UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Training Students to Recognize Comparisons in Science Texts

Permalink

https://escholarship.org/uc/item/02z8t65q

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Mo, Fanyi
Zheng, Yinyuan
Holt, Sebastian
et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Training Students to Recognize Comparisons in Science Texts

Fanyi Mo (fanyimo2026@u.northwestern.edu)
Department of Psychology, Northwestern University

Yinyuan Zheng (zhengyinyuan@u.northwestern.edu)
Department of Psychology, Northwestern University

Sebastian Holt (sebastian.holt@northwestern.edu)
Department of Psychology, Northwestern University

Olivia Harbison (oliviaharbison2024@u.northwestern.edu)
Department of Psychology, Northwestern University

Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, Northwestern University

Bryan Matlen (bmatlen@gmail.com)
Science, Technology, Engineering & Mathematics (STEM), WestEd

Benjamin Jee (bjee@worcester.edu)
Department of Psychology, Worcester State University

Florencia Anggoro (fanggoro@holycross.edu)
Department of Psychology, College of the Holy Cross

Abstract

Science texts frequently use comparisons to communicate concepts, based on research that has established comparison's benefits in learning. However, applying these findings to student learning relies on a critical premise yet to be examined—namely, that students recognize comparisons in their readings. The current work aims to examine (1) students' ability to recognize comparisons in science texts, and (2) whether this ability can be trained. In two experiments across two different institutions, college students were asked to mark all the comparisons in expository science texts, with some receiving instructions and practice on identifying comparisons. Without training, students missed many of the intended comparisons. However, targeted training reliably improved comparison detection. The findings showed that detecting comparisons is a bottleneck in learning from science texts, but that this ability can be improved with brief training.

Keywords: Analogical comparison; comparison spotting; cognitive training; STEM learning

Introduction

Comparison is a mainstay in science instruction. A wealth of research suggests that making comparisons—especially analogical comparisons—improves relational learning and generalization in the lab and in the classroom (e.g., Alfieri et al., 2013; Goldwater & Schalk, 2016; Richland & Simms, 2015; Vendetti et al., 2015). For example, a meta-analysis of 57 learning studies in both laboratory and classroom settings (Alfieri et al., 2013) revealed that comparison activities with analogous cases commonly led to greater learning outcomes than did traditional instruction, including other forms of case-based instruction. The benefits of comparison-based instruction have been demonstrated across age and content area, including mathematics (Hattikudur & Alibali, 2010; Mix et al., 2024; Richland & McDonough, 2010; Rittle-Johnson & Star, 2007, 2009; Sullivan & Barner, 2014), biology (Emmons et al., 2018; Kurtz & Gentner, 2013; Legare et al., 2018; Seiler & Huggins, 2018; Shtulman et al., 2019), geoscience (Jee et al., 2013; Resnick et al., 2017), and physics and engineering (Clement, 1993; Dixon & Dohn, 2003; Gentner et al., 2016; Nokes-Malach et al., 2013).

Consistent with researchers' recommendations, comparison is often used in instruction to explain unfamiliar domains (Brown & Salter, 2010; Jee et al., 2010; Jee & Anggoro, 2019, 2021; Keri & Elbatarny, 2021; Orgill & Brodner, 2006). For example, in analyzing eight college-level biochemistry textbooks, Orgill (2006) found 158 analogical comparisons, yielding an average of approximately 20 per book. Thus, the ability to engage productively with analogical comparisons is important for STEM mastery and scientific literacy.

Because of its importance in learning, a large body of research has investigated factors that support students' processing of analogical comparisons, including conceptual alignment (Christie et al., 2010; Goldwater & Gentner, 2015; Kurtz et al., 2001; Mason, 1994; Orton et al., 2012), visual alignment (Begolli & Richland, 2016; Matlen et al., 2022; Zheng et al., 2022; Simms et al., 2023), language (Christie

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

& Gentner, 2014; Son & Goldstone, 2005; Zheng et al., 2025), and gestural support (Alibali et al., 2014; Richland, 2015; Wakefield & Goldin-Meadow, 2019; Yuan et al., 2017). However, applying these findings to student learning relies on a critical premise—namely, that students recognize and process analogical comparisons as they read. When directly asked, most students show clear appreciation of analogies in science learning (Keri & Elbatarny, 2021). Nevertheless, there is evidence that students vary in the degree to which they take advantage of available comparisons (e.g., Alexander et al., 1989; Murphy et al., 2021; Richland et al., 2006; Rottman et al., 2012). For example, Zepeda and Nokes-Malach (2021) surveyed college students about their study strategies before each of three noncumulative exams. Although the use of analogy was correlated with high test scores, students reported using analogical comparisons only 7% of the time on Exam 1, 11% on Exam 2, and 6% on Exam 3. The overall average number of times a student reported using analogical comparison across the exams was only 0.24 (out of a possible three total exams). This suggests that analogical comparison is not a salient part of students' normal study habits.

Why do students neglect comparisons? In some cases, they may simply not take time to think through a comparison. But it is possible that students sometimes fail even to *notice* an intended comparison. This may be particularly likely with analogical comparisons, in which the elements being compared lack obvious surface similarity. The goals of the current work are (1) to assess college students' ability to identify intended comparisons in science text; (2) to test whether this ability can be improved; and (3) to assess whether these patterns would vary across SES and individual differences.

