40% items with the lowest relevance score, resulting in 453 final phrases that were used to create the network. A table with all extracted terms can be found on the Open Science Framework (<https://osf.io/cz6th/files/8an4f>).

Finally, we used a modularity-based clustering technique to graphically cluster noun phrases based on co-occurrence in the same article (resulting in Figure 1D in the main paper). With modularity-based clustering, noun phrases are grouped by maximizing a so-called ‘modularity score’, which measures how densely connected the items are within their own group compared to random connections. For the clustering analysis, we did not deviate from default settings: We used the default *association strength* method for normalizing the strength of the links between items, visualization parameters 2 for attraction and 1 for repulsion, a clustering resolution of 1.00, and a minimal cluster size of 1 [see 2].

Supplementary Note 4: Co-citation

The co-citation analysis was based on the literature search on the testing effect and SLDs combined (Search B, see Supplementary Table 1). We selected papers that were cited at least once, resulting in a set of 19,204 (out of 23,850) research items. Finally, we only included items that were co-cited with at least one other research item in the search set, resulting in a set of 16,331 research items to create the network.

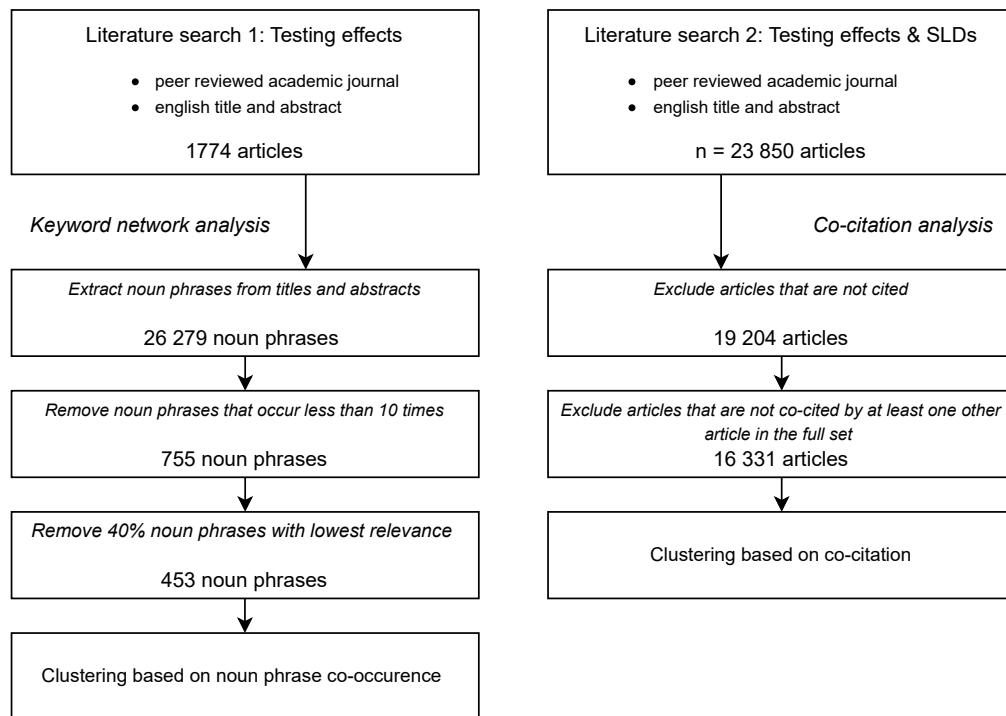
As in the thematic analysis, a modularity-based clustering technique was used to graphically cluster research items, based on the number of times they were cited by the same other source (resulting in Figure 3 in the main paper). We used the same, default clustering parameters as for the thematic analysis.

In total, we found 61 clusters of research items. Nineteen clusters contained more than 10 research items, and topic interpretations of each of these clusters are listed in Supplementary Table 3.

