COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13496

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13496

Allow Me to Explain: Benefits of Explaining Extend to Distal Academic Performance

Anahid S. Modrek,[a] Tania Lombrozo[b]

[a]*Department of Psychology, California State University, San Bernardino*
[b]*Department of Psychology, Princeton University*

Received 7 November 2023; received in revised form 31 July 2024; accepted 24 August 2024

Abstract

How does the act of explaining influence learning? Prior work has studied effects of explaining through a predominantly proximal lens, measuring short-term outcomes or manipulations within lab settings. Here, we ask whether the benefits of explaining extend to academic performance over time. Specifically, does the quality and frequency of student explanations predict students' later performance on standardized tests of math and English? In Study 1 ($N = 127$ 5th−6th graders), participants completed a causal learning activity during which their explanation quality was evaluated. Controlling for prior test scores, explanation quality directly predicted both math and English standardized test scores the following year. In Study 2 ($N =$ 20,384 10th graders), participants reported aspects of teachers' explanations and their own. Controlling for prior test scores, students' own explanations predicted both math and English state standardized test scores, and teacher explanations were linked to test performance *through* students' own explanations. Taken together, these findings suggest that benefits of explaining may result in part from the development of a metacognitive explanatory skill that transfers across domains and over time. Implications for cognitive science, pedagogy, and education are discussed.

Keywords: Explanations; State standardized test scores; Metacognition; Long-term academic performance; Pedagogy

Correspondence should be sent to Anahid S. Modrek, PhD, Department of Psychology, CSUSB, 5500 University Parkway, San Bernardino, CA 92407, USA. E-mail: anahid.modrek@csusb.edu

1. Introduction

Cognitive, educational, and developmental scientists have established many benefits of explanations for learning. Explanations can support the acquisition of new content and the development of domain-specific skills (e.g., Chi, 2018; Chi & Fonseca, 2011; Rittle-Johnson, 2006; Webb, 1982b). Explanations can also support the development of domain-general skills, such as metacognitive monitoring (Chi, 2018; Moreno & Mayer, 2000; Rozenblit & Keil, 2002), which are especially promising targets for educational interventions (Duncan et al., 2007): if explanations foster general skills for learning, explaining could have learning benefits that accrue over time and that transfer across domains. Because prior work has overwhelmingly focused on short-term effects of explanations within particular content domains (e.g., Chi & Fonseca, 2011; Lombrozo, 2016; Rittle-Johnson, 2006), this possibility remains unexplored. Here, we ask: Can the benefits of explanations extend to long-term, distal outcomes, such as academic performance on standardized tests? And do such benefits generalize across domains? We address these questions using two existing datasets: one that allows us to evaluate explanation quality and quantity in a causal learning task with elementary-school-aged children, and another that involves a large sample of adolescents who reported aspects of teachers' explanations and their own. In each case, we test whether explanation frequency or quality predicts later standardized test scores (both math and English), controlling for prior standardized test performance and various demographics.

This work builds on previous research demonstrating that explanations support both content learning and skill learning. Contributing to content learning, explanations facilitate the acquisition and transfer of material by scaffolding problem representations that support generalization, with effects documented in children as young as 3–6 years (e.g., Legare & Lombrozo, 2014; Walker & Lombrozo, 2017; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2017) and in adulthood (e.g., Edwards, Williams, Gentner, & Lombrozo, 2019; Lombrozo, 2016; Williams & Lombrozo, 2010). For example, 8- to 11-year-old children who practiced addition problems were more likely to succeed in solving transfer subtraction problems if prompted to explain the initial addition problems (Rittle-Johnson, 2006; see also Webb, 1982a).

Supporting skill learning, explaining can improve metacognitive calibration in both adults (e.g., Rozenblit & Keil, 2002) and school-aged children (Mills & Keil, 2004), direct processing to explanation-relevant content in both adults (e.g., Williams & Lombrozo, 2013) and children (e.g., Legare & Lombrozo 2014; Vasil, Ruggeri, & Lombrozo, 2022; Walker et al., 2017; Walker, Lombrozo, Legare, & Gopnik, 2014), and support abstract causal reasoning that directs further inquiry by age 8 (Astington & Gopnik, 1991; Chuey et al., 2021; Goddu, Lombrozo, & Gopnik, 2020; Gopnik, 2000; Kushnir, Vredenburgh, & Schneider, 2013; Ruggeri, Xu, & Lombrozo, 2019). Moreover, explaining can itself be cultivated as a skill: beyond individual differences in self-explanation (Ainsworth & Loizou, 2003; Aleven & Koedinger, 2002; Chi, 2018; Renkl, 1997), there is evidence from high school and college students that self-explanation can be trained (Bielaczyc, Pirolli, & Brown, 1995; McNamara, 2017; McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007).

Given the strong link between metacognitive skills and academic achievement (Aleven & Koedinger, 2002; Blair & Raver, 2014; Duckworth, Kirby, Gollwitzer, & Oettingen, 2013;

Mills & Keil, 2004), as well as consistent, robust associations between metacognitive skills and long-term academic achievement (Duncan et al., 2007), we might expect that being a frequent and adept explainer similarly improves academic performance, with potentially distal effects (e.g., on later standardized test scores). However, this has not been tested in prior work. The only study investigating effects of explanation that included standardized test scores (that we know of) comes from Chi, De Leeuw, Chiu, and LaVancher (1994), who showed eighth-graders learned more about the circulatory system when prompted to explain as they studied an expository text. The benefits of being prompted to explain held when controlling for prior standardized test performance, and within the group of students prompted to explain, gains were comparable (around 30%) for those scoring highest and lowest on the standardized test. Our datasets allow us to evaluate whether explanation quantity and skill predict later performance not only *controlling* for prior test scores, but also investigating standardized test performance itself as a long-term, distal outcome.