Cues to comparison. One potential challenge in detecting comparisons in texts is the variety of linguistic cues used to signal them. Some cues (such as "similar", "like", or "likewise") are likely to be obvious even for novices, but others (such as "equivalent of" or "correspondingly") may be more challenging. Students may also miss signals inviting contrastive comparison, such as "in contrast." Another challenge might be cues such as "while" or "however," which can signal comparison but also have other functions. For example, in (a) below, *while* signals comparison, but in (b), it indicates simultaneous activity. In addition to overt cues, comparisons often have parallel content between the two things being compared, as can be seen in (a) and (b). Indeed, in some cases, no overt cues are used: the intent to compare is signaled chiefly by parallel semantic structure, as in (c). We suspect that these comparisons are particularly likely to be missed.

(a) The hummingbird flaps its wings rapidly to hover, while the albatross uses long wings to glide effortlessly.
(b) The hummingbird sipped nectar from the flower while the crickets chirped in the bushes.
(c) For most fish, the brain is located in the head; for octopuses, it is distributed throughout the tentacles and head.

Experiment 1

The first goal of E1 was to explore whether college students do, in fact, miss intended comparisons in science texts. To this end, we included a Baseline condition in which students were given only one task: to highlight all the comparisons in a series of short texts. If under these focused conditions, students fail to detect some intended comparisons, this will flag comparison-detection as a likely point of failure in students' learning.

Our second goal was to discover whether we can improve students' ability to detect comparisons in texts. To address this question, we ran a Training condition in which students were given instructions and practice in identifying comparisons. After training, these students completed a comparison-detection task like that given to the Baseline group.

In addition to the training factor, we also varied Institution. The subjects were students from two universities: a highly selective private institution and a less selective public institution (here labeled A and B, respectively). As Table 1 shows, their demographic characteristics vary considerably. Thus, this contrast allows us to assess the generality of our prospective findings on students' propensity to detect comparisons across a range of institutional backgrounds, as well as to discover whether our training plan can succeed across different populations. To this end, we also included a short form of Raven's Progressive Matrix (RPM), a standard assessment for fluid intelligence (Carpenter et al., 1990; Gray & Holyoak, 2020).

The basic plan was to give subjects a series of texts involving comparison, and to ask students either to choose which of two texts expresses a comparison, or to highlight all the comparisons in each text. The Training group was given lessons on comparison and feedback on practice items. The Baseline group was also instructed to highlight comparisons, but was not given training and feedback. Thus, the Baseline group serves to assess whether students can spot comparisons when directly told to do so. The Training group assesses whether this ability improves with training.

The procedure was identical at the two institutions. In both cases, the Training group received a lesson on the importance of comparison cues. Students were instructed on the cues that commonly signal comparisons in science texts. Then, they received a series of practice items to distinguish comparison statements from non-comparisons using these cues. The activity ended by asking students to identify comparisons in short science passages. Feedback and explanations were provided throughout the instruction phase.

Table 1: Institution Differences.

University	A	B
Median Family Income	$171,200	$98,600
% Receiving Financial Aid	54%	95%
% First-generation Students	15%	47%
% Scoring Above 1400 on SAT	92%	2%

In contrast, the Baseline (No training) subjects received a minimal introduction and were told that their task was to identify comparisons in texts and given a practice item. After training, the two groups received a test on the same materials, in which they were asked to spot comparisons in unfamiliar science texts. Neither group received feedback in the Test phase.

We had three predictions: Prediction (1) was that college students would miss some intended comparisons in texts, *even when being directly told to mark them*. Prediction (2) was that participants at University A would outperform those at University B. This prediction is based on the general finding that SES and SAT scores tend to predict academic skills (Burton & Ramist, 2001; Caro et al., 2009). Prediction (3) was that students in the Training group would be better at recognizing comparisons than would the Baseline group. We also wanted to discover whether the training effect (if any) would differ between the two institutions and if so, how.

Methods

Participants. 84 University A students and 84 University B students participated in the study. Among participants from University A, 25% identified as male, 71% as female, and 4% as other; 38% were White, 24% Asian, 14% Multiracial, 11% Latino/Latina, and 10% Black. For University B, 23% identified as male, 76% as female, and 1% other/unknown; subjects were drawn from a participant pool in which 62% were White, 4% Asian, 3% Multiracial, 21% Latino/Latina, 8% Black, and 2% other/unknown. The participants were randomly assigned to two conditions: Training (N_A= 40; N_B= 44) and Baseline (N_A= 44; N_B= 40).

Materials and design. All instructions and materials were presented on Qualtrics. We chose passages containing many analogical comparisons, taken from college-level popular science articles and books. Participants used the highlighting function to mark comparisons in the texts. The design was a 2 x 2 between-subject design: Condition (Training vs. Baseline) by Institution (A vs. B).

Learning Phase.

Training group. The participants were told that the study aimed to improve students' ability to notice comparisons in texts.

Choice questions. Next, they were given a forced-choice task in which they had to click on which of two sentences expressed a comparison. There were three questions, all using highly familiar comparison cues, such as *like* or *as.* Next, participants were told that some comparison cues are more challenging, as they do not always signal comparison (e.g., *yet, however*, and, *whereas).* Participants then answered four forced-choice questions involving subtle comparison cues. They received feedback on all seven forced-choice examples.

