

The DuckyHelper Learning-Science Library

An annotated bibliography of the 100+ papers DuckyHelper's tutor is built on

The DuckyHelper team, MingLLM, Inc. October 2026 • Version 1.0 • hello@duckyhelper.com

This is the full list of the papers in DuckyHelper's learning-science library (146 entries), grouped into sixteen shelves: the twelve principles the tutor follows and four research topics. Every entry is a real paper. Its title, authors, year and venue were matched against Crossref or OpenAlex. The note under each entry is our own one-line summary in plain words, never a quote. A quote is printed only where we found it word for word in the paper's PDF, its text-recognized scan, or its published abstract; the source of each check is given in brackets. A paper can sit on more than one shelf; it is listed once, on its main shelf. The interactive version is at duckyhelper.com/research/library.

1 One-on-one tutoring

Why a good tutor beats a lecture, and what the tutors do.

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13(6), 4-16. <https://doi.org/10.3102/0013189X013006004>

Tutored students scored about two standard deviations above students taught in class.

"the average student under tutoring was about two standard deviations above the average of the control class" [checked in the scanned PDF]

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25(4), 471-533. https://doi.org/10.1207/s15516709cog2504_1

Students learned just as well when tutors prompted instead of explaining.

"students learned just as effectively even when tutors were suppressed from giving explanations and feedback" [checked in the PDF]

Lepper, M. R., & Woolverton, M. (2002). The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. *Improving Academic Achievement* (Academic Press, book chapter). <https://doi.org/10.1016/B978-012064455-1/50010-5>

The best math tutors tend motivation and emotion, not just cognition.

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The Promise of Tutoring for PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *American Educational Research Journal*, 61(1), 74-107. <https://doi.org/10.3102/00028312231208687>

Tutoring programs show consistent, substantial gains, about 0.29 standard deviations.

“tutoring programs yield consistently substantial positive impacts on learning” [checked in the published abstract]

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4), 197-221. <https://doi.org/10.1080/00461520.2011.611369>

Step-based tutoring systems came nearly as close as human tutors.

“the effect size of intelligent tutoring systems was 0.76, so they are nearly as effective as human tutoring” [checked in the PDF]

2 AI answers and learning

What happens to learning when an AI hands over the answer.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>

Unguarded GPT-4 help lifted practice grades, then cut exam grades 17%.

“unfettered access to GPT-4 can harm educational outcomes” [checked in the PDF]

3 Retrieval practice

Pulling an idea out of memory makes it stick better than reading it again.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, 14(1), 4-58. <https://doi.org/10.1177/1529100612453266>

Of ten study techniques, practice testing and spacing were rated most useful.

“Practice testing and distributed practice received high utility assessments” [checked in the PDF]

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*, 331(6018), 772-775. <https://doi.org/10.1126/science.1199327>

Practicing recall beat concept mapping, even on a concept-mapping test.

“practicing retrieval produces greater gains in meaningful learning than elaborative studying with concept mapping” [checked in the PDF]

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, 319(5865), 966-968. <https://doi.org/10.1126/science.1152408>

Repeated testing, not repeated studying, drove long-term vocabulary recall.

“Repeated studying after learning had no effect on delayed recall, but repeated testing produced a large positive effect” [checked in the PDF]

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, 17(3), 249-255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>

Taking recall tests beat restudying for remembering material days later.

“Testing is a powerful means of improving learning, not just assessing it.” [checked in the PDF]

4 Spaced practice

Coming back to an idea after a gap beats cramming it in one go.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354-380. <https://doi.org/10.1037/0033-2909.132.3.354>

Across 317 experiments, spacing helps, and longer retention calls for longer gaps.

“the ISI producing maximal retention increased as retention interval increased” [checked in the PDF]

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot (book; English: *Memory*, 1913). <https://archive.org/details/berdasgedchtnis00ebigoog>

First experiments on memory, charting the forgetting curve and effects of repetition.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide (NCER 2007-2004), U.S. Department of Education. <https://ies.ed.gov/ncee/wwc/PracticeGuide/1>

Seven research-based recommendations, including spacing, worked examples and quizzing.

5 Worked examples

Studying a solved example first helps beginners learn a new kind of problem.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research*, 70(2), 181-214. <https://doi.org/10.3102/00346543070002181>

Good worked examples are integrated and placed right next to matched practice.

“examples should be presented in close proximity to matched practice problems” [checked in the published abstract]

Renkl, A. (2014). Toward an Instructionally Oriented Theory of Example-Based Learning. *Cognitive Science*, 38(1), 1-37. <https://doi.org/10.1111/cogs.12086>

Examples are a very effective way to start learning a new skill.

“Learning from examples is a very effective means of initial cognitive skill acquisition” [checked in the PDF]

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Conventional problem solving eats working memory needed to learn the underlying schemas.

“conventional problem-solving activity is not effective in schema acquisition” [checked in the PDF]

Sweller, J., & Cooper, G. A. (1985). The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra. *Cognition and Instruction*, 2(1), 59-89. https://doi.org/10.1207/s1532690xi0201_3

Studying worked examples beat problem solving for novices: faster, with fewer errors.

“worked examples, as expected, require considerably less time to process than conventional problems” [checked in the PDF]

6 Scaffolding and hints

Give just enough help, one step at a time, and fade it as skill grows.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The Expertise Reversal Effect. *Educational Psychologist*, 38(1), 23-31. https://doi.org/10.1207/S15326985EP3801_4

Techniques that help novices can backfire with more experienced learners.

“Instructional techniques that are highly effective with inexperienced learners can lose their effectiveness”
[checked in the PDF]

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist*, 41(2), 75-86. https://doi.org/10.1207/s15326985ep4102_1

For novices, minimal guidance is less effective than strong, explicit guidance.