1.1. The present study

In sum, prior work has documented relationships between explanation and skills important for learning (such as metacognitive monitoring), as well as links between metacognitive skills and academic achievement. However, prior work has not (a) tested the frequency or quality of explanations as predictors of academic achievement (e.g., on standardized test scores for math and English), (b) studied these associations longitudinally while controlling for prior performance, or (c) done so with large, diverse samples. These gaps represent important avenues to pursue. Beyond informing our understanding of how explaining shapes learning, finding links between explanation and later academic achievement has important implications for education. Thus, we posit and test the hypothesis that explanation frequency and quality both reflect and support learning skills (such as metacognitive monitoring) that contribute to academic achievement over time and across domains.

We utilize two unique datasets that allow us to test our hypotheses with parallel models and different samples, with both math and English Language Arts (ELA) state standardized test scores, while also controlling for prior scores. Our first dataset is a sample of 127 participants (ages 10–12 years) who were involved in a study investigating different aspects of self-regulation (i.e., cognitive vs. behavioral) as predictors of inductive learning and academic achievement (Modrek, Kuhn, Conway, & Arvidsson, 2019). While aspects of these data have been previously published, our own analyses are based on unpublished data that have not been previously analyzed, and that offer a direct test of our hypotheses. Specifically, by coding explanations generated during an inductive learning task (Study 1), the data allow us to evaluate the quality and frequency of explanations and their association with both math and ELA state standardized test scores across 2 years. Our second dataset (Study 2) offers a conceptual replication and extension with a sample of 20,384 participants (primarily aged 15) from a quasi-experiment across the United States, part of a larger project with The William and Flora Hewlett Foundation and American Institutes for Research (AIR) focused on promoting opportunities for deeper learning (e.g., complex problem-solving) with deeper learning schools. This quasi-experiment includes non-charter/non-magnet schools that underwent professional development focused on deeper learning, as well as demographically matched non-charter/non-magnet schools that did not receive this professional development. Within this larger project, student participants completed several scales (Locus of Control, Self-Efficacy, Belonging, etc.) and additional ratings, totaling approximately 200 items. Among the items that did not belong to formal scales were questions about explanation. These items have not been previously analyzed or published; we use them here as a proxy for explanation frequency in classrooms. In partnership with AIR, our dataset includes student participants' math and ELA state standardized test performance, along with their prior scores (to serve as controls). Taken together, these two datasets offer a unique opportunity to test our hypotheses concerning the link between explanation and state standardized test performance across both math and ELA (while controlling for prior scores).

Given that we used existing datasets, our procedures were not preregistered. However, all hypotheses were theory-driven and formulated prior to testing; analyses and results are thus confirmatory.

2. Study 1

Participants provided explanations during an inductive learning activity administered early in the academic year. State standardized test scores were obtained approximately 2 years later and included both the year during which the activity was administered and the following year. This allowed us to test the hypothesis that explanation quantity and quality, as assessed during the inductive learning activity, predicts later academic achievement (controlling for prior standardized test scores).

2.1. Methods

2.1.1. Participants

Participants were 127 students aged 10–12 years (at baseline) recruited from fifth and sixth grade classes in a metropolitan school district in New York ($M_{age} = 11.3$, $SD = 0.67$; 55% female; 45% male). The sample was 52% Caucasian, 12% Asian, 10% African-American, 7% Hispanic, and 19% mixed background. Many were bilingual (56%), with common languages including Russian, Hebrew, Italian, Greek, and Mandarin. According to the state Department of Education, school performance ranking was in the 47th percentile. Approximately 10% of students qualified for free/reduced-price lunch (FRPL).

2.1.2. Materials and procedure

2.1.2.1. Inductive learning task: Participants completed an inductive learning task requiring them to examine data consisting of cases that varied on multiple dimensions to identify associations with an outcome variable. Similar tasks have been used in studies investigating inquiry learning, causal learning, inductive inference, and problem-solving (Dean & Kuhn 2007; Greiff et al., 2013; Holyoak & Cheng 2011). The task was administered individually to each participant and took on average 60 minutes. There were three phases (introduction, induction, application), described below. Explanations were collected at four points during the induction phase. (For complete task procedures and instructions, see Modrek et al., 2019.)

During the introduction phase, participants learned about a space foundation recruiting astronauts. The foundation was investigating four factors (fitness, family size, education, and parents' health) that might affect astronaut performance. Participants reported their prior hypotheses about each factor: whether it would have an effect on performance, and if so, in what direction (e.g., whether large vs. small family size predicts better performance).

During the induction phase, participants received data about how astronauts performed in a simulator. The dataset contained exemplars of astronauts with varying fitness, family size, education, and parents' health. Participants could compare or request specific cases to review. Each factor was investigated serially, and the interviewer invited (but did not require) participants to conclude whether the factor made a difference. After each factor, participants were asked to explain how they drew each inference. This was repeated for all four factors until participants reported being satisfied with their level of inquiry. By the end of this phase, participants had a summary sheet indicating their final inferences about all factors.

During the application phase, participants were presented with profiles of new astronauts, which they evaluated with access to their summary sheet. Participants had to predict how well each astronaut would perform, reported on a Likert scale. This allowed us to assess the extent to which participants relied on learning in the induction phase, versus initial hypotheses based on prior beliefs.

