



Asking young children to “do science” instead of “be scientists” increases science engagement in a randomized field experiment

Marjorie Rhodes^{a,1}, Amanda Cardarelli^a, and Sarah-Jane Leslie^b

^aDepartment of Psychology, New York University, New York, NY 10003; and ^bDepartment of Philosophy, Princeton University, Princeton, NJ 08544

Edited by Eva Pomerantz, University of Illinois at Urbana–Champaign, Champaign, IL, and accepted by Editorial Board Member Renée Baillargeon March 10, 2020 (received for review November 8, 2019)

Subtle features of common language can imply to young children that scientists are a special and distinct kind of person—a way of thinking that can interfere with the development of children’s own engagement with science. We conducted a large field experiment (involving 45 prekindergarten schools, 130 teachers, and over 1,100 children) to test if targeting subtle properties of language can increase science engagement in children’s daily lives. Despite strong tendencies to describe scientists as a special kind of person (in a baseline control condition), brief video-based training changed the language that teachers used to introduce science to their students. These changes in language were powerful enough to predict children’s science interest and behavior days later. Thus, subtle features of language shape children’s beliefs and behaviors as they unfold in real world environments. Harnessing these mechanisms could promote science engagement in early childhood.

cognitive development | generic language | science education

We often speak of scientists as a special and distinct kind of person. Commonplace descriptions such as, “You’re thinking like a real scientist,” “Scientists think about problems and get ideas to solve them,” or “A great scientist wouldn’t let a problem get the best of him” contain two linguistic cues—category labels and generic descriptions of categories—that can imply to a young child that being a scientist is a special kind of identity (1–6). These quotes are from popular children’s television shows, where over half of references to science include at least one of these linguistic markers (7). Here we considered that this commonplace tendency to speak about scientists as a special kind of person might have detrimental consequences for the early development of children’s science beliefs and behaviors. To evaluate this possibility, we tested whether changing how we talk about science can increase science engagement in early childhood.

The quotes above sound fun and engaging—they certainly do not include exclusionary or discouraging content. Yet, while the content of these sentences does not sound problematic, their linguistic form still could be. This is because children expect noun labels (in this case, “scientist”) to pick out categories that are stable and distinct (3, 5, 8). Further, when children hear groups described in abstract ways that generalize over the entire category—as in “Scientists discover new things”—they assume the categories contain people who are fundamentally similar to one another and different from others (4, 6, 9–11).

Thus, these linguistic markers could lead children to think that only some people can be scientists, that people who are scientists are fundamentally different from those who are not, and that whether one is a scientist is fixed and stable (12, 13). This categorical way of thinking about scientists can then become problematic for children because it invites the question of whether they themselves are members of the scientist group (14, 15). Indeed, introducing children (ages 4–9) to science with identity-cuing linguistic markers in laboratory studies leads to less immediate subsequent persistence in science activities, compared to introducing science with action-oriented language

(e.g., “Let’s do science! Doing science means exploring the world and discovering new things”; ref. 14). Among older children (ages 7–9), identity-cuing language also leads to less science interest and lower science self-efficacy (15).

Here we tested whether it is possible to harness these linguistic mechanisms to increase child engagement in science in daily life. For this to be the case, it must be possible for adults to change the way they talk, despite the prevalence of identity-cuing language about science (7, 16). Further, language must have sufficiently powerful effects on children’s beliefs and behaviors to cut through other aspects of children’s experience and extend across time. Testing the immediate consequences of language in scripted laboratory experiments cannot reveal if linguistic mechanisms are powerful enough to shape development as it unfolds in daily life, when both the nature of the input and the contexts in which it is delivered are considerably more variable. Therefore, we ran a field experiment in a large public prekindergarten program. We randomly assigned teachers to a brief video-based training to encourage more action-focused language or to a standard-practice baseline control. We then recorded and analyzed the language that teachers used to teach about science in their classrooms and tested for the consequences on children’s science beliefs and behaviors several days later.