“minimally guided instruction is less effective and less efficient than instructional approaches that place a strong emphasis on guidance” [checked in the PDF]

Razzaq, L., & Heffernan, N. T. (2006). Scaffolding vs. Hints in the ASSISTment System. ITS. https://doi.org/10.1007/11774303_63

Compares breaking a missed problem into scaffolding questions against giving hints on request in ASSISTments.

Rosenshine, B. (2012). Principles of Instruction: Research-Based Strategies That All Teachers Should Know. *American Educator*. <https://www.aft.org/sites/default/files/periodicals/Rosenshine.pdf>

Ten research-based principles, like small steps and checking every student's understanding.

“Present new material in small steps with student practice after each step” [checked in the PDF]

Stamper, J. C., Eagle, M., Barnes, T., & Croy, M. (2011). Experimental Evaluation of Automatic Hint Generation for a Logic Tutor. *AIED*. https://doi.org/10.1007/978-3-642-21869-9_45

Hints generated automatically from past students' data helped students persist in a logic course.

“hints help students persist in deductive logic courses” [checked in the PDF]

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press (book). <https://www.hup.harvard.edu/books/9780674576292>

Learners can do more with guidance than alone: the zone of proximal development.

“the distance between the actual developmental level as determined by independent problem solving”
[checked in the scanned PDF]

Wood, D., Bruner, J. S., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89-100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>

Good tutors handle what is beyond the learner and leave the rest to them.

7 Pictures and words

People learn more from words with pictures than from words alone.

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3), 183-198. <https://doi.org/10.1016/j.learninstruc.2006.03.001>

The DeFT framework for when learning with several representations (text, graphs, pictures) helps.

“Multiple (external) representations can provide unique benefits when people are learning complex new ideas” [checked in the PDF]

Ainsworth, S., Prain, V., & Tytler, R. (2011). Drawing to Learn in Science. *Science*, 333(6046), 1096-1097. <https://doi.org/10.1126/science.1204153>

Argues that drawing should be a key part of science education.

“Emerging research suggests drawing should be explicitly recognized as a key element in science education” [checked in the PDF]

Bobek, E., & Tversky, B. (2016). Creating visual explanations improves learning. *Cognitive Research: Principles and Implications*, 1(1), 27. <https://doi.org/10.1186/s41235-016-0031-6>

Students who created visual explanations of a bicycle pump and of chemical bonding learned more.

“creating a visual explanation increased understanding particularly for participants of low spatial ability” [checked in the PDF]

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3(3), 149-210. <https://doi.org/10.1007/BF01320076>

Memory has verbal and imagery codes, and teaching can use both.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, 106(2), 1047-1058. <https://doi.org/10.1016/j.cognition.2007.04.010>

Children who gestured while learning a new math idea kept what they learned better.

“requiring children to gesture while learning the new concept helped them retain the knowledge they had gained during instruction” [checked in the PDF]

Fiorella, L., & Mayer, R. E. (2016). Effects of observing the instructor draw diagrams on learning from multimedia messages. *Journal of Educational Psychology*, 108(4), 528-546. <https://doi.org/10.1037/edu0000065>

Watching an instructor draw diagrams while explaining helped low prior knowledge learners transfer more than seeing finished diagrams.

“watching the instructor draw diagrams (by viewing the instructor’s full body) resulted in significantly better transfer test performance than viewing already-drawn diagrams for learners with low prior knowledge” [checked in the published abstract]

Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and Imagined Activity Can Enhance Young Children’s Reading Comprehension. *Journal of Educational Psychology*, 96(3), 424-436. <https://doi.org/10.1037/0022-0663.96.3.424>

Young readers who moved toys (or imagined moving them) to act out a story understood and remembered it better.

“Both actual manipulation and imagined manipulation resulted in markedly better (compared with rereading) memory for and comprehension of the text material” [checked in the PDF]

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture’s Role in Speaking, Learning, and Creating Language. *Annual Review of Psychology*, 64(1), 257-283. <https://doi.org/10.1146/annurev-psych-113011-143802>

A review of how gesture helps people speak, think and learn.

“the gestures speakers produce when they talk are integral to communication” [checked in the PDF]

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing Gives Children New Ideas About Math. *Psychological Science*, 20(3), 267-272. <https://doi.org/10.1111/j.1467-9280.2009.02297.x>

Children told to make correct gestures during a math lesson learned more than children making partial or no gestures.

“children required to produce correct gestures learned more than children required to produce partially correct gestures, who learned more than children required to produce no gestures” [checked in the PDF]

Hegarty, M. (2011). The Cognitive Science of Visual-Spatial Displays: Implications for Design. *Topics in Cognitive Science*, 3(3), 446-474. <https://doi.org/10.1111/j.1756-8765.2011.01150.x>

A review of the cognitive science of graphs, maps and diagrams, and what it means for designing them.

“This paper reviews cognitive science perspectives on the design of visual-spatial displays” [checked in the PDF]

Larkin, J. H., & Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11(1), 65-100. <https://doi.org/10.1111/j.1551-6708.1987.tb00863.x>

Diagrams can make problem solving easier because they put related information in one place.

“Diagrammatic representations are indexed by location in a plane” [checked in the PDF]

Mayer, R. E., & Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist*, 38(1), 43-52. https://doi.org/10.1207/S15326985EP3801_6

People learn by linking pictures and words, if neither channel is overloaded.