2.1.2.2. State standardized test scores: Math and English state standardized test scores (ranging from 1 to 4) were obtained for all participants, across 2 years: both the end of the academic year during which the inductive learning task was administered, as well as the following year.

2.1.3. Scoring

Participants were first given a score (0/1) based on whether they provided an explanation justifying *how* they drew an inference. Next, participants received a separate score (0/1) based on whether they explained their inference in line with the observed data. Making no reference to the data would result in a score of 0. Finally, participants received a third score (0/1) indicating whether they provided an explanation that referenced the data and evidence *accurately* and *consistently* across the task. These three scores resulted in a single composite score of 0–3 designed to assess the quality and quantity of explanations generated in the task (see Table 1). For reliability, a second rater independently coded 10% of the explanation data, achieving a kappa of .815 with the ratings of the first author.

2.2. Results and discussion

We first report descriptive statistics for our variables of interest: explanation scores (M = 1.080, SD = 1.004), math standardized test scores (M = 3.702, SD = .612), and English standardized test scores (M = 3.479, SD = .641). We next test our focal hypothesis that

Table 1
Coding of participant explanations in Study 1

Score	*Example*
Mechanistic justification	
0: Participant simply restated an association without offering an explanation.	"…you have to go to college if you want to be a good astronaut…" "…I mean I just think that you need to have the overall package and lots of things going for you."
1: Participant provided a mechanism underlying an association to justify an inference.	"…when you are more fit I guess maybe it's easier to do things so it's better off having excellent than average so…"
Data reference	
0: Participant provided no reference to data.	"…I guess when you are fit it's easier to do things."
1: Participant explained inferences using data.	"…so both of these tables uh have the same things except for [points to data] 12 which has uh fair performance, and she did poorly but not as bad as them."
Interpreted data evidence correctly and consistently	
0: Participant either referenced data and did so incorrectly or did not consistently reference data across their four explanations.	[*Initially appeals to data*] "… it's better to have a smaller family than a larger family. Because Yollanda had a small family and she did well and Cory had a large family and he didn't do so well…" [*Later draws incorrect conclusion from data, showing incorrect/inconsistent use of data*] "…because they had excellent and they still did really bad. I still think that it kind of makes a difference, because whether health was fair or excellent they still did bad…"
1: Participant provided explanations that referenced the data *accurately* and did so *consistently* across all four explanations.	"…like I want to find a card that has like small family. [*participant pointing to card that allows for control of variables comparison to be made*] It doesn't make a difference. Well, again, john has a small family um but jake has a large family and they both did very well they both have excellent everything and both went to college but one has small family and one has large but both had same performance…" followed by a later explanation given to justify their conclusion "…..well I went based off of the cards and tried to make sure everything else was the same besides for what we looked at."
Final explanation score range: 0–3	

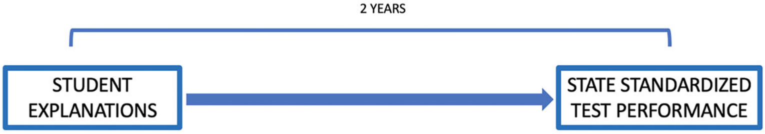

Fig. 1. Theoretical model tested in Study 1.

Table 2
Explanations predicting state standardized test scores for math and ELA across 2 years

	Coef.	Std. Error	z	*p*	[95% Conf. Interval]	
				English		
English (Time 2)						
Explanations	.1280295	.0638299	2.01	.045	.0029252	.2531339
English (Time 1)	.6638868	.0485702	13.67	.000	.5686909	.7590828
Gender	−.0333797	.0634345	−0.53	.599	−.1577091	.0909497
Bilingualism	.0433802	.0629637	0.69	.491	−.0800264	.1667867
Age	.0523568	.0651318	0.80	.421	−.0752992	.1800127
Intercept	.784727	1.082957	0.72	.469	−1.337829	2.907283
				Math		
Math (Time 2)						
Explanations	.1492751	.0558829	2.67	.008	.0397466	.2588037
Math (Time 1)	.714542	.0389678	18.34	.638	.6004788	.7909174
Gender	−.0653756	.0531368	−1.23	.219	−.1695218	.0387706
Bilingualism	.0505556	.0561943	0.90	.368	−.0595832	.1606944
Age	.0860296	.0558787	1.54	.124	−.0234907	.1955499
Intercept	.190344	.9681453	0.20	.844	−1.707186	−1.707186

Note. Results are standardized. $N = 127$. Participants self-reported gender, with 55% identifying as female and 45% identifying as male (coded as 1 or 0, respectively). If participants self-reported being bilingual, a researcher asked them to recite a basic introductory greeting. If participants showed basic proficiency, their bilingualism score was 1, and if they either reported not being bilingual, or were unable to provide a basic introductory greeting, their bilingualism score was 0; about half (56%) were bilingual.

explanation quality and quantity (as reflected in the explanation score) predicts state standardized test performance, even controlling for prior scores (see theoretical model, Fig. 1).

All models included age, gender, and bilingualism as covariates, as these factors had significant associations with explanation, math, or English scores (age & explanation: $r = .229$, $p < .01$; age & English scores: $r = .176, p < .05$; gender & English scores: $r = .241, p < .05$; bilingualism & English scores: $r = .217, p < .05$).

We employed a linear regression model providing a saturated, perfect fit to the data using a restricted maximum likelihood approach as a form of maximum likelihood estimation that does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function calculated from a transformed set of data. We entered year 2 state standardized test performance as the dependent variable, controlling for respective prior scores, followed by the aforementioned covariates. Separate models were run for English and math scores.