Comparison detection. Participants were then introduced to their main task. They were given texts and told to mark all the comparisons they could find. They were shown how to highlight comparisons in Qualtrics. After performing this task for two short passages[1], they were given feedback in the form of an expert's highlighting of the same passages.

Baseline (No-training) group. Participants were told that comparisons are often used in explanation and that their task would be to mark comparisons in texts. They were then given an example and shown how to highlight it. Next, they performed the same Comparison detection task as the Training group, but without feedback.

Test Phase. After the above training, both groups went on to the test phase, in which they performed identical tasks. No feedback was given to either group.

Choice questions. Both groups answered four forced-choice questions, in which they were asked to choose which of two sentences expressed a comparison. The four pairs were matched in comparison cues and overall sentence structure.

Comparison detection. Next, subjects were asked to highlight all comparisons in two longer passages on biological topics (roughly 400 words). Each passage contained eight possible comparisons.

Ravens' Matrix. Participants then completed the short form of Raven's Progressive Matrices test (RPM).

Analysis Plan. There were two independent variables: institution (A vs. B) and condition (Training vs. Baseline), and two dependent variables of interest: detection accuracy and forced-choice accuracy. For all passages, we coded whether participants' highlights matched the correct comparisons (determined a priori). Detection accuracy was calculated as the proportion of comparisons detected[2]. We fit mixed-effects logistic regression models to the trial-level accuracy data. Fixed effects included passage, condition, institution, and condition-institution interaction. All models included random intercepts for participants and for items. In a further analysis, we included the Raven's score as a covariate in the mixed-effects model to control for variation in fluid intelligence.

[1] The first of these short passages proved to be too difficult (95% of the participants missed one of the comparisons). Therefore, we include only the second passage in the analyses below.

[2] Three coders each coded about half of the responses. The interrater reliability was 87%.

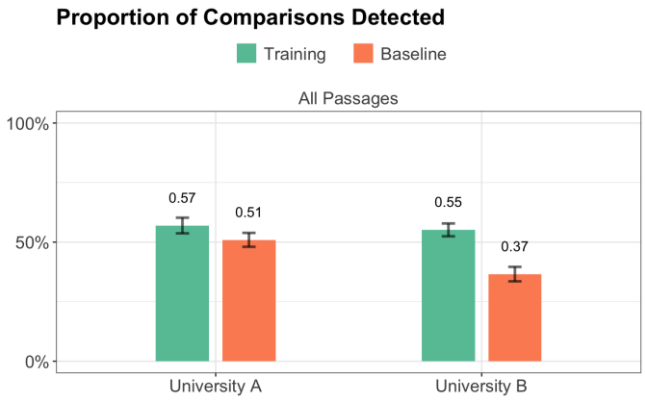


Figure 1: Proportion of comparisons detected by Training group vs. Baseline group (error bars represent standard errors).

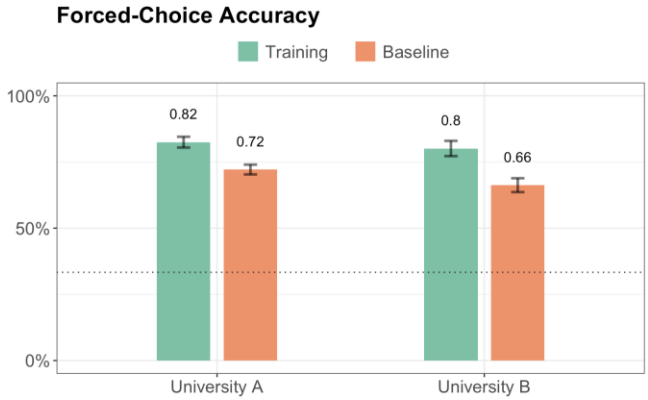


Figure 2: Forced-choice performance (error bars represent standard errors; the dotted bar is chance level).

Results

Baseline Performance. Consistent with Prediction 1, the results bore out our concern that even college students often fail to notice comparisons in texts that they read. Across both institutions, students marked less than half of the comparisons present in the passages (45%) (Figure 1). There was also an effect of institution (Prediction 2): students at University B were more likely to miss comparisons than those at University A: proportion detected: 54% in A and 46% in B, $t = 2.39$, $p = .02$, $d = 0.37$, averaged across conditions.

Effects of Training. As predicted, the Training group performed better than the Baseline group across all measures (Prediction 3).

Comparison detection. A mixed-effects model over the proportion of comparisons revealed two main effects: condition and institution. The Training group performed significantly better than the Baseline group ($\beta = 1.07$, $SE = 0.24$, $z = 4.40$, $p < .001$). Further, we found that participants from University A outperformed those from University B ($\beta = 0.82$, $SE = 0.24$, $z = 3.37$, $p < .001$). The interaction effect was also significant ($\beta = -0.72$, $SE = 0.34$, $z = -2.10$, $p = .036$), reflecting a larger training effect at University B than at University A. T-tests showed that for University B, the training group was more likely than the Baseline group to identify comparisons, $t = 4.56$, $p < .001$, $d = 1$; for University A, training did not have a significant effect, $t = 1.36$, $p = .18$, $d = 0.30$. The Passage effects were not significant.