“meaningful learning involves cognitive processing including building connections between pictorial and verbal representations” [checked in the PDF]

Schmeck, A., Mayer, R. E., Opfermann, M., Pfeiffer, V., & Leutner, D. (2014). Drawing pictures during learning from scientific text: testing the generative drawing effect and the prognostic drawing effect. *Contemporary Educational Psychology*, 39(4), 275-286. <https://doi.org/10.1016/j.cedpsych.2014.07.003>

Tests whether drawing pictures while reading a science text improves learning, and whether drawing quality predicts it.

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4), 247-262. <https://doi.org/10.1006/ijhc.2002.1017>

A review that asks whether animation beats static graphics, and finds the benefits often come from other differences.

“The assumption is that graphics can facilitate comprehension, learning, memory, communication and inference” [checked in the PDF]

Van Meter, P., & Garner, J. (2005). The Promise and Practice of Learner-Generated Drawing: Literature Review and Synthesis. *Educational Psychology Review*, 17(4), 285-325. <https://doi.org/10.1007/s10648-005-8136-3>

A review of research on students making their own drawings to learn.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, 69(9), 1752-1776. <https://doi.org/10.1080/17470218.2015.1094494>

Drawing a word's meaning made it easier to recall than writing the word out, across seven experiments.

“Drawn words were better recalled than written” [checked in the published abstract]

8 Feedback

Quick, specific feedback about the task is one of the strongest levers in learning.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward Meta-cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, 16(2), 101-128. [https://doi.org/10.3233/irg-2006-16\(2\)02](https://doi.org/10.3233/irg-2006-16(2)02)

Extends the Cognitive Tutor with a model of good and bad help seeking so it can coach how students ask for help.

“an intelligent tutoring system that provides guidance with respect to students’ meta-cognitive abilities can help them to become better learners” [checked in the PDF]

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research*, 73(3), 277-320. <https://doi.org/10.3102/00346543073003277>

A review of on-demand help in learning software: learners often do not use help well.

“learners are not using help facilities effectively” [checked in the published abstract]

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4(2), 167-207. https://doi.org/10.1207/s15327809jls0402_2

Tutors giving immediate feedback matched classroom results in a third of the time.

“the tutor provides immediate feedback, consisting of short and directed error messages” [checked in the scanned PDF]

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-Task Behavior in the Cognitive Tutor Classroom: When Students “Game the System”. *CHI*. <https://doi.org/10.1145/985692.985741>

Students who abused hints and feedback to get answers learned much less than students who did not.

“students who frequently game the system score substantially lower on a post-test than students who never game the system” [checked in the PDF]

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-like Events. *Review of Educational Research*, 61(2), 213. <https://doi.org/10.2307/1170535>

A meta-analysis of 58 effect sizes on feedback after test-like questions.

“The present meta-analysis reviewed 58 effect sizes from 40 reports” [checked in the published abstract]

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. <https://doi.org/10.1080/0969595980050102>

A landmark review of classroom formative assessment and the learning gains from frequent feedback.

“innovations designed to strengthen the frequent feedback that students receive about their learning yield substantial learning gains” [checked in the PDF]

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, 65(3), 245. <https://doi.org/10.2307/1170684>

A theory of how feedback works inside self-regulated learning, including the feedback learners give themselves.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>

Feedback strongly shapes learning, but its type and delivery decide the effect.

“Feedback is one of the most powerful influences on learning and achievement” [checked in the PDF]

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254-284. <https://doi.org/10.1037/0033-2909.119.2.254>

Feedback helps on average, but over a third of interventions hurt performance.

“FI effectiveness decreases as attention moves up the hierarchy closer to the self and away from the task” [checked in the PDF]

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218. <https://doi.org/10.1080/03075070600572090>

Seven principles of good feedback practice that help students take control of their own learning.

“seven principles of good feedback practice that support self-regulation” [checked in the PDF]

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2), 267-280. <https://doi.org/10.1016/j.learninstruc.2010.07.004>

The Help Tutor gave students immediate feedback on how they used hints in a geometry tutor.

“The present research investigated whether immediate metacognitive feedback on students' help-seeking errors can help students acquire better help-seeking skills” [checked in the PDF]

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144. <https://doi.org/10.1007/bf00117714>

A theory of formative assessment: learners need to know the goal, where they are, and how to close the gap.

“Feedback is defined in a particular way to highlight its function in formative assessment” [checked in the PDF]

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153-189. <https://doi.org/10.3102/0034654307313795>

Formative feedback works best when supportive, timely and specific.

“formative feedback should be nonevaluative, supportive, timely, and specific” [checked in the PDF]

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>

A meta-analysis of 435 studies on feedback; effects vary a lot with what information the feedback carries.

“Overall results based on a random-effects model indicate a medium effect ($d = 0.48$) of feedback on student learning” [checked in the PDF]

9 Self-explanation

Explaining an idea in your own words builds understanding.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science*, 13(2), 145-182. https://doi.org/10.1207/s15516709cog1302_1

Students who explain worked examples to themselves learn with real understanding.

““Good” students learn with understanding: They generate many explanations” [checked in the published abstract]

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994a). Eliciting Self-Explanations Improves Understanding. *Cognitive Science*, 18(3), 439-477. https://doi.org/10.1207/s15516709cog1803_3

Prompting eighth graders to self-explain deepened their understanding of circulation.

“Generating explanations to oneself (self-explaining) facilitates that integration process” [checked in the scanned PDF]

10 Interleaving

Mixing different kinds of problems in practice helps learners tell them apart.

Bjork, R. A. (1994). Memory and Metamemory Considerations in the Training of Human Beings. *Metacognition: Knowing about Knowing* (MIT Press, book chapter). <https://doi.org/10.7551/mitpress/4561.003.0011>

Conditions that make practice harder can make learning last longer.