As hypothesized, explanations predicted state standardized test performance, for both English and math, controlling for respective prior scores (see Table 2). Given that the explanations elicited during the inductive learning task did not share content with the math or English standardized tests, this suggests that explanatory tendency or skill, as reflected in the explanation score from the inductive learning task, explained unique variance in later academic achievement. Controlling for prior test scores helps rule out the plausible alternative that explanatory tendency and skill instead shared a common cause with later test performance (such as general academic ability).

We also disaggregated explanation scores to test whether explanation quality and quantity each contributed to predictions. We entered each individual score in our models in lieu of the composite. For English, the first and third scores were significant predictors (95% CI [.0778972, .4167099] $p = .004$; 95% CI [.0033406, .1918635] $p = .042$, respectively); the second was not (95% CI [−.1004695, .2718672] $p = .367$). For math, the results were the same: the first and third scores were significant predictors (95% CI [.0865342, .3733355] $p = .002$; 95% CI [.0211586, .1808575] $p = .013$, respectively); the second score was not (95% CI [−.0706251, .251285] $p = .271$). This suggests that while explanation quality and quantity both mattered, simply explaining (the first score) was sufficient to predict later academic performance.

3. Study 2

A large sample of 10th-grade students completed questionnaires early in the academic year. State standardized test scores for math and English were obtained for that same year as well as prior years. This allowed us to test our hypothesis that explanations during the school year would predict standardized test scores for that same year, even controlling for the prior year's test performance. Additional measures allowed us to assess the role of students' reports of teachers' explanations, and to compare the effects of reported explanation to those of other constructive activities (such as combining different ideas).

3.1. Methods

3.1.1. Participants

Participants were 20,384 10th-grade students, primarily aged 15 ($M_{age} = 15.977$, $SD = 0.848$), sampled from 24 non-magnet/non-charter schools across the United States. Across the 24 schools, racial demographics were mixed, resulting in a diverse sample. On average, 60% of students qualified for FRPL. (For additional demographics, please see Table S1.) About half the schools reported participating in professional development focused on deep learning (which we control for in our models, described below).

3.1.2. Materials and procedure

3.1.2.1. Explanations: Participants answered six items about explanation (see Tables 3 and 4): three about their own explanations (e.g., "I interpret data and *explain* what the results

Table 3
Structural equation models for math and ELA standardized test scores predicted by teacher explanation scale, then student explanation scale, after controlling for prior scores

	Coef.	Std. Error	z	$p>\|z\|$	[95% Conf. Interval]	
				English		
Student Explanations						
Teacher Explanations	.3778202	.0367543	10.28	.000	.305783	.4498574
English (8th grade)	.0915569	.0717385	1.28	.202	−.049048	.2321618
English (7th grade)	.0901034	.0716018	1.26	.208	−.0502336	.2304403
School Professional Development	−.0571506	.0412428	−1.39	.166	−.1379851	.0236838
Intercept	.8474783	.3566683	2.38	.017	.1484212	1.546535
English (10th **grade)**						
Student Explanations	.0565916	.0284449	1.99	.047	.0008406	.1123426
Teacher Explanations	−.0038417	.0280234	−0.14	.891	−.0587666	.0510831
English (8th grade)	.377042	.0441966	8.53	.000	.2904182	.4636658
English (7th grade)	.4661155	.0434488	10.73	.000	.3809574	.5512735
School Professional Development	−.0210007	.0261147	−0.80	.421	−.0721846	.0301833
Intercept	−5.41368	.1543291	−35.08	.000	−5.71616	−5.111201
var(e.StudentExplanations)	*.8156301*	*.0300874*			*.7587412*	*.8767844*
var(e.English10thgrade)	*.3244561*	*.0196059*			*.2882175*	*.3652511*
				Math		
Student Explanations						
Teacher Explanations	.3812397	.0366281	10.41	.000	.30945	.4530294
Math (8th grade)	.0552628	.0580121	0.95	.341	−.0584387	.1689644
Math (7th grade)	.1210385	.0576156	2.10	.036	.0081141	.233963
School Professional Development	−.0643845	.0411533	−1.56	.118	−.1450434	.0162744
Intercept	.9994371	.3368445	2.97	.003	.3392339	1.65964
Math (10th **grade)**						
Student Explanations	.0633838	.0320145	1.98	.048	.0006364	.1261311
Teacher Explanations	.0235085	.0315947	0.74	.457	−.0384161	.085433
Math (8th grade)	.2548453	.0405246	6.29	.000	.1754185	.334272
Math (7th grade)	.5380281	.0376613	14.29	.000	.4642133	.611843
School Professional Development	−.0515285	.0293676	−1.75	.079	−.1090879	.0060309
Intercept	−4.546492	.1746181	−26.04	.000	−4.888737	−4.204246
var(e.StudentExplanations)	*.8178218*	*.029976*			*.7611305*	*.8787355*
var(e.Math10thgrade)	*.4133481*	*.0240285*			*.3688369*	*.4632308*

Note. Results are standardized. *Var(e.-)* are variances as exogenous variables.

mean"; Cronbach's alpha .54), and three about their teachers' explanations (e.g., "My teacher explains difficult things clearly"; Cronbach's alpha .74). All items were rated on a frequency scale (1-Never, 2-Some of the time, 3-Most of the time, 4-All of the time), and together produced a Cronbach's alpha of .72, demonstrating adequate reliability. In answering these questions, students were instructed to think about their core classes (not one specific teacher).[1]