Results

To avoid contamination of the language manipulation across neighboring classrooms, we randomly assigned teachers to

Significance

Language implying that scientists have a special kind of identity (e.g., “Let’s turn on our special scientist brains!”) is prevalent in input to young children and has immediate negative consequences for children’s science behavior in laboratory studies. To test if these effects of language are powerful enough to shape child behavior as it unfolds in the natural course of development, we conducted a large field experiment with prekindergarten teachers and their students. Brief video-based training led teachers to change their language and increased children’s science persistence several days later but did not affect children’s feelings of science self-efficacy. These data reveal tools that could be used to increase science engagement in daily life.

Author contributions: M.R., A.C., and S.J.L. designed research; A.C. performed research; M.R. analyzed data; and M.R., A.C., and S.J.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. E.P. is a guest editor invited by the Editorial Board.

Published under the [PNAS license](#).

Data deposition: All data for this study are available on the Open Science Framework: <https://osf.io/pe7k5/>.

¹To whom correspondence may be addressed. Email: marjorie.rhodes@nyu.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1919646117/-DCSupplemental>.

condition at the level of the school in which they taught (Baseline Control = 23 schools, 62 teachers, 565 children; Experimental = 22 schools, 68 teachers, 582 children; all from a free public prekindergarten program serving children beginning in the calendar year when they turn four). The schools were drawn from 11 different districts run by a central department of education in a large city; the districts varied geographically and with respect to the economic and other demographic characteristics of their surrounding neighborhoods. Therefore, random assignment of schools to conditions was stratified by district. Across the sample, the schools that participated were comprised of children who were racially and ethnically diverse (~36% Hispanic, 22% Asian, 19% White, 15% Black, 8% other or unknown) and who came from diverse economic backgrounds [participating districts ranged from including ~50–90% of families who were eligible for social programs based on economic disadvantage (17)]. Demographic features of participating schools and teachers (based on available data) did not vary by condition (*SI Appendix*).

For the field experiment, we asked teachers to teach a new lesson about friction to their prekindergarten classes. In the lesson, children learned about friction by experiencing how cars travel down a ramp at different speeds when the ramp is covered in different materials (e.g., more slowly on carpet and more quickly on wrapping paper). Teachers received a brief training video that showed how to set up and implement the lesson with their class. If their school was assigned to the experimental condition, the video included a teacher who modeled action-focused language—for example, who said to the class in the video, “Today we are going to do science! The first part of doing science is observing with our senses”). The video provided explanation of the importance of describing how to do science (by observing, making guesses, and checking) and examples of the teacher implementing the lesson using action-focused language. In the control condition, the video showed how to set up and give the lesson, but the teacher did not model particular language and the video did not emphasize the importance of describing the process of science to the same degree. Because the video in the control condition did not include these examples, it was briefer than the video in the experimental condition (for details, see *Materials and Methods*). In the control condition, teachers were not instructed to use any particular language in the lesson, and thus, we expected them to use whatever language felt natural to them.

After watching the video (between 0–4 d later), teachers gave the lesson to their class while wearing an audio recorder and we transcribed and coded these transcripts for the linguistic markers of interest. The training was successful: Teachers produced fewer identity-references to scientists, odds ratio (OR) = 0.19, CI = 0.07, 0.53, $z = -3.2$, $P = 0.0014$, and more action-based descriptions of science, OR = 44.43, CI = 21.93, 90.04, $z = 10.53$, $P < 0.001$, in the experimental than the control condition (Fig. 1). As shown in Fig. 1, as proportions of total relevant references to science, over 75% of teachers’ descriptions of science were identity-references in the control condition (and less than 25% were action-based), but in the experimental condition, ~90% of relevant references were action-based, and less than 10% were identity-references. Note that we did not explicitly tell teachers in the experimental condition not to use identity-cuing language, nor did we encourage teachers in the control condition to use it. Rather, identity-cuing language appeared by standard practice, but was easily inhibited in the experimental condition simply by modeling an alternate way to speak.