Kornell, N., & Bjork, R. A. (2008). Learning Concepts and Categories: Is Spacing the “Enemy of Induction”? *Psychological Science*, 19(6), 585-592. <https://doi.org/10.1111/j.1467-9280.2008.02127.x>

Interleaving artists’ paintings improved learning each style, though massing felt better.

“Surprisingly, induction profited from spacing, even though massing apparently created a sense of fluent learning” [checked in the PDF]

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, 35(6), 481-498. <https://doi.org/10.1007/s11251-007-9015-8>

Mixing different problem types in practice improved later math test scores.

11 AI and LLM tutors

Recent research on large language models as tutors: what works and how to measure it.

Chevalier, A., et al. (2024). Language Models as Science Tutors. *ICML*. <https://proceedings.mlr.press/v235/chevalier24a.html>

TutorEval tests whether language models can answer questions about long textbook chapters the way a science tutor would, and TutorChat trains for it.

“fine-tuning base models with existing dialogue datasets leads to poor performance on TutorEval” [checked in the published abstract]

Daheim, N., Macina, J., Kapur, M., Gurevych, I., & Sachan, M. (2024). Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors. *EMNLP*. <https://doi.org/10.18653/v1/2024.emnlp-main.478>

A dataset of student math solutions with the first wrong step marked by teachers; tutors that first find the error give better feedback.

“they struggle to precisely detect student’s errors and tailor their feedback to these errors” [checked in the PDF]

De Simone, M., Tiberti, F., Rodriguez, M. B., Manolio, F., Mosuro, W., & Dikoru, E. J. (2025). From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria. World Bank Policy Research Working Paper. <https://doi.org/10.1596/1813-9450-11125>

A randomized trial in Nigeria where secondary students used Microsoft Copilot (GPT-4) for English as an after-school tutor for six weeks.

“The effect on English, the main outcome of interest, was of 0.23 standard deviations” [checked in the PDF]

Dinucu-Jianu, D., Macina, J., Daheim, N., Hakimi, I., Gurevych, I., & Sachan, M. (2025). From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning. EMNLP. <https://doi.org/10.18653/v1/2025.emnlp-main.15>

Uses reinforcement learning with simulated students to turn an LLM into a tutor that guides instead of giving away answers.

“effective pedagogy which requires strategically withholding answers” [checked in the PDF]

Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024). Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI-Math Tutor in Ghana. AIED. https://doi.org/10.1007/978-3-031-64315-6_34

A field experiment with about 500 students in Ghana using Rori, a WhatsApp math tutor, one hour a week during study hall.

“an AI-powered math tutor accessible via WhatsApp” [checked in the PDF]

Jurenka, I., et al. (2024). Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv (Google DeepMind tech report). <https://arxiv.org/abs/2407.12687>

Educators and learners preferred a pedagogy-tuned Gemini tutor over prompting alone.

“LearnLM-Tutor is consistently preferred over a prompt tuned Gemini by educators and learners” [checked in the PDF]

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), 17458. <https://doi.org/10.1038/s41598-025-97652-6>

Students learned more in less time with an AI tutor than in class.

“students learn significantly more in less time when using the AI tutor” [checked in the PDF]

Kochmar, E., Maurya, K., Petukhova, K., Srivatsa, K. A., Tack, A., & Vasselli, J. (2025). Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors. BEA Workshop @ ACL. <https://doi.org/10.18653/v1/2025.bea-1.77>

A shared task where over 50 teams built systems to judge AI tutor replies on finding, locating and guiding a student’s mistake.

“mistake identification, precise location of the mistake, providing guidance, and feedback actionability” [checked in the PDF]

LearnLM Team, Google (2024). LearnLM: Improving Gemini for Learning. arXiv (Google tech report). <https://arxiv.org/abs/2412.16429>

Training on teaching instructions made Gemini the experts' pick for learning.

"tuned to present information by default, rather than engage users in service of learning" [checked in the PDF]

LearnLM Team, Google (2025). Evaluating Gemini in an arena for learning. arXiv (Google tech report). <https://arxiv.org/abs/2505.24477>

Educators ran blind, head-to-head tutoring comparisons of leading AI models.

"experts preferred Gemini 2.5 Pro in 73.2% of these match-ups" [checked in the PDF]

Liu, J., Huang, Z., Xiao, T., Sha, J., Wu, J., Liu, Q., Wang, S., & Chen, E. (2024). SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models. NeurIPS 2024. <https://doi.org/10.52202/079017-2721>

An LLM tuned on Socratic teaching dialogues out-taught GPT-4.

"actively engaging students in the thought process required for genuine problem-solving mastery" [checked in the PDF]

Macina, J., Daheim, N., Chowdhury, S., Sinha, T., Kapur, M., Gurevych, I., & Sachan, M. (2023a). MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. Findings of EMNLP 2023. <https://doi.org/10.18653/v1/2023.findings-emnlp.372>

LLMs solve math well but reveal solutions too early when tutoring.

"While models like GPT-3 are good problem solvers, they fail at tutoring" [checked in the PDF]

Macina, J., Daheim, N., Hakimi, I., Kapur, M., Gurevych, I., & Sachan, M. (2025). MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors. EMNLP. <https://doi.org/10.18653/v1/2025.emnlp-main.11>

An open benchmark for how well LLMs teach math, with a reward model that tells expert from novice tutor replies; solving skill did not translate into good teaching.

"subject expertise, indicated by solving ability, does not immediately translate to good teaching" [checked in the PDF]

Macina, J., Daheim, N., Wang, L., Sinha, T., Kapur, M., Gurevych, I., & Sachan, M. (2023b). Opportunities and Challenges in Neural Dialog Tutoring. EACL. <https://doi.org/10.18653/v1/2023.eacl-main.173>

Tests generative language models on two language-learning tutoring datasets and maps where they still fall short of human tutors.