3.1.2.2. Non-explanation/comparison items: Participants also answered items about other activities that plausibly support cognitive enrichment (see Table 4); these were analyzed in an effort to distinguish cognitive benefits that are potentially unique to explanation

Table 4
Correlation table of explanation items and comparison/non-explanation items

			Explanation items					
			Students			About Teachers		
			“I explain how writers use tools like symbolism and metaphor to communicate meaning.”	“I write a few sentences to explain how I solved a math problem.”	“I interpret data and explain what the results mean.”	“My teacher asks us to explain our thinking.”	“My teacher explains difficult things clearly.”	“My teacher has several good ways to explain things.”
Comparison/ Non-explanation items	Students	“I make observations or collect data outside of the classroom for assignments.”	0.189**	0.221**	0.274**	0.319**	0.320**	0.318**
		“I combine many ideas and pieces of information into something new and more complex.”	0.292**	0.257**	0.294**	0.385**	0.377**	0.382**
	About Teachers	“My teacher gives tests that require us to use different sources of information for our answers.”	0.193**	0.221**	0.212**	0.408**	0.374**	0.432**
		“My teacher helps me learn to use different sources of information.”	0.194**	0.200**	0.193**	0.391**	0.376**	0.423**

Note. Overall composite scale score for both scales correlated with one another $r = .558$ ($p < .01$).
** significant at $p < .01$.

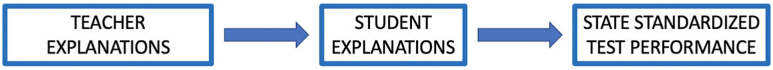

Fig. 2. Theoretical model tested in Study 2.

(see Kastens & Liben, 2007). To match the explanation items as closely as possible, the first two items reflected the students' own cognitive efforts (e.g., "I combine many ideas and pieces of information into something new and more complex."), and the second two items concerned teacher efforts (e.g., "My teacher helps me learn to use different sources of information."). These were rated on the same frequency scale used for explanation items. Together, these items produced a Cronbach's alpha of .68, demonstrating acceptable reliability. Again, students were instructed to think about their core classes when answering these questions, not one specific teacher.

3.1.2.3. State standardized test scores: State standardized test scores for math and English were obtained for all participants for 7th, 8th, and 10th grades.

3.2. *Results and discussion*

We first report descriptive statistics of our variables of interest: teacher explanations ($M =$ 9.480, $SD = 2.040$), student explanations ($M = 7.438$, $SD = 2.029$), math state standardized test scores ($M = .063$, $SD = .982$; standardized, with range of -2.947 to 1.737), English state standardized test scores ($M = .003$, $SD = .953$; standardized, with range -3.028 to 1.833), and non-explanation/comparison items ($M = 11.922$, $SD = 2.626$).

Our focal hypothesis was that more frequent explaining would predict stronger academic performance, as assessed by state standardized test scores. Study 2 additionally allowed us to test the role of teacher explanations. Specifically, we posited that teacher explanations would scaffold students' explaining, in turn predicting state standardized test performance (see theoretical model, Fig. 2), even after controlling for prior scores.

We tested this model for both math and English scores. We included additional covariates when (a) there was a priori theoretical motivation for doing so, and (b) the covariate showed a potential confounding effect. Thus, participation in deeper learning professional development was included as a covariate in all models given significant associations with explanation frequency ($t = 6.894$, $p < .001$) and state standardized test performance in math and English ($t = 6.575$, $p < .001$; $t = 5.224$, $p < .001$). We utilized structural equation modeling to estimate fully saturated models in path analyses. We utilized maximum likelihood estimation for missing data (a common statistical tool used to address implicit imputation of missing data, valid under the assumption that data are missing at random).[2] The model provided a perfect fit to the data. We entered participants' state standardized test scores from 10th grade as the dependent variable, controlling for prior scores and whether the student was in a school with deeper learning professional development.

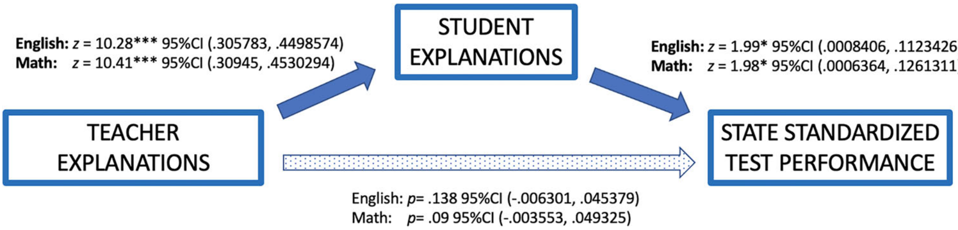
English: z = 10.28*** 95%CI (.305783, .4498574)
Math: z = 10.41*** 95%CI (.30945, .4530294)
STUDENT EXPLANATIONS
English: z = 1.99* 95%CI (.0008406, .1123426)
Math: z = 1.98* 95%CI (.0006364, .1261311)
TEACHER EXPLANATIONS
STATE STANDARDIZED TEST PERFORMANCE
English: p= .138 95%CI (-.006301, .045379)
Math: p= .09 95%CI (-.003553, .049325)

Fig. 3. Model depicting indirect, not direct, effect of teachers' explanations predicting state standardized test performance.
Note. * significant at $p < .05$, ** significant at $p < .01$, and *** significant at $p < .001$.

We first tested teacher explanations predicting student explanations in turn predicting state standardized test performance, controlling for prior scores. As seen in Table 3, the predicted path was significant in both models.