A table with all 204 research items that were either part of the testing effect research cluster (3), and co-cited with a paper in any of the other SLD clusters, or part of any of the other SLD clusters, and co-cited with a paper in the retrieval practice field, can be found on the Open Science Framework (<https://osf.io/cz6th/files/3pumx>).

Supplement 5: DLD and Testing effects - empirical work

A summary of all eight original empirical studies that investigated the effectiveness of testing/retrieval practice against passive study techniques in learners with SLDs is shown in Supplementary Table 4. All eight studies focused on DLD.



Supplementary Figure 1. Workflow for literature search and clustering analyses. The flowchart summarizes the steps used to identify themes in the testing effects literature (left), as well as the steps to complete the co-citation coupling analysis to examine overlap between research on the testing effect, and research on SLDs.

Supplementary Table 1. Web of Science searches

Search terms	N items	Search date	Search engine
A. 'retrieval practice' (all fields) OR 'testing effect' (all fields) OR 'test-enhanced learning' (all fields)	1,774	16-12-2024	Web of Science
B. 'specific learning disorder' (all fields) OR 'specific learning disabilities' (all fields) OR 'developmental language disorder' (all fields) OR 'dyslexia' (all fields) OR 'dyscalculia' (all fields) OR 'dysgraphia' (all fields) OR 'special educational needs' (all fields) OR 'retrieval practice' (all fields) OR 'testing effect' (all fields) OR 'test-enhanced learning' (all fields)	23,850	16-12-2024	Web of Science

Supplementary Table 2. Journal categories by editorial statement

Journal	N items	Focus
Memory	81	Basic
Journal of Experimental Psychology: Learning Memory and Cognition		Mixed
Memory & Cognition	80	Basic
Educational Psychology Review	77	Applied
Applied Cognitive Psychology	54	Applied
Journal of Applied Research in Memory and Cognition	48	Applied
Frontiers in Psychology	40	Mixed
Journal of Experimental Psychology: Applied	39	Applied
Journal of Memory and Language	39	Mixed
Psychonomic Bulletin & Review	34	Basic
Journal of Educational Psychology	28	Applied
Teaching of Psychology	23	Basic
Journal of Cognitive Psychology	22	Mixed
Quarterly Journal of Experimental Psychology	22	Basic
Medical Teacher	19	Applied
Advances in Physiology Education	17	Applied
Experimental Psychology	17	Basic
Medical Education	17	Applied
Learning and Instruction	16	Applied
Cognitive Research-Principles and Implications	15	Applied
Medical Science Educator	15	Applied
Psychological Science	15	Basic

Supplementary Table 3. Cluster interpretations for clusters with more than 10 research items each

Cluster	N items	Interpretation
1	1526	Language disorders
2	1504	Special education
3	1352	Testing effects
4	1290	Dyslexia
5	1155	Speech, developmental language disorder
6	1122	Visual perception, eye movements
7	1116	Structural brain imaging
8	1102	Auditory perception
9	1024	Educational practice
10	971	Arithmetic, mathematics
11	900	Brain mechanisms, functional processing
12	709	Genetics
13	709	Applied interventions
14	586	Logographics
15	527	Motor skills
16	363	Screening
17	126	Autism
18	96	Medication
19	74	Neuroimaging, memory

Supplementary Table 4. Summary of Retrieval Practice Studies in Developmental Language Disorder (DLD)

Author (year)	Title	N (DLD/TD)	Context	Key result(s)
Levlin <i>et al.</i> [3]	Evaluating the effect of rich vocabulary instruction and retrieval practice on the classroom vocabulary skills of children with (developmental) language disorder	11 / –	Classroom	Retrieval practice enhanced vocabulary learning relative to rich instruction.
Leonard <i>et al.</i> [4]	Retrieval practice and word learning by children with developmental language disorder: Does expanding retrieval provide additional benefit?	14 / 14	Laboratory	Short spacing within-session influenced short-term, but not long-term, retention.