"dialog tutoring has largely remained unaffected by these advances" [checked in the PDF]

Maurya, K. K., Srivatsa, K. A., Petukhova, K., & Kochmar, E. (2025). Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors. NAACL 2025. <https://doi.org/10.18653/v1/2025.naacl-long.57>

An eight-dimension benchmark for judging how well LLMs actually tutor.

"highlighting which LLMs are good tutors and which ones are more suitable as question-answering systems" [checked in the PDF]

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. PLOS ONE, 19(5), e0304013. <https://doi.org/10.1371/journal.pone.0304013>

ChatGPT-written hints produced learning gains comparable to human tutor hints.

"only the ChatGPT condition produces statistically significant learning gains compared to a no-help control" [checked in the PDF]

Scarlato, A., Liu, N., Lee, J., Baraniuk, R., & Lan, A. (2025). Training LLM-Based Tutors to Improve Student Learning Outcomes in Dialogues. AIED. https://doi.org/10.1007/978-3-031-98414-3_18

Trains an LLM tutor to pick replies that a model predicts will help the student answer correctly, while keeping the replies pedagogically sound.

“they are not trained to maximize student learning throughout the course of a dialogue” [checked in the PDF]

Schmucker, R., Xia, M., Azaria, A., & Mitchell, T. (2024). Ruffle&Riley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System. AIED. https://doi.org/10.1007/978-3-031-64302-6_6

A conversational tutor where two LLM agents, a student and a professor, let the learner teach and be taught from a lesson text.

“orchestration in a learning-by-teaching format via two LLM-based agents” [checked in the PDF]

Sonkar, S., Ni, K., Chaudhary, S., & Baraniuk, R. (2024). Pedagogical Alignment of Large Language Models. Findings of EMNLP. <https://doi.org/10.18653/v1/2024.findings-emnlp.797>

Trains LLMs with preference learning to guide students through problems rather than reply with the answer at once.

“often provide immediate answers rather than guiding students through the problemsolving process” [checked in the PDF]

Tack, A., & Piech, C. (2022). The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. EDM. <https://educationaldatamining.org/edm2022/proceedings/2022.EDM-short-papers.54/>

Compares Blender and GPT-3 replies with real teachers' replies in the same dialogues on speaking like a teacher, understanding a student and helping a student.

“speak like a teacher, understand a student, help a student” [checked in the PDF]

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024b). Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. arXiv preprint (v1). <https://arxiv.org/abs/2410.03017v1>

AI guidance for human tutors raised student topic mastery by 4 points.

“less likely to give away the answer to the student” [checked in the PDF]

Wang, R., Zhang, Q., Robinson, C., Loeb, S., & Demszky, D. (2024a). Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes. NAACL 2024. <https://doi.org/10.18653/v1/2024.naacl-long.120>

Adding expert remediation decisions made LLM tutor replies far more preferred.

“the expert’s decision-making model is critical for LLMs to close the gap” [checked in the PDF]

12 How tutors talk

Datasets and studies of real tutoring and classroom conversation.

Boyer, K. E., et al. (2010). Dialogue Act Modeling in a Complex Task-Oriented Domain. SIGDIAL. <https://aclanthology.org/W10-4356/>

A classifier for dialogue acts in computer-programming tutoring that uses both the chat and the student’s task actions.

“The task-oriented domain involves tutoring in computer programming exercises” [checked in the PDF]

Caines, A., Yannakoudakis, H., Edmondson, H., Allen, H., Pérez-Paredes, P., Byrne, B., & Buttery, P. (2020). The Teacher-Student Chatroom Corpus. NLP4CALL. <https://doi.org/10.3384/ecp2017510>

A corpus of one-to-one online chat lessons between English teachers and learners, with scaffolding and correction annotated.

“the teacher was able to focus exclusively on the linguistic abilities and errors of the student” [checked in the PDF]

Demszky, D., & Hill, H. (2023). The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. BEA Workshop, ACL 2023. <https://doi.org/10.18653/v1/2023.bea-1.44>

Dialogic teacher moves track with better observation scores and learning outcomes.

“these moves are correlated with better classroom observation scores and learning outcomes” [checked in the PDF]

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). Can Automated Feedback Improve Teachers’ Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in a Large-Scale Online Course. Educational Evaluation and Policy Analysis, 46(3), 483-505. <https://doi.org/10.3102/01623737231169270>

A randomized trial with 1,136 instructors: automated feedback on uptake of student ideas made instructors build on students more.

“We find that M-Powering Teachers improves instructors’ uptake of student contributions by 13%” [checked in the PDF]

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. ACL. <https://doi.org/10.18653/v1/2021.acl-long.130>

Measures how teachers build on what students say, and links that uptake to outcomes.

“teachers’ uptake of student contributions has been linked to higher student achievement” [checked in the PDF]

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. Applied Cognitive Psychology, 9(6), 495-522. <https://doi.org/10.1002/acp.2350090604>

Everyday tutoring works through joint problem solving, questions and explanation with examples.

“Naturalistic one-to-one tutoring is more effective than traditional classroom teaching methods” [checked in the published abstract]

Kakarla, S., et al. (2024). Using Large Language Models to Assess Tutors’ Performance in Reacting to Students Making Math Errors. AAAI Workshop on AI for Education. <https://arxiv.org/abs/2401.03238>

Checks whether GPT-3.5 and GPT-4 can grade how real tutors respond when a student makes a math error.