To scrutinize our hypothesis, we reran the models in alternative directions (student explanations predicting teacher explanations predicting test performance). In this order, the model was not significant for math or ELA (95% CI [−.0134111, .0333852] $p = .403$; 95% CI [−.0166654, .0309812] $p = .556$).

Next, we tested a direct versus indirect effect from the focal predictor variable, teachers' explanations. We found that teacher explanations did not have a direct link to standardized test performance after controlling for prior scores (see Fig. 3; math: 95% CI [−.0033839, .0392599] $p = .099$; ELA: 95% CI [−.0048268, .032601] $p = .146$).

These analyses support our interpretation (reflected in Table 3) that students' own explanations, predicted by teacher explanations, are what predict later test performance.

We next fit identical models, but replacing the explanation items with the corresponding comparison items to test whether the effects reported above were potentially unique to explanation, or if instead the explanation items were serving as proxies for more general cognitive enrichment. Notably, explanation items and comparison items all showed significant, consistent correlations with each other (see Table 4).

However, when we used comparison items in our models, none reached significance (see Fig. 4). This suggests that associations found for explanations do not extend to similar, correlated comparison items.

In sum, student explanations directly predict both math and English performance, after controlling for prior ability. Testing direct versus indirect effects suggests the link between teacher explanations and students' standardized scores are fully explained by students' own explanations.

4. Discussion

Two studies—one assessing explanations directly, one using self-report in an educational setting—found that explaining predicted standardized test performance. This result was

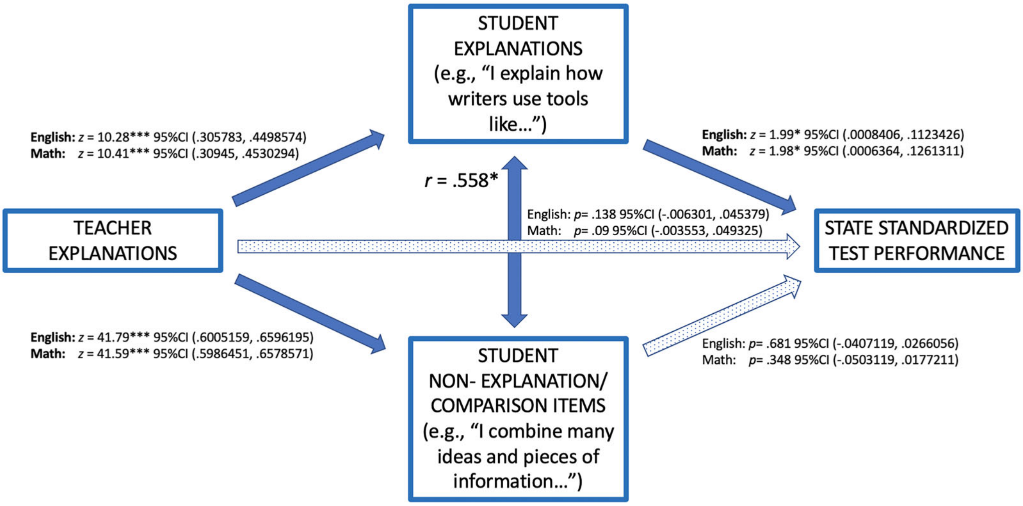
STUDENT EXPLANATIONS (e.g., "I explain how writers use tools like...")
English: z = 10.28*** 95%CI (.305783, .4498574)
Math: z = 10.41*** 95%CI (.30945, .4530294)
English: z = 1.99* 95%CI (.0008406, .1123426)
Math: z = 1.98* 95%CI (.0006364, .1261311)
r = .558*
TEACHER EXPLANATIONS
English: p= .138 95%CI (-.006301, .045379)
Math: p= .09 95%CI (-.003553, .049325)
STATE STANDARDIZED TEST PERFORMANCE
English: z = 41.79*** 95%CI (.6005159, .6596195)
Math: z = 41.59*** 95%CI (.5986451, .6578571)
STUDENT NON- EXPLANATION/ COMPARISON ITEMS (e.g., "I combine many ideas and pieces of information...")
English: p= .681 95%CI (-.0407119, .0266056)
Math: p= .348 95%CI (-.0503119, .0177211)

Fig. 4. Teacher explanations predicting student explanation versus student non-explanation items and, in turn, standardized test performance.
Note. * significant at $p < .05$, ** significant at $p < .01$, and *** significant at $p < .001$.

obtained for both math and English, suggesting domain-general effects. This result also held while controlling for prior scores, which challenges the idea that associations between explanation quality or quantity and test performance simply reflect a common cause, such as intelligence or test-taking skill. (That said, it remains possible that explanation and test performance were both shaped by some unmeasured variable, such as broader attitudes toward learning.) Finally, in Study 2, these results were found to be unique to explanation, suggesting that other forms of active student engagement do not necessarily have comparable effects. Though there are important limitations that come with utilizing brief survey questionnaires as we did in Study 2, it is striking, given the sparsity of the assessment, that we find the sizable and consistent relationships that we do.

These findings have both theoretical and practical implications. Theoretically, an effect of explanation skill on a distal academic outcome sheds light on the mechanisms by which explaining might contribute to learning over time. In Study 1, the content of the explanations elicited during the inductive learning task had no (or extremely little) shared content with the material assessed in math and English standardized tests. In Study 2, students reported the general prevalence of explanatory activities across their academic experience. It thus seems unlikely that the benefits of explanation stemmed from any *particular* content that was the target of explanation. Instead, we suspect that the measures across both studies reflected students' general tendency to engage in explanation, and that this tendency was responsible for their superior academic outcomes.