Approximately 3 d after the lesson, researchers visited the children’s classrooms to assess the children’s science beliefs and behaviors. The measures were set up on tablets in the classroom. Approximately half of the children who completed study measures ($n = 545$; 255 control, 290 experimental) completed a measure of science persistence, in particular how long they chose

to persist on a science task thematically related to the target lesson, which we considered to be a behavioral measure of science engagement. This task (similar to ref. 14) presented children with a tablet-based video game, which children could navigate entirely on their own. In the game, children were asked to predict how far a car would go down a ramp, using the information they had learned a few days earlier. As in previous studies, the game was rigged: Children were shown the outcomes of their first two guesses: They always got their first guess right and their second guess wrong. We then measured how long children chose to continue persisting on this game after the wrong guess (how many more trials out of six they chose to play before choosing to stop and do something else in their classroom). No feedback as to the accuracy of the child’s guess was given on the six test trials. We selected this measure because young children need to practice new skills in order to develop them; therefore, measures of behavioral engagement are important dimensions of classroom behavior in early childhood [as longer engagement provides more opportunities to learn (18)] and predict achievement across various samples, ages, and domains (19).

Teachers’ identity-based references to scientists during the lesson predicted less behavioral persistence on the science game 3 d later (OR = 0.92, CI = 0.85, 0.98, $z = -2.48$, $P = 0.013$; Fig. 2). Further, there was an overall causal effect of treatment assignment in an intention-to-treat analysis. Children who attended a school assigned to the experimental condition showed more task persistence than children assigned to the control condition, OR = 1.44, CI = 1.01, 2.06, $z = 2.02$, $P = 0.04$; see Fig. 3. Effects on behavior were reliable but small in magnitude: Children in the experimental condition were about 6 percentage points more likely to persist on the first optional trial, and about 4 percentage points more likely to persist past the halfway point on the task, than children in the control condition. Although this effect is small in magnitude, it provides a conservative estimate of the effects of language on children’s persistence as this analysis does not account for variation in teacher language or child attention or participation during the target lesson, and outcomes were measured in a naturalistic setting several days later. Further, this provides a conservative estimate of the potential impact of language-focused interventions more generally, since the present field experiment targeted only a single lesson.

The other half of children who completed the study measures were asked self-report assessments of their science interest and efficacy. Linguistic cues shape these variables in older children (15), but had not been included in previous research on the effects of language on science behaviors among children this young. Teachers who produced more identity-references to scientists had students who expressed less interest in science, OR = 0.96, CI = 0.93, 0.99, $t = -2.53$, $P = 0.01$ (Fig. 4). However, there was no causal effect of condition on children’s interest in science in an intention-to-treat analysis and no effects of teacher language or condition-assignment for children’s self-efficacy. Children responded very positively to these measures across the board (the median response was the highest scale point for all scale measures), as is common in children’s self-reports of their own capabilities (20, 21). While it is possible that language affects children’s science behavior but not their beliefs or attitudes in early childhood, it is also possible that the measures we selected were not sufficiently sensitive to provide a powerful test of these variables in children this young.

Discussion

We found that although speaking about science by emphasizing the identity category of scientists was very common in our control condition, teachers could easily change the way they talk to adopt a more action-oriented approach. Further, when teachers used less identity-cuing language, children in their classes