Author (year)	Title	N (DLD / TD)	Context	Key result(s)
Leonard <i>et al.</i> [5]	Can retrieval practice facilitate verb learning in children with developmental language disorder and their peers with typical language development?	11 / 12	Laboratory	Retrieval practice facilitated verb learning, but benefits did not generalize to inflection learning.
Leonard <i>et al.</i> [6]	After initial retrieval practice, more retrieval produces better retention than more study in the word learning of children with developmental language disorder	13 / 13	Laboratory	Continued retrieval after initial practice boosted word learning relative to restudy.
Leonard <i>et al.</i> [7]	Adjective learning in young typically developing children and children with developmental language disorder: A retrieval-based approach	14 / 13	Laboratory	Repeated spaced retrieval led to better adjective learning compared to restudy.
Leonard <i>et al.</i> [8]	Retrieval-based word learning in young typically developing children and children with developmental language disorder I: The benefits of repeated retrieval	10 / 10	Laboratory	Repeated retrieval significantly benefited word retention relative to traditional word learning.
Gordon <i>et al.</i> [9]	A word-learning intervention pilot study utilizing principles of retrieval- and criterion-based learning for children with developmental language disorder	7 / 0	clinical-style intervention	Retrieval-based learning was feasible for learners with DLD.
Haebig <i>et al.</i> [10]	Retrieval-based word learning in young typically developing children and children with developmental language disorder II: A comparison of retrieval schedules	16 / 16	Laboratory	Retrieval practice with contextual reinstatement improved long-term word retention.

References

1. Birkle, C., Pendlebury, D. A., Schnell, J. & Adams, J. Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies* **1**, 363–376 (2020).
2. Arruda, H., Silva, E. R., Lessa, M., Proença Jr, D. & Bartholo, R. VOSviewer and bibliometrix. *Journal of the Medical Library Association: JMLA* **110**, 392 (2022).
3. Levlin, M., Wiklund-Hörnqvist, C., Sandgren, O., Karlsson, S. & Jonsson, B. Evaluating the effect of rich vocabulary instruction and retrieval practice on the classroom vocabulary skills of children with (developmental) language disorder. *Language, Speech, and Hearing Services in Schools* **53**, 542–560 (2022).
4. Leonard, L. B., Christ, S. L., Deevy, P., Karpicke, J. & Kueser, J. B. Retrieval Practice and Word Learning by Children With Developmental Language Disorder: Does Expanding Retrieval Provide Additional Benefit? *Journal of Speech, Language, and Hearing Research* **67**, 1530–1547 (2024).
5. Leonard, L. B. *et al.* Can retrieval practice facilitate verb learning in children with developmental language disorder and their peers with typical language development? *Journal of Speech, Language, and Hearing Research* **66**, 1309–1333 (2023).
6. Leonard, L. B., Deevy, P., Karpicke, J. D., Christ, S. L. & Kueser, J. B. After initial retrieval practice, more retrieval produces better retention than more study in the word learning of children with developmental language disorder. *Journal of Speech, Language, and Hearing Research* **63**, 2763–2776 (2020).
7. Leonard, L. B. *et al.* Adjective learning in young typically developing children and children with developmental language disorder: A retrieval-based approach. *Journal of Speech, Language, and Hearing Research* **62**, 4433–4449 (2019).
8. Leonard, L. B. *et al.* Retrieval-based word learning in young typically developing children and children with developmental language disorder I: The benefits of repeated retrieval. *Journal of Speech, Language, and Hearing Research* **62**, 932–943 (2019).
9. Gordon, K. R., Storkel, H. L., Lowry, S. L. & Sultani, M. J. A word-learning intervention pilot study utilizing principles of retrieval-and criterion-based learning for children with developmental language disorder. *American Journal of Speech-Language Pathology* **33**, 530–551 (2024).
10. Haebig, E. *et al.* Retrieval-based word learning in young typically developing children and children with development language disorder II: A comparison of retrieval schedules. *Journal of Speech, Language, and Hearing Research* **62**, 944–964 (2019).