Thinking of explanation as a general strategy or metacognitive skill helps explain why effects may have been both long-lasting and domain-general. Metacognitive skills have been identified as critical predictors of long-term outcomes, specifically in relation to academic achievement (Hinshaw, 1992; Claessens, Duncan, & Engel, 2009; Modrek & Ramirez, 2021). In a mega study utilizing six datasets totaling over 50,000 children followed longitudinally in the United States and United Kingdom, metacognitive skills—specifically attention skills—were found to be one of the strongest predictors of long-term achievement, for both math and reading (Duncan et al., 2007). Metacognitive skills are also known to transfer (Kornell, Son, & Terrace, 2007): because they are domain-general, they support potential applications across contexts, subjects, and tasks (Kuhn & Modrek, 2023; Kuhn & Modrek, 2021; Modrek & Sandoval, 2020; Stebner et al., 2022).

While prior work has not investigated the distal impacts of explanation on academic achievement as we do here, it does support the idea of explaining as a metacognitive skill (Chi, 2000; Chi & Bassok, 1989). Fergusson-Hessler and de Jong (1990) investigated the study strategies of high- and low-achieving students and found that while both categories of students engaged in an equal number of study processes, the high-achieving students tended to use metacognitive strategies (such as explaining relationships), whereas low-achieving students were more likely to use superficial processing (e.g., restating). Research also finds that when learners are prompted to explain, it engages their metacognitive awareness (Mills & Keil, 2004; Renkl, 1997; Rozenblit & Keil, 2002).

At a practical level, these results have potential implications for education. Both studies suggest that increasing the frequency of student explanation could have beneficial effects, consistent with prior work (Ing et al., 2015; Kuhn, Modrek, & Sandoval, 2020). Going beyond prior work, our results suggest that the benefits could extend to distal academic achievement. Study 1 additionally suggests that explanation quality might matter, and Study 2 offers hints about how explanatory frequency might be achieved—not only by explicitly prompting students to explain, but by having teachers model explanation. In thinking about implementation, it will be important to identify boundary conditions on these effects (e.g., Kuhn & Katz, 2009; Williams, Lombrozo, & Rehder, 2013; Legare & Lombrozo, 2014), and to consider the opportunity cost (e.g., Schwartz & Martin, 2004; Schworm & Renkl, 2019): if teachers and students spend more time explaining, what activities might be displaced?

A broader agenda of the present work was to identify contributors to distal outcomes. Psychological scientists long identified self-control as a longitudinal predictor of standardized test scores (Mischel et al., 2011; Shoda, Mischel, & Peake, 1990; Mischel, 2004)—a conclusion disputed by more recent evidence suggesting self-control does not have a robust or direct effect on standardized test performance (Watts, Duncan, & Quan, 2018). For instance, with national samples comprising nearly 20,000 adolescent participants, Baldwin et al. (2022) found that self-control had no direct relation to state standardized test performance. They found it was *strategies* that directly predicted state standardized test performance (e.g., time spent practicing), and that self-control's relation to standardized test performance was fully explained by such strategies. In the current project, we shed light on mechanisms driving effective learning that contribute to academic outcomes over time. Specifically, our findings suggest long-term benefits of fostering explanation skills in learners.

Acknowledgments

This work was funded by The William and Flora Hewlett Foundation and supported by the American Educational Research Association (AERA) Deeper Learning Fellowship. The views and opinions expressed herein are those of the authors only, and do not represent the official views and opinions of The William and Flora Hewlett Foundation and AERA. We thank the American Institute for Research (AIR) for their research collaborations. We thank Emily Karras, Zumana Noor, and Marina Campbell from our research team, Jillian Anderson for her assistance with scoring, and all the teachers and schools who participated in this work. We are grateful for the Micro-Sabbatical program at Princeton University's Department of Psychology for supporting the work presented here.

Compliance with ethical standards

The authors declare that they have no conflict of interest. The rights of the participants were protected, and applicable human research subject guidelines were followed in this research. All participants provided informed consent.

All study procedures and instruments were approved by Columbia University's and American Institutes for Research's institutional review boards (IRB) and the school district(s) prior to participant recruitment and data collection.

Notes

[1] Note that the explanation items were not derived from a previously validated measure; instead, they were pulled from a list of items given to participants to assess activity frequency. The items were all highly, significantly correlated with one another (see Table 4). When we originally pulled all six items that pertained to explanations, they produced a Cronbach's alpha of .72, suggesting they captured explanatory frequency reliably. However, when we divided the items depending on whether they concerned student or teacher explanations, we did find a lower alpha for the three student explanation items of .54. Given that we still had theoretical reasons to group these items, and that they are significantly, highly correlated with one another, our primary analyses still focus on a composite score of the three student explanation items. However, to ensure that this composite did not obscure meaningful variation with respect to our hypotheses, we ran exploratory analyses on each model with individual student item explanation scores, and found that each item alone (e.g., "I interpret data and explain what the results mean.") replicated our findings with the composite score.

[2] In an exploratory analysis, we tested whether imputation was affecting the results. In this analysis, we dropped all participants who were missing even one item from the survey response and reran all our models on this subset of the sample. We found the same results, including consistent results for both math and English. These exploratory analyses increase our confidence that imputation is not meaningfully influencing our results.

References

Ainsworth, S., & Loizou, A. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, *27*, 669–681.

Aleven, V., & Koedinger, K. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*, 147–179.

Astington, J. W., & Gopnik, A. (1991). Theoretical explanations of children's understanding of the mind. *British Journal of Developmental Psychology*, *9*(1), 7–31.