Choice Question Accuracy. A mixed-effects model over the number of correct responses revealed only a significant effect of condition, $\beta = 1.31$, $SE = 0.35$, $z = 3.78$, $p < .001$. Neither the main effect of institution, $\beta = 0.51$, $SE = 0.32$, $z = 1.59$, $p = 0.11$, nor the interaction, $\beta = -0.24$, $SE = 0.48$, $z = -0.50$, $p = .61$, was significant.

Including Raven's Score. The Raven's Matrix score was significantly correlated with Institution, $r = .51$, $p < .001$, but not with Condition, $r = .05$, $p = .56$. University A participants were 79% accurate on the Raven's task, while University B participants were 55% accurate.

We next reran the two mixed-effects models above, including Raven's score as a covariate. Raven's score was found to be significant both for comparison detection accuracy ($\beta = 0.28$, $SE = 0.10$, $z = 2.93$, $p = .003$) and for choice question accuracy ($\beta = 0.49$, $z = 3.38$, $SE = 0.14$, $p < .00$). The important question here is whether including the Raven's score changed the pattern. The answer is no. The pattern of results remains the same: the training effect remains significant for both comparison detection and the choice task.

Table 2: Mean detection rates for a sample of cues across institutions (University A vs University B) and training conditions (Trained (T) vs Baseline (B)), in descending order of detection rate by the University A Baseline group.

Cue	UA-B	UA-T	UB-B	UB-T
like	0.81	0.83	0.75	0.81
just as	0.69	0.80	0.75	0.80
however	0.57	0.60	0.50	0.80
whereas	0.55	0.60	0.45	0.73
equivalent	0.45	0.59	0.35	0.50
both + and	0.25	0.13	0.28	0.39
meanwhile	0.25	0.25	0.20	0.20

Comparison Cues. We conducted an item-level examination to determine which cues were most and least noticed in the passages. Because cue type is confounded with the content

of the comparison(s) associated with it, this analysis is exploratory. There were 21 comparisons across the four passages, and 15 distinct comparison cues. (Some cues occurred more than once.) Table 2 shows the mean detection rate for a sample of cues. As expected, detection rates range widely over cues. Highly familiar comparison cues, such as *like* and *just as*, are correctly highlighted by most participants, but others, such as *however* and *meanwhile*, are often missed.

The results of E1 bore out our three predictions: 1. Without training, college students often failed to detect comparisons in science texts, despite being explicitly told to mark all comparisons.
2. Comparison detection rates were higher at University A than at University B.
3. Comparison-detection can be improved with training. A brief training session in detecting comparison cues improved students' ability to detect comparisons.

A fourth finding was that the training effect was greater for University B than for University A. There was no significant gain at University A.

Experiment 2

The results of E1 offer encouragement for the idea that comparison detection can be improved. However, this improvement occurred only for students at University B. In E2, we asked whether University A would also show a training effect with more challenging materials. A second motivation for E2 was concern that our comparison-detection task, which encouraged students to find all comparisons in a text, might lead them to neglect the materials' conceptual content. To address this concern, we included a control group that was asked to simply comprehend the same training materials, without the comparison-focused task. We used a pretest-posttest design to compare the two groups. In addition, we made improvements to the materials and task: we developed new materials containing a larger set of comparison cues; we altered the detection task to allow for nonadjacent comparisons (which occur fairly often in explanatory texts), and we asked participants to list the comparisons on a separate page, rather than highlighting them as in E1.

Methods

Participants. 25 undergraduate students from University A participated in the study. Participants completed the study for course credits. 13 students were randomly assigned to the Training condition and 12 to the Control condition.

Design. We used a between-subject (Training vs. Control), pretest-posttest design. The Training group received explicit instruction and feedback on detecting comparisons. The Control group read the same materials but was asked to focus on comprehending the materials. Both groups spent approximately 30-40 minutes on the task.

The study was presented on Qualtrics and consisted of three phases: Pretest, Training/Control, and Posttest. The comparison detection tasks were done on expository science passages adapted from popular science sources and textbooks.

Pretest. All participants first completed the comparison-listing task for a short science passage with seven possible analogical comparisons. They listed as many comparisons as they could identify on a physical worksheet. They received no feedback.

Training/Control. The two groups read the same materials but performed different tasks. We first describe the procedure for the Training group:

Training group. The group was told that the study aims to improve students' ability to notice comparisons in texts.

Choice questions. They next completed seven forced-choice questions in which they were asked to select the comparison, as in Expt. 1.

Comparison detection. Next, participants completed a Comparison detection task for a longer science passage on biology with 20 possible comparisons. Participants were instructed to list them on a physical worksheet in the same format as the pretest worksheet. After completing the task, they received the list of the 20 intended comparisons. After checking their list with the feedback list, they handed in their worksheet.

Control group. Participants received the same materials as the Training group but were given non-comparison-focused tasks. On the Choice questions, they were told to choose the more plausible sentence. On the Comparison detection passage, they were told to read it for comprehension. After reading, participants answered three comprehension questions, none of which involved comparisons. No feedback was provided to the Control group.

Posttest. The posttest assessed learners' ability to recognize comparisons in novel passages, as well as their comprehension of the training materials. The Training and Control groups received identical tasks. No feedback was given.

Choice questions. Both groups answered four Choice questions for which they were asked to choose the comparison, as in Expt. 1.