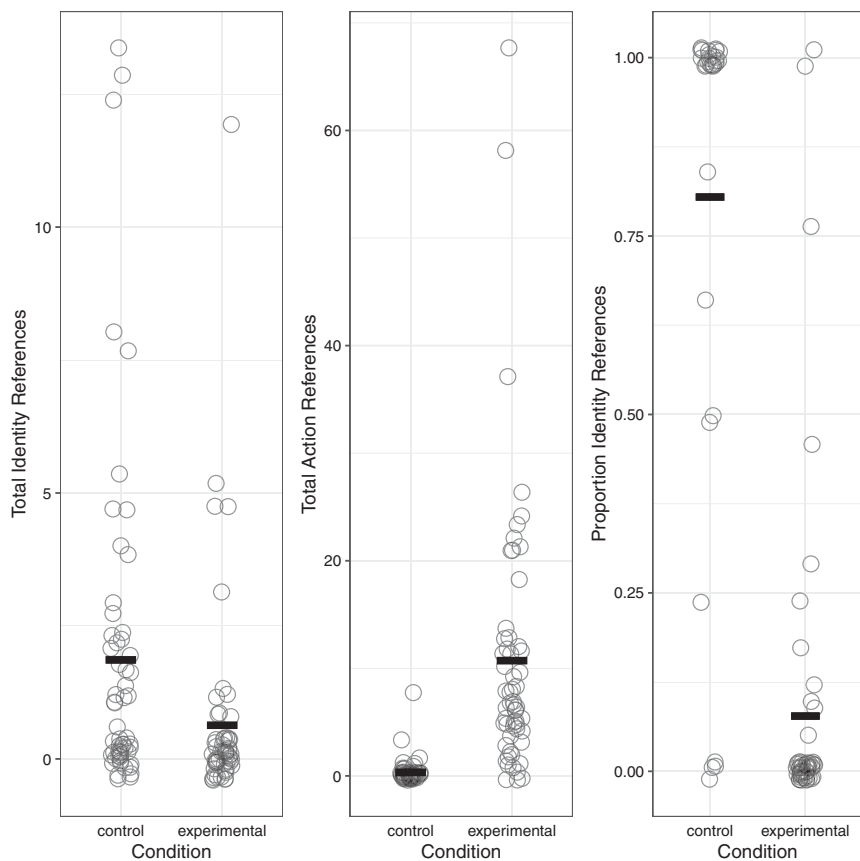


Fig. 1. (Left) The total number of identity-references produced by teachers, which included both the use of category labels (e.g., “Today are we are going to be scientists”) and generic claims (e.g., “Scientists have a really cool job”); Control, $M = 1.86$, $SD = 3.18$; Experimental, $M = 0.63$, $SD = 1.94$). (Center) The total number of action-based descriptions of science (e.g., “Today we are going to *do* science to learn about friction”); Control, $M = 0.33$, $SD = 1.16$; Experimental, $M = 10.72$, $SD = 12.7$). (Right) The proportion of teacher language that used identity cues out of the total identity-based or action-focused language (Control, $M = 0.80$, $SD = 0.36$; Experimental, $M = 0.08$, $SD = 0.22$). For each figure, the circles represent the scores of individual teachers, and the bars represent the means by condition.

showed more science persistence and interest several days later. Thus, changing a subtle but powerful feature of the way we speak about science could increase science engagement in early childhood. We suspect that the positive consequences of the action-focused language stem primarily from it serving as a comfortable alternative to an identity-cuing way of speaking; it is possible that action-focused language also provides its own unique benefits as an approach for early science education, but unique positive effects of action-focused language (beyond serving as a substitute for identity-cuing speech) were not documented in the present data. This project targeted a single lesson, and the effects on child behavior (several days later) were modest in magnitude and scope (influencing children’s persistence behaviors but not their self-evaluations, for example). The present work shows a proof-of-concept that relatively subtle features of language are powerful enough to shape development in children’s daily lives, opening the door to examine how important these effects are in explaining variation in development across time. For instance, future work will need to examine if more sustained input (e.g., over the course of a school year) leads to accumulating effects over time.

To maximize participation, children completed study measures anonymously, and we did not have access to demographic information for individual children. Thus, while the present sample was diverse with respect to race, ethnicity, and economic background, one limitation of the present approach is that we were unable to test if the effects of the intervention varied by features

of children’s identities and background. We suggest that identity-cuing language becomes particularly problematic for children when they have reason to question if they are members of the referenced group. This means that language will interact with other aspects of children’s beliefs and experience, as such “reasons to question” can come from children’s own experiences of difficulty, other features of the input that children receive, as well as from social stereotypes about what group members are typically like. All of this makes identity-cuing language problematic for science, for several reasons. First, experiencing difficulty and setbacks is an inherent part of science. Second, children receive explicit input suggesting that scientists are special and unique [e.g., that there is such a thing as a “special science brain” (7)] which could give them reason to doubt if they have the necessary qualities to succeed. Third, social stereotypes about what scientists look like are early developing and pervasive (22, 23), providing “reasons to question” particularly for children from underrepresented groups. Indeed, previous laboratory studies indicate that the negative implications of identity-cuing language extend broadly in early childhood, but become especially problematic among older children from underrepresented groups (14, 15).