“Rather than drawing direct attention to the error, tutors should guide the students to identify and correct their mistakes on their own” [checked in the PDF]

Lin, J., Singh, S., Sha, L., Tan, W., Lang, D., Gašević, D., & Chen, G. (2022). Is it a good move? Mining effective tutoring strategies from human-human tutorial dialogues. Future Generation Computer Systems, 127, 194-207. <https://doi.org/10.1016/j.future.2021.09.001>

Mines human tutoring dialogues to find which tutor moves go with better learning.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative Discourse Idealized and Realized: Accountable Talk in the Classroom and in Civic Life. *Studies in Philosophy and Education*, 27(4), 283-297. <https://doi.org/10.1007/s11217-007-9071-1>

Describes Accountable Talk, classroom discussion practices that hold students to reasoning, evidence and each other's ideas.

"Classroom discussion practices that can lead to reasoned participation by all students are presented and described by the authors" [checked in the PDF]

Miller, P., & DiCerbo, K. (2024). LLM Based Math Tutoring: Challenges and Dataset. EdArXiv. <https://doi.org/10.35542/osf.io/5zwv3>

A dataset for checking whether LLM tutors get the math right while tutoring, plus a look at the kinds of student interactions.

"evaluating their performance in real-time math tutoring scenarios presents distinct challenges" [checked in the PDF]

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding Tutor Learning: Knowledge-Building and Knowledge-Telling in Peer Tutors' Explanations and Questions. *Review of Educational Research*, 77(4), 534-574. <https://doi.org/10.3102/0034654307309920>

Peer tutors mostly tell knowledge instead of building it with students.

"Peer tutors, even when trained, focus more on delivering knowledge rather than developing it" [checked in the PDF]

Stasaski, K., Kao, K., & Hearst, M. A. (2020). CIMA: A Large Open Access Dialogue Dataset for Tutoring. BEA Workshop, ACL 2020. <https://doi.org/10.18653/v1/2020.bea-1.5>

A tutoring dialogue dataset where tutors choose between hints and questions.

"One-to-one tutoring is often an effective means to help students learn" [checked in the PDF]

Suresh, A., et al. (2022). The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. LREC. <https://aclanthology.org/2022.lrec-1.497/>

567 K-12 math lesson transcripts labelled for teacher and student talk moves from accountable talk theory.

"sustained classroom discourse is a critical component of equitable, engaging, and rich learning environments" [checked in the PDF]

Wang, R. E., & Demszky, D. (2023). Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. BEA Workshop @ ACL. <https://aclanthology.org/2023.bea-1.53/>

Tests whether ChatGPT can score classroom transcripts and give teachers useful coaching feedback.

"We explore whether generative AI could become a cost-effective complement to expert feedback by serving as an automated teacher coach" [checked in the PDF]

Wang, R., & Demszky, D. (2024). Edu-ConvoKit: An Open-Source Library for Education Conversation Data. NAACL (Demo). <https://doi.org/10.18653/v1/2024.naacl-demo.6>

An open-source library for cleaning, annotating and analyzing conversation data from classrooms and tutoring.

"an open-source library designed to handle pre-processing, annotation and analysis of conversation data in education" [checked in the PDF]

Wang, Z., et al. (2020). Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge. arXiv (NeurIPS 2020 Competition). <https://arxiv.org/abs/2007.12061>

The NeurIPS 2020 challenge on Eedi diagnostic questions, where wrong answer choices reveal specific misconceptions.

“the answers that the students give to these questions reveal key information about the specific nature of misconceptions” [checked in the PDF]

13 Intelligent tutoring systems

Forty years of computer tutors, knowledge tracing and what they taught us.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., & Heo, J. (2020). EdNet: A Large-Scale Hierarchical Dataset in Education. AIED. https://doi.org/10.1007/978-3-030-52240-7_13

A large public dataset of student interactions from an English test-prep tutor, for knowledge tracing research.

“there is no public large-scale benchmark dataset” [checked in the PDF]

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modelling and User-Adapted Interaction, 4(4), 253-278. <https://doi.org/10.1007/bf01099821>

Bayesian knowledge tracing: a tutor estimates, skill by skill, how likely the student is to know it, and updates after every answer.

Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-Aware Attentive Knowledge Tracing. KDD. <https://doi.org/10.1145/3394486.3403282>

Attentive knowledge tracing with interpretable parts from cognitive and psychometric models, including decay of what was learned.

“predicting future learner performance given their past performance in educational applications” [checked in the PDF]

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. Cognitive Systems Research, 1(1), 35-51. [https://doi.org/10.1016/S1389-0417\(99\)00005-4](https://doi.org/10.1016/S1389-0417(99)00005-4)

A computer tutor that holds natural-language dialogue like a human tutor.

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. International Journal of Artificial Intelligence in Education, 24(4), 470-497. <https://doi.org/10.1007/s40593-014-0024-x>

ASSISTments, a free platform where teachers assign problems with hints and students get immediate feedback, also used for research.

“The system gives immediate feedback to students while they are working” [checked in the PDF]

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes To School in the Big City. International Journal of Artificial Intelligence in Education. https://www.researchgate.net/publication/2834538_Intelligent_Tutoring_Goes_To_School_in_the_Big_City

An algebra tutor in Pittsburgh city schools raised standardized test scores.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. Cognitive Science, 36(5), 757-798. <https://doi.org/10.1111/j.1551-6709.2012.01245.x>

The Knowledge-Learning-Instruction framework: match the kind of instruction to the kind of knowledge and learning involved.

“we describe the Knowledge-Learning-Instruction (KLI) framework” [checked in the PDF]

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, 86(1), 42-78. <https://doi.org/10.3102/0034654315581420>

A meta-analysis of 50 controlled evaluations of intelligent tutoring systems.