Transfer Questions. Next, participants answered two near-transfer questions and five far-transfer questions, applying training materials to new contexts. Near-transfer questions utilized modified biological cases, while far-transfer questions extended the concepts to social scenarios (e.g., Participants should choose "parasitism" as the best fit to the scenario: "Sarah's roommate, Mark, eats her groceries,…, and never pays his share of the bills, leaving Sarah to cover the costs")

Comparison detection. Then, participants were asked to list all comparisons they could find in a passage on a biological

topic, using a physical worksheet in the same format as the pretest worksheet.

Comprehension questions. Then, they answered three questions about examples from the training passage.

Ravens' Matrix. Participants then completed the short form of Raven's Progressive Matrices test (RPM).

Analysis Plan. There were four dependent variables of interest: performance on the comparison detection, performance on the choice questions, near-transfer performance, and far-transfer performance. We fitted mixed-effects logistic regression models to the trial-level accuracy data. Fixed effects included condition and pretest performance. All models included random intercepts for participants and for items. We then included the Raven's score as a covariate in the models to control for individual variation in fluid intelligence.

Results

The key finding was that, as predicted, the training group outperformed the control group in recognizing comparisons in texts. Moreover, comprehension and transfer performance did not differ significantly across groups.

Comparison detection. As predicted, training improved students' ability to identify comparisons in texts (Figure 3). A mixed-effects model confirmed that training significantly increased accuracy (β = 0.95, *SE* = 0.36, *z* = 2.61, *p* = .009). Pretest performance was also a significant positive predictor (β = 0.43, *SE* = 0.18, *z* = 2.32, *p* = .020).

Choice questions. Pretest performance was a significant positive predictor of performance (β = 0.66, *SE* = 0.32, *z* = 2.08, *p* = .038). The effect of training was nonsignificant (β = 1.35, *SE* = 0.70, *z* = 1.94, *p* = .052).

Transfer accuracy. Mixed-effects models did not find training to be a significant predictor for either near transfer (β = 0.50, *SE* = 0.36, *z* = 1.39, *p* = .17) or far transfer (β = 0.47, *SE* = 0.37, *z* = -0.5, *p* = .62). The pretest covariate was not a significant predictor either: near transfer (β = −0.04, *SE* = 0.70, *z* = -0.01, *p* = .96), far transfer (β = −0.01, *SE* = 0.34, *z* = -0.04, *p* = .97)[3]. The item-level variability was large: near transfer (SD = 1.23); far transfer (SD = 3.26).

Comprehension performance. Mixed-effects models did not find training to be a significant predictor of comprehension performance (β = 0.86, SE = 1.26, z = 0.69, p = .49). The pretest covariate was a significant predictor (β = 2.56, SE = 0.96, z = 2.67, p = .01). Participant-level variability was moderate (SD = 2.08), while item-level variability was negligible (SD = 0.00).

Including Raven's Score. The two groups did not differ in their Raven's scores (Training: M = 0.64; Control: M = 0.61, *t = 0.47, p = 0.64*). We reran the analyses for all four tasks, including Raven's score as a covariate. The pattern of results was largely unchanged, except that for Choice question accuracy, there was now a significant effect of training, (β = 1.46, *SE* = 0.67, *z* = 2.18, *p* = .029). For the Comparison-listing performance, the training effect remained significant (β = 0.97, *SE* = 0.36, *z* = 2.65, *p* = .008). For near and far transfer, the effects of training remained nonsignificant.

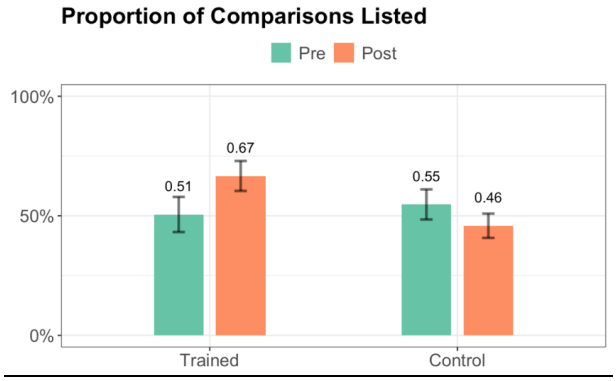


Figure 3: Proportion of comparisons listed (error bars represent standard errors)

Comparison Cues. See Table 3 for example detection rates.

Pretest passage. As expected, the familiar comparison cue (e.g., *just as*) was detected reliably, with consistently high mean rates for both the trained and control groups.

Table 3: Example mean rates of detection for comparison cues across conditions for pretest and posttest.

Cue	Control	Trained	Phase
equivalent	0.91	0.77	pretest
just as	0.82	0.79	pretest
or	0.36	0.62	posttest
mimic	0.64	0.92	posttest

Posttest Passage. The trained group outperformed the control group on all but one comparison cue, and several cues showed large differences. These included contrastive cues as well as cues that also have non-comparison functions. For example, "or" was used contrastively in the passage ("*to protect **or** to attack*"), and "or" canonically encodes logical disjunction, not comparison. (However, as noted earlier, cue detectability depends on the comparison context as well as on the cue itself, so these observations are preliminary.)

Discussion

Despite the large amount of work on how students learn from comparisons, little is known about whether and how they notice comparisons in the first place. The present work addresses this gap by directly measuring students' ability to detect comparisons in science texts and investigating whether targeted training can improve that ability. To our knowledge, this is the first work to investigate this issue.