Thus, identity-cuing language—currently the most common way of presenting science to young children—undermines science persistence in early childhood (as shown here) and could also feed into developmental trajectories that contribute to social disparities in science achievement over time. Further, this

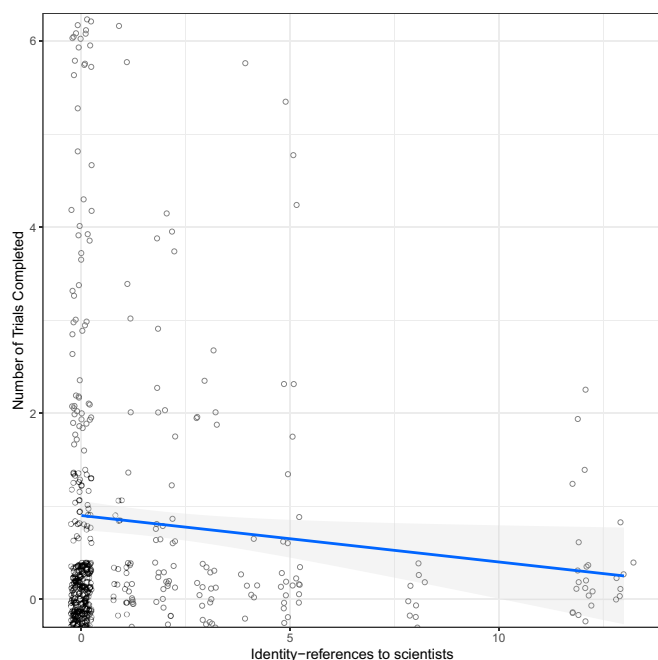


Fig. 2. Shows that the number of identity-references produced by teachers related negatively to children’s persistence. Circles depict the responses of individual children, and the blue line shows the predicted values from the regression model described in the text, with the 95% CI shown in the shaded region. We conducted a similar analysis predicting persistence from action-based language; this slope went in the other direction, but was not significant (OR = 1.02, CI = 1, 1.04, $z = 1.74$, $P = 0.082$).

framework for how language shapes the development of children’s beliefs and behaviors suggests that similar processes could also operate in other domains of development, particularly when language interacts with other features of children’s experience to lead them to doubt their membership in an important group (24–26). Although disengagement from science is a multifaceted problem with no single solution, the present study suggests that a fairly simple change to the way we speak could perhaps prevent young children from disengaging before they have the opportunity to begin doing science.

Materials and Methods

This study was reviewed and approved by the Institutional Review Boards of New York University and the New York City Department of Education. Teachers provided written consent for their participation. Families of children in participating schools received a letter describing the project and had the opportunity to ask for their child to not be included in the project. Because the study was conducted as part of regular classroom activities, researchers did not interact directly with children, and no identifying information on children was collected, we received a waiver of the need to document informed consent from families.

The following can be found in the Open Science Framework repository for this project (<https://osf.io/pe7k5/>): 1) the training videos shown to teachers in both conditions, 2) the printed lesson plans given to teachers in both conditions, 3) a brief prelesson survey given to teachers, 4) animated versions of all study measures given to children, 5) the postlesson survey given to teachers, 6) the coding guide used to code teacher language data, 7) our analysis plan (posted in advance of data analysis), 8) all data files and code used for analyses. For more information on the participants and inclusion criteria, see *SI Appendix*.