Baldwin, C. R., Haimovitz, K., Shankar, P., Gallop, R., Yeager, D., Gross, J. J., & Duckworth, A. L. (2022). Self-control and SAT outcomes: Evidence from two national field studies. *PLos One*, *17*(9), e0274380.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction*, *13*(2), 221–252.

Blair, C., & Raver, C. C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. *PLoS One*, *9*(11), e112393.

Chi, M. T. (2018). Learning from examples via self-explanations. In L. Resnick (Ed.), *Knowing, learning, and instruction* 251–282. Routledge.

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M. T. H., & Bassok, M. (1989). Learning from examples via self-explanations. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 251–28).

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*(4), 415–427.

Chuey, A., McCarthy, A., Lockhart, K., Trouche, E., Sheskin, M., & Keil, F. (2021). No guts, no glory: Underestimating the benefits of providing children with mechanistic details. *npj Science of Learning*, *6*(1), 30.

Dean, Jr, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91*(3), 384–397.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428.

Duckworth, A. L., Kirby, T. A., Gollwitzer, A., & Oettingen, G. (2013). From fantasy to action: Mental contrasting with implementation intentions (MCII) improves academic performance in children. *Social Psychological and Personality Science*, *4*(6), 745–753.

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38.

Fonseca, B. A., & Chi, M. T. (2011). Instruction based on self-explanation. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* 310–335. Routledge.

Fergusson-Hessler, M., & de Jong, T. (1990). Studying physics texts: Differences in study processes between good and poor performers. *Cognition and Instruction*, *7*, 41–54.

Goddu, M. K., Lombrozo, T., & Gopnik, A. (2020). Transformations and transfer: Preschool children understand abstract relations and reason analogically in a causal task. *Child Development*, *91*(6), 1898–1915.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system.

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence*, *41*(5), 579–596.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin*, *111*(1), 127.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Ing, M., Webb, N. M., Franke, M. L., Turrou, A. C., Wong, J., Shin, N., & Fernandez, C. H. (2015). Student participation in elementary mathematics classrooms: The missing link between teacher practices and student achievement? *Educational Studies in Mathematics*, *90*, 341–356.

Kastens, K. A., & Liben, L. S. (2007). Eliciting self-explanations improves children's performance on a field-based map skills task. *Cognition and Instruction*, *25*(1), 45–74.

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, *18*(1), 64–71.

Kuhn, D., & Modrek, A. (2023). The broad reach of multivariable thinking. *Informal Logic*, *43*(1), 1–22.

Kuhn, D., & Modrek, A. (2021). Mere exposure to dialogic framing enriches argumentive thinking. *Applied Cognitive Psychology*, *35*(5), 1349–1355.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, *103*(3), 386–394.

Kuhn, D., Modrek, A. S., & Sandoval, W. A. (2020). Teaching and learning by questioning. In L. Butler, S. Ronfard, & K. Corriveau (Eds.), *The questioning child: Insights from psychology and education* (pp. 232–251). Cambridge, UK: Cambridge University Press.

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology*, *49*(3), 446.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., Kross, E., Teslovich, T., Wilson, N. L., Zayas, V., & Shoda, Y. (2011). 'Willpower' over the life span: decomposing self-regulation. *Social Cognitive and Affective Neuroscience*, *6*(2), 252–256.

Mischel, W. (2004). Toward an integrative science of the person. *Annu. Rev. Psychol.*, *55*(1), 1–22.

Modrek, A. S., & Ramirez, G. (2021). Cognitive regulation outdoes behavior regulation in predicting state standardized test scores over time. *Metacognition and Learning*, *16*(1), 113–134.

Modrek, A. S., & Sandoval, W. A. (2020). Can autonomy play a role in causal reasoning? *Cognitive Development*, *54*, 100849.

Modrek, A. S., Kuhn, D., Conway, A., & Arvidsson, T. S. (2019). Cognitive regulation, not behavior regulation, predicts learning. *Learning and Instruction*, *60*, 237–244.

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

McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* 397–421. Psychology Press.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87*(1), 1–32.

Moreno, R., & Mayer, R. E. (2000). A learner-centered approach to multimedia explanations: Deriving instructional design principles from cognitive theory. *Interactive Multimedia Electronic Journal of Computer-enhanced Learning*, *2*(2), 12–20.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*, 1–29.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Ruggeri, A., Xu, F., & Lombrozo, T. (2019). Effects of explanation on children's question asking. *Cognition*, *191*, 103966.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Schworm, S., & Renkl, A. (2019). Learning by solved example problems: Instructional explanations reduce self-explanation activity. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 816–821). Routledge.

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology*, *26*(6), 978.

Stebner, F., Schuster, C., Weber, X. L., Greiff, S., Leutner, D., & Wirth, J. (2022). Transfer of metacognitive skills in self-regulated learning: Effects on strategy application and content knowledge acquisition. *Metacognition and Learning*, *17*(3), 715–744.

Vasil, N., Ruggeri, A., & Lombrozo, T. (2022). When and how children use explanations to guide generalizations. *Cognitive Development*, *61*, 101144.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*(1), 229–246.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, *29*(7), 1159–1177.

Webb, N. M. (1982a). Peer interaction and learning in cooperative small groups. *Journal of Educational Psychology*, *74*(5), 642.

Webb, N. M. (1982b). Group composition, group interaction, and achievement in cooperative small groups. *Journal of Educational Psychology*, *74*(4), 475.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*(1), 55–84.

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*(4), 1006.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. School demographics.