We examined whether students vary in their ability to spot comparisons in the materials they read and, if so, whether training would improve their performance. There are four

main findings. Our first and perhaps most surprising finding is that even college students often miss intended comparisons in texts. Despite being directly told to mark all comparisons in science texts, students in the Baseline condition missed about half the comparisons.

The second finding is that without training, students at the more selective institution were better able to identify comparisons than those at the less selective institution. This could be because the students at the more selective institution are better acquainted with the cues that signal comparison, or because they are more aware of the value of attending to comparisons.

The third and most heartening finding is that a fairly brief training session on comparison cues improved college students' ability to detect comparisons. Further, we saw no evidence that this task inhibited comprehension of the text. The effect of training was robust across all tasks and persisted after controlling for individual differences in fluid intelligence. Importantly, the training generalized to new materials. The test passages differed in topic and tone from the training materials. This suggests that training in detecting comparisons could yield domain-general benefits in an important aspect of analogical learning.

Finally, we found evidence for an interaction such that students from University B benefited more from the training than did those from University A. This is not surprising, since the University B students began at a lower level. However, the encouraging finding is that training largely eliminated the baseline gap found between students at the two universities (Figures 1 and 2). This suggests that training of this kind may help to remediate gaps in students' skills at detecting comparisons. Further, E2 shows that with more targeted training, University A students can improve, suggesting that training may be broadly helpful in helping students achieve comparison detection.

Conclusion

These findings highlight a critical gap in our understanding of how students process comparisons in science texts. To our knowledge, this is the first investigation of students' ability to recognize intended comparisons in text. It is dismaying to discover that even university students miss many of the comparisons in texts, even when directly instructed to find comparisons. Because comparisons across examples and across categories are prevalent in science writing, this lack of sensitivity is a serious concern. On the positive side, the effectiveness of our brief training session offers reason to believe that this situation can be improved. The ability to notice comparisons in ongoing contexts appears to be a malleable subskill in analogical reasoning.

References

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. *Psychonomic Bulletin & Review, 6*(2), 328-337.

Alexander, P. A., Pate, P. E., Kulikowich, J. M., Farrell, D. M., & Wright, N. L. (1989). Domain-specific and strategic knowledge: Effects of training on students of differing ages or competence levels. *Learning and Individual Differences, 1*(3), 283-325.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist, 48*(2), 87-113.

Alibali, M. W., Nathan, M. J., Wolfgram, M. S., Church, R. B., Jacobs, S. A., Johnson Martinez, C., & Knuth, E. J. (2014). How teachers link ideas in mathematics instruction using speech and gesture: A corpus analysis. *Cognition and instruction, 32*(1), 65-100.

Begolli, K. N., & Richland, L. E. (2016). Teaching mathematics by comparison: Analog visibility as a double-edged sword. *Journal of Educational Psychology, 108*(2), 194.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment, 19*(3), 354-369.

Blanchette, I., & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory & Cognition, 28*(1), 108-124.

Burton, N. W., & Ramist, L. (2001). Predicting Success in College: SAT® Studies of Classes Graduating since 1980. Research Report No. 2001-2. *College Entrance Examination Board*.

Caro, D. H., McDonald, J. T., & Willms, J. D. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education/Revue canadienne de l'éducation, 32*(3), 558-590.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological review, 97*(3), 404.

Chen, Z. (2002). Analogical problem solving: A hierarchical analysis of procedural similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 81-98.

Chen, Z., & Mo, L. (2004). Schema Induction in Problem Solving: A Multidimensional Analysis. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 30, 583-600.

Chi, M. T., J. D., & De Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and instruction, 4*(1), 27-43.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development, 11*(3), 356-373.

Christie, S. & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science, 38*, 383-397.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of research in science teaching, 30*(10), 1241-1257.

Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive science*, *15*(1), 89-132.

Compton, J., & van der Linden, S. (2022). The analogy of/and inoculation theory to mental immunity. *This View of Life*.

Corkill, A. J., & Fager, J. J. (1995). Individual differences in transfer via analogy. *Learning and Individual Differences*, *7*(3), 163-187.

Curtis, R. V., & Reigeluth, C. M. (1984). The use of analogies in written text. *Instructional Science*, *13*(2), 99-117.

Dixon, J. A., & Dohn, M. C. (2003). Redescription disembeds relations: Evidence from relational transfer and use in problem solving. *Memory & cognition*, *31*, 1082-1093.

Doumas, L. A., & Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. *PloS one*, *8*(6), e63889.

Dunbar, K. (2001). The analogical paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 313–334). The MIT Press.

Emmons, N., Lees, K., & Kelemen, D. (2018). Young children's near and far transfer of the basic theory of natural selection: An analogical storybook intervention. *Journal of Research in Science Teaching*, *55*(3), 321-347.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, *41*(1), 1-63.

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, *41*(5), 1152-1201.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155-170.

Gentner, D. (2002). Analogy in scientific discovery: The case of Johannes Kepler. In *Model-based reasoning: Science, technology, values* (pp. 21-39). New York, NY: Springer US.

Gentner, D., Levine, S. C., Ping, R., Isaia, A., Dhillon, S., Bradley, C., & Honke, G. (2016). Rapid learning in a children's museum via analogical comparison. *Cognitive science*, *40*(1), 224-240.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American psychologist*, *52*(1), 45.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology*, *15*(1), 1-38.