Procedure. Once all of the prekindergarten schools signed up to join the Science Initiative, the schools were randomly assigned to either the control or the experimental condition. The sample was randomized at the level of the school, not at the level of individual teacher, to avoid possible contamination between the different conditions across neighboring classrooms (e.g., if

teachers from neighboring classrooms happened to observe each other’s lessons or materials). The random assignment of schools to condition was stratified by district, as the schools came from 11 districts across a large city that varied with respect to their demographic features (*SI Appendix*). Teachers were sent a brief training video to view online, which varied by condition, the Thursday before the week in which they were to implement the target lesson (Fig. 5). In both conditions of the video, a model teacher demonstrated how to implement a science lesson with a prekindergarten class, which introduced the concept of friction by having children observe and test how fast toy cars move down a ramp when the ramp is covered in different materials—for example, more slowly when covered with carpet, more quickly when covered in wood. The video for teachers in the control condition showed how to set up and implement the lesson but did not model any specific language or examples of implementation. For the experimental condition, the video contained numerous examples of action-focused language (e.g., “Today we are going to do science. Let’s start doing science by using our hands to observe these materials”) and showed the model teacher giving parts of the lesson to a small group of children. We used the training video in the experimental condition to model action-focused language, but did not provide a direct script or explicitly tell teachers not to use identity-focused language. We made these choices because we thought a more direct approach might not give teachers sufficient autonomy to teach the lesson as they felt comfortable and because we thought it might be more complicated to ask teachers to both remember what to say (e.g., action-focused language) and to remember what not to say (e.g., identity-cuing language).

After watching the video, teachers completed several comprehension questions to help them encode the material. Teachers could watch the video anytime between when it was sent (on Thursdays) and when they taught the lesson (Monday or Tuesday of the following week). We received confirmation that most teachers (87%) watched the video successfully; the remainder relied on the printed lesson plan alone (which also included the condition-manipulation), or perhaps watched the video along with a teacher from a neighboring classroom (in which case, we would not have received a confirmation of them watching the video). To provide a conservative test of the

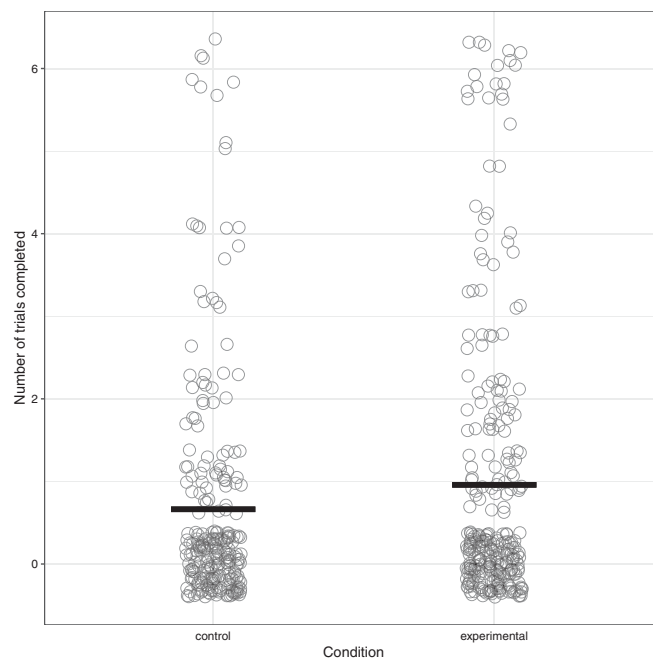


Fig. 3. Shows the effect of condition-assignment on children’s behavior in the science persistence task several days later. These data are from an intention-to-treat analysis and thus estimate the causal effect of condition-assignment, regardless of teacher language production or child attention or participation in the lesson. As described in the text, children in the experimental condition persisted longer on the task than children in the control condition (numbers of trials completed: Experimental, $M = 0.96$, $SD = 1.69$; Control, $M = 0.66$, $SD = 1.35$). Circles represent the responses of individual participants and dark lines represent the mean number of trials completed by condition.