“The median effect of intelligent tutoring in the 50 evaluations was to raise test scores 0.66 standard deviations over conventional levels” [checked in the published abstract]

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901-918. <https://doi.org/10.1037/a0037123>

A meta-analysis of 107 effect sizes comparing learning with intelligent tutoring systems to other kinds of instruction.

“107 effect sizes involving 14,321 participants were extracted and analyzed” [checked in the PDF]

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, 24(4), 427-469. <https://doi.org/10.1007/s40593-014-0029-5>

Conversational tutoring produced learning gains in physics, computer literacy and more.

“AutoTutor is a natural language tutoring system that has produced learning gains across multiple domains” [checked in the PDF]

Pandey, S., & Karypis, G. (2019). A Self-Attentive model for Knowledge Tracing. EDM. <https://arxiv.org/abs/1907.06837>

A self-attention model for knowledge tracing that handles sparse student data better than earlier neural models.

“Knowledge tracing is the task of modeling each student’s mastery of knowledge concepts” [checked in the PDF]

Pavlik, P. I., Jr., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. AIED. <https://doi.org/10.3233/978-1-60750-028-5-531>

Performance Factors Analysis, a simple alternative to knowledge tracing that counts a student’s past successes and failures per skill.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep Knowledge Tracing. NeurIPS. <https://papers.nips.cc/paper/5654-deep-knowledge-tracing>

Uses recurrent neural networks to model what a student knows from their sequence of answers.

“we explore the utility of using Recurrent Neural Networks (RNNs) to model student learning” [checked in the PDF]

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online Mathematics Homework Increases Student Achievement. *AERA Open*, 2(4), 2332858416673968. <https://doi.org/10.1177/2332858416673968>

A randomized trial with 2,850 seventh graders in Maine: online homework with timely feedback and hints raised math scores.

“Results showed that the intervention significantly increased student scores on an end-of-the-year standardized mathematics assessment” [checked in the PDF]

Zhang, J., Shi, X., King, I., & Yeung, D.-Y. (2017). Dynamic Key-Value Memory Networks for Knowledge Tracing. WWW. <https://doi.org/10.1145/3038912.3052580>

A memory network for knowledge tracing that outputs a mastery level for each concept.

“directly output a student’s mastery level of each concept” [checked in the PDF]

14 Explanations, analogies and misconceptions

How explanations and analogies work, and why some ideas are hard to change.

Chi, M. T. H., Slotta, J. D., & De Leeuw, N. (1994b). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction*, 4(1), 27-43. [https://doi.org/10.1016/0959-4752\(94\)90017-5](https://doi.org/10.1016/0959-4752(94)90017-5)

Argues that some science ideas are hard because students treat processes, like heat or current, as if they were things.

diSessa, A. (1993). Toward an Epistemology of Physics. *Ethics & Behavior*, 10(2), 105-225. https://doi.org/10.1207/s1532690xci1002&3_2

Describes intuitive physics as many small pieces of knowledge (p-prims) that learning has to reorganize.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170. [https://doi.org/10.1016/S0364-0213\(83\)80009-3](https://doi.org/10.1016/S0364-0213(83)80009-3)

Analogies work by mapping relations between domains, not surface features.

“Relations between objects, rather than attributes of objects, are mapped from base to target” [checked in the scanned PDF]

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2), 393-408. <https://doi.org/10.1037/0022-0663.95.2.393>

Comparing two cases side by side helped learners pull out the shared idea and use it later.

“Experiment 2 showed a marked advantage for comparing two cases over studying the 2 cases separately” [checked in the PDF]

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12(3), 306-355. [https://doi.org/10.1016/0010-0285\(80\)90013-4](https://doi.org/10.1016/0010-0285(80)90013-4)

A solved story helps crack an analogous problem, mostly when learners are hinted.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141-158. <https://doi.org/10.1119/1.2343497>

The Force Concept Inventory, a test that shows how common sense beliefs about force survive regular physics teaching.

“Instruction that does not take them into account is almost totally ineffective, at least for the majority of students” [checked in the PDF]

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, 126, 198-212. <https://doi.org/10.1016/j.jecp.2014.03.001>

Asking young children to explain a toy helped them learn how it works more than only watching it.

“explanation promotes causal learning” [checked in the PDF]

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464-470. <https://doi.org/10.1016/j.tics.2006.08.004>

A review of how explanations shape learning and generalization by fitting new facts into what we already believe.

“explanations accommodate novel information in the context of prior beliefs, and do so in a way that fosters generalization” [checked in the PDF]

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2), 211-227. <https://doi.org/10.1002/sce.3730660207>

A classic theory of conceptual change: learners give up an idea only when it stops working and a new one makes sense.

“learning is the result of the interaction between what the student is taught and his current ideas or concepts” [checked in the PDF]

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive Supports for Analogies in the Mathematics Classroom. *Science*, 316(5828), 1128-1129. <https://doi.org/10.1126/science.1142103>

Compares how teachers in Hong Kong, Japan and the US support analogies in math lessons.

“Variations in the effective use of analogies in math instruction across countries may contribute to performance differences in the TIMSS studies” [checked in the PDF]

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26(5), 521-562. https://doi.org/10.1207/s15516709cog2605_1

People think they understand how things work much better than they do, until they try to explain it.

“People feel they understand complex phenomena with far greater precision, coherence, and depth than they really do” [checked in the PDF]

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24(4), 535-585. [https://doi.org/10.1016/0010-0285\(92\)90018-w](https://doi.org/10.1016/0010-0285(92)90018-w)

Children’s ideas about the shape of the Earth show consistent mental models that mix what they are told with what they see.