Glynn, S. M. (2008). 5.1 Making science concepts meaningful to students: teaching with analogies. *Four Decades of Research in Science Education-from Curriculum Development to Quality Improvement: From Curriculum Development to Quality Improvement*, *113*.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the non-independence of features in similarity judgment. *Cognitive Psychology*, *23*, 222-264.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the learning sciences*, *14*(1), 69-110.

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition*, *137*, 137-153.

Goldwater, M. B., Gentner, D., LaDue, N. D., & Libarkin, J. C. (2021). Analogy generation in science experts and novices. *Cognitive Science*, *45*(9), e13036.

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729.

Gray, M. E., & Holyoak, K. J. (2020). Individual differences in relational reasoning. *Memory & Cognition*, 48. 96-110.

Gray, M. E., & Holyoak, K. J. (2021). Teaching by analogy: From theory to practice. *Mind, Brain, and Education*, *15*(3), 250-263.

Green, A. E., Fugelsang, J. A., Kraemer, D. J. M., Shamosh, N. A., & Dunbar, K. (2006) Frontopolar cortex mediates abstract integration in analogy. *Brain Research*, 1096, 125-137.

Hattikudur, S., & Alibali, M. W. (2010). Learning about the equal sign: Does comparing with inequality symbols help?. *Journal of experimental child psychology*, *107*(1), 15-30.

Harrison, A. G., & Treagust, D. F. (2006). Teaching and learning with analogies: Friend or foe?. *Metaphor and analogy in science education*, 11-24.

Halford, G. S. (1992). Analogical reasoning and conceptual complexity in cognitive development. *Human Development*, *35*(4), 193-218.

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society

Holyoak, K. J. (2012). Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, 234-259.

Honke, G., & Kurtz, K. J. (2018). Similarity is as similarity does? A critical inquiry into the effect of Thematic Association on Similarity. *Cognition,* 186, 115–138. https://doi.org/10.31234/osf.io/v4xtc

Honke, G., Kurtz, K. J., & Laszlo, S. (2019). Similarity judgments predict N400 amplitude differences between taxonomic category members and Thematic Associates. *Neuropsychologia*, *141*.

Honomichi, R. D, & Chen, Z. (2006) Learning to align relations: The effects of feedback and self-explanation. *Journal of Cognition and Development*, *7*(4), 527–550.

Hoyos, C., Horton, W. S., Simms, N. K., & Gentner, D. (2020). Analogical comparison promotes theory-of-mind development. *Cognitive Science*, *44*(9), e12891.

Hummel, J. E. (2010). Symbolic versus associative learning. Cognitive Science, 34(6), 958-965.

Hummel, J. E., & Holyoak, K. J. (1997). LISA: A computational model of analogical inference and schema induction. *Psychological Review.*

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review. 110*(3), 220-264.

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T. F., Tikoff, B., ... & Sageman, B. (2010). Commentary: Analogical thinking in geoscience education. *Journal of Geoscience Education*, *58*(1), 2-13.

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T. F., & Sageman, B. (2013). Finding faults: Analogical comparison supports spatial concept learning in geoscience. *Cognitive processing*, *14*, 175-187.

Kok, E. M., de Bruin, A. B., Robben, S. G., & van Merriënboer, J. J. (2013). Learning radiological appearances of diseases: does comparison help?. *Learning and Instruction*, *23*, 90-97.

Krawczyk, D. C., Holyoak, K. J., & Hummel, J. E. (2004). Structural constraints and object similarity in analogical mapping and inference. *Thinking & reasoning*, *10*(1), 85-104.

Krawczyk, D. C., Holyoak, K. J., & Hummel, J. E. (2005). The one-to-one constraint in analogical mapping and inference. *Cognitive science*, *29*(5), 797-806.

Kurtz, K. J., & Gentner, D. (2013). Detecting anomalous features in complex stimuli: the role of structured comparison. *Journal of Experimental Psychology: Applied*, *19*(3), 219.

Kurtz, K. J., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*(4), 417-446.

Legare, C. H., Opfer, J. E., Busch, J. T., & Shtulman, A. (2018). A field guide for teaching evolution in the social sciences. *Evolution and Human Behavior*, *39*(3), 257-268.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6*(4), 586-597.

Lovett, A. & Forbus, K.D. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, 124 (1): 60-90.

Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Portland, OR.

Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive science*, *33*(7), 1192-1231.

Mannion, J., & Mercer, N. (2016). Learning to learn: improving attainment, closing the gap at Key Stage 3. *The Curriculum Journal*, *27*(2), 246-271.

Markman, A. B. (1997). Constraints on analogical inference. *Cognitive science*, *21*(4), 373-418.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431-467.

Mason, L. (1994). Analogy, metaconceptual awareness and conceptual change: A classroom study. *Educational Studies*, *20*(2), 267-291.

Matlen, B. J., Gentner, D., & Franconeri, S. L. (2020). Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 46(5), 443-457.

McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Processes*, *54*(7), 479-492.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278.

Mix, K. S. (2008a). Children's equivalence judgments: Cross-mapping effects. *Cognitive development, 23*(1), 191-203.

Mix, K. S. (2008b). Surface similarity and label knowledge impact early numerical comparisons. *British Journal of Developmental Psychology*, *26*(1), 13-32.