“many children said that the earth is round but also stated that it has an end or edge from which people could fall” [checked in the PDF]

Wittwer, J., & Renkl, A. (2008). Why Instructional Explanations Often Do Not Work: A Framework for Understanding the Effectiveness of Instructional Explanations. *Educational Psychologist*, 43(1), 49-64. <https://doi.org/10.1080/00461520701756420>

A framework for why instructional explanations often fail and how to design ones that work.

“research shows that they often do not contribute to learning” [checked in the published abstract]

15 Memory, motivation and metacognition

The limits of working memory, knowing what you know, and wanting to learn.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. *Psychology of Learning and Motivation*. [https://doi.org/10.1016/S0079-7421\(08\)60452-1](https://doi.org/10.1016/S0079-7421(08)60452-1)

The working memory model: a small active workspace for thinking, not just a short-term store.

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223-241. <https://doi.org/10.1016/j.ijhcs.2009.12.003>

Tracks learners' feelings in three computer learning environments; boredom was more persistent and more harmful than frustration.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention. *Child Development*, 78(1), 246-263. <https://doi.org/10.1111/j.1467-8624.2007.00995.x>

Seventh graders who believed intelligence can grow did better in math over junior high.

"the belief that intelligence is malleable (incremental theory) predicted an upward trajectory in grades over the two years of junior high school" [checked in the PDF]

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87-114. <https://doi.org/10.1017/s0140525x01003922>

Working memory holds about three to five chunks, fewer than the famous seven.

"it is only three to five chunks" [checked in the PDF]

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671-684. [https://doi.org/10.1016/s0022-5371\(72\)80001-x](https://doi.org/10.1016/s0022-5371(72)80001-x)

Memory depends on how deeply information is processed, not on which store it sits in.

"An alternative framework for human memory research is then outlined in terms of depth or levels of processing" [checked in the PDF]

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906-911. <https://doi.org/10.1037/0003-066x.34.10.906>

The paper that named metacognition: knowing and monitoring your own thinking.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. <https://doi.org/10.1037/0022-3514.77.6.1121>

People with the least skill tend to overrate their skill the most, because judging it needs the same skill.

"participants scoring in the bottom quartile on tests of humor, grammar, and logic grossly overestimated their test performance and ability" [checked in the PDF]

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63(2), 81-97. <https://doi.org/10.1037/h0043158>

The classic paper on the limits of short-term memory and on chunking information to stretch them.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement. *Psychological Science*, 26(6), 784-793. <https://doi.org/10.1177/0956797615571017>

Short online growth mindset and purpose lessons raised grades for students at risk of dropping out.

"delivered brief growth-mind-set and sense-of-purpose interventions through online modules to 1,594 students in 13 geographically diverse high schools" [checked in the PDF]

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33-40. <https://doi.org/10.1037/0022-0663.82.1.33>

Self-efficacy, self-regulation and strategy use predicted classroom performance for 173 seventh graders.
"Self-efficacy and intrinsic value were positively related to cognitive engagement and performance" [checked in the PDF]

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68-78. <https://doi.org/10.1037/0003-066x.55.1.68>

Self-determination theory: people are most motivated when they feel competent, autonomous and related to others.

"The findings have led to the postulate of three innate psychological needs" [checked in the PDF]

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To What Extent and Under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses. *Psychological Science*, 29(4), 549-571. <https://doi.org/10.1177/0956797617739704>

Two meta-analyses found weak overall links between growth mindset and achievement, stronger for some groups.

"Overall effects were weak" [checked in the PDF]

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., .. Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774), 364-369. <https://doi.org/10.1038/s41586-019-1466-y>

A national randomized experiment: a short online growth mindset lesson raised grades for lower-achieving students.

"improved grades among lower-achieving students and increased overall enrolment to advanced mathematics courses" [checked in the PDF]

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, 41(2), 64-70. https://doi.org/10.1207/s15430421tip4102_2

An overview of self-regulated learning: setting goals, using strategies, and checking your own progress.

16 Active learning

Why doing beats listening, and why effort can feel like not learning.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49(4), 219-243. <https://doi.org/10.1080/00461520.2014.965823>

Learning rises from passive to active to constructive to interactive engagement.

"from passive to active to constructive to interactive, their learning will increase" [checked in the PDF]

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251-19257. <https://doi.org/10.1073/pnas.1821936116>

Students in active classes learned more but felt they learned less.

“when students experience the increased cognitive effort associated with active learning, they initially take that effort to signify poorer learning” [checked in the PDF]

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410-8415. <https://doi.org/10.1073/pnas.1319030111>

Across 225 studies, active learning raised exam scores and cut failure rates.

“student performance on examinations and concept inventories increased by 0.47 SDs under active learning” [checked in the PDF]

Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, 26(3), 379-424. <https://doi.org/10.1080/07370000802212669>

Students who struggled first with hard problems later transferred knowledge better.

“suggesting a latent productivity in what initially seemed to be failure” [checked in the PDF]

National Research Council; Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.) (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. National Academies Press (book). <https://doi.org/10.17226/9853>

Synthesis of learning science: how experts differ from novices and how transfer works.

Nuthall, G. (2007). *The Hidden Lives of Learners*. NZCER Press (book). <https://www.nzcer.org.nz/nzcerpress/hidden-lives-learners>

Decades of classroom recordings show peers and private worlds shape what students learn.

Schwartz, D. L., & Bransford, J. D. (1998). A Time For Telling. *Cognition and Instruction*, 16(4), 475-522. https://doi.org/10.1207/s1532690xc1604_4

Comparing contrasting cases first prepares students to learn from a later lecture.