Mix, K. S., Bower, C. A., Hancock, G. R., Crespo, S., & Smith, L. B. (2024). Place value learning in kindergarten. *Journal of Educational Psychology*.

Murphy, A. N., Zheng, Y., Shivaram, A., Vollman, E., & Richland, L. E. (2021). Bias and sensitivity to task constraints in spontaneous relational attention. *Journal of Experimental Child Psychology*, *202*, 104981.

Nokes-Malach, T. J., VanLehn, K., Belenky, D. M., Lichtenstein, M., & Cox, G. (2013). Coordinating principles and examples through analogy and self-explanation. *European Journal of Psychology of Education*, *28*, 1237-1263.

Orton, J. M., Anggoro, F. K., & Jee, B. D. (2012). Mutual alignment comparison facilitates abstraction and transfer of a complex scientific concept. *Educational Studies*, *38*(4), 473-477.

Resnick, I., Davatzes, A., Newcombe, N. S., & Shipley, T. F. (2017). Using analogy to learn about phenomena at scales outside human perception. *Cognitive Research: Principles and Implications*, *2*, 1-17.

Richland, L. E. (2015). Linking gestures: Cross-cultural variation during instructional analogies. *Cognition and Instruction*, *33*(4), 295-321.

Richland, L. E., & McDonough, I. M. (2010). Learning by analogy Discriminating between potential analogs. Contemporary Educational Psychology, 35(1), 28-43.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of experimental child psychology*, *94*(3), 249-273.

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177-192.

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128-1129.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural

knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561.

Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology*, *101*(3), 529.

Rittle-Johnson, B., & Star, J. R. (2011). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons. In *Psychology of Learning and Motivation* (Vol. 55, pp. 199–225). Elsevier.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 629-639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*(3), 456-468.

Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory & Language*, *37*(2), 240-267.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive science*, *36*(5), 919-932.

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*(6), 1019-1050.

Seiler, K. P., & Huggins, J. (2018). From cheese curls to fatty acid structure: using "commonplace" analogies to teach science to nonmajors. *Advances in Physiology Education*, *42*(2), 393-395.

Sheldrake, M. (2021). *Entangled life: How fungi make our worlds, change our minds & shape our futures*. Random House Trade Paperbacks.

Shtulman, A., Neal, C., & Lindquist, G. (2019). Children's ability to learn evolutionary explanations for biological adaptation. In *Young Children's Developing Understanding of the Biological World* (pp. 123-137). Routledge.

Simms, N. K., Matlen, B. J., Jee, B. D., & Gentner, D. (2023). Spatial alignment supports comparison of life science images. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000471

Son, J. Y., & Goldstone, R. L. (2005). Relational words as handles: They bring along baggage. *Proceedings of the Twenty-seventh Annual Conference of the Cognitive Science Society*. Hillsdale, New Jersey: Lawrence Erlbaum Associates. (pp. 2050-2055)

Sullivan, J., & Barner, D. (2014). The development of structural analogy in number-line estimation. *Journal of Experimental Child Psychology*, *128*, 171-189.

Treagust, D. F., Harrison, A. G., & Venville, G. J. (1996). Using an analogical teaching approach to engender conceptual change. *International journal of science education*, *18*(2), 213-229.

Treagust, D. F., Harrison, A. G., & Venville, G. J. (1998). Teaching science effectively with analogies: An approach for preservice and in-service teacher education. *Journal of Science Teacher Education*, *9*(2), 85-101.

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, *81*(6), 1768-1786.

Trench, M., & Minervino, R. A. (2015). The role of surface similarity in analogical retrieval: Bridging the gap between the naturalistic and the experimental traditions. *Cognitive Science*, 39(6), 1292-1319.

Trench, M., & Minervino, R. A. (2020). Distant connections: The memory basis of creative analogy. New York: Springer. https://doi.org/10.1007/978-3-030-52545-3 776

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: a meta-analysis of training studies. *Psychological bulletin*, *139*(2), 352.

Vendetti, M. S., Matlen, B. J., Richland, L. E., & Bunge, S. A. (2015). Analogical reasoning in the classroom: Insights from cognitive science. *Mind, Brain, and Education*, *9*(2), 100-106.

Vendetti, M. S., Wu, A., & Holyoak, K. J. (2014). Far-out thinking generating solutions to distant analogies promotes relational thinking. *Psychological Science*, DOI:0956797613518079.

Wakefield, E. M., & Goldin-Meadow, S. (2019). Harnessing our hands to teach mathematics: How gesture can be used as a teaching tool in the classroom.Yang, W., Liu, H., Chen, N., Xu, P., & Lin, X. (2020). Is early spatial skills training effective? A meta-analysis. *Frontiers in psychology*, *11*, 1938.

Yuan, L., Uttal, D., & Gentner, D. (2017). Analogical processes in children's understanding of spatial representations. *Developmental psychology*, *53*(6), 1098.

Zheng, Y. S., Goldwater, M., Gentner, D. (2025). Structural alignment and linguistic contrast help children learn a key principle of spatial construction. *Cognitive Science 49*, 12. **https://doi.org/10.1111/cogs.70149**

Zheng, Y., Matlen, B., & Gentner, D. (2022). Spatial alignment facilitates visual comparison in children. *Cognitive Science*, *46*(8), e13182.