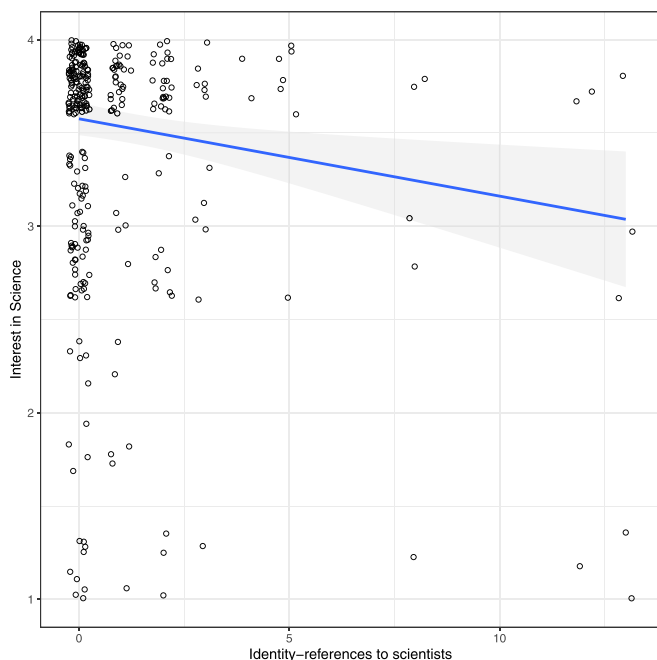


Fig. 4. Shows that the number of identity-references produced by teachers related negatively to children's interest in science. Circles depict the responses of individual children, and the blue line shows the predicted values from the regression model described in the text, with 95% CI (shown in the shaded region).

causal effect of condition-assignment, analyses included all teachers and classrooms, regardless of whether we received confirmation that the teacher had watched the training video (as is common in intention-to-treat tests of field-based interventions).

Teachers were also provided with all of the physical materials to teach the lesson to their classes. This pack of materials included a ramp to attach different materials to, seven different materials to observe and test the cars on, a rainbow mat to see the distance the cars traveled as they rolled down the ramp, a lesson plan that reiterated the contents of the instructional video (which was condition-specific), and a poster that the teacher could use to write the children's observations, predictions, and results on. Additionally, teachers were given a wearable audio recorder and asked to audio record the lesson. Approximately 0–4 d after watching the training video (depending on when they chose to watch it—some teachers watched it right before they taught the lesson, some watched it days before and then again the day before, and so on), teachers were asked to teach and audio record the target lesson in their classrooms on either the Monday or Tuesday of the specified week of implementation. Ninety-one percent of teachers successfully recorded their lessons—the remainder were evenly split in whether they did not record because they had technical difficulties, such as forgetting to turn on the recorder, or did not feel comfortable recording themselves.

Once the audio recorders were collected, the recordings were transcribed by research assistants. As described earlier, the training videos differed by condition in whether the teacher modeled action-focused language and whether the video included examples of the teacher teaching the lesson. Because the experimental video included these additional examples, it was longer than the video in the control condition, although both were relatively brief (~3 min in the control condition and 6 min in the experimental condition). Despite these differences in the length of the training video, the length of the lesson that teachers taught in their classrooms did not differ by condition (mean (M) = 20.34 min, SD = 8.14 min; M control condition = 20.52 min, M experimental condition = 20.16 min).

Research assistants blind to condition coded the language produced by the teachers line-by-line for the use of science language into four different categories, including 1) identity labels and generic statements about scientists ("Today we are going to be scientists!"; "Put on your scientist hat"; "Scientists work hard to solve problems") (M proportion of all science statements = 0.22, SD = 0.36), 2) action-focused terminology about doing

science ("Today we are going to be doing science!"; "Doing science is to use your senses") (M = 0.44, SD = 0.43), 3) other phrases mentioning science but not describing the activity (most of these referred to times or places, as in "We will continue the activity in the science center"; "It's science time!") (M = 0.32, SD = 0.37), and 4) use of the word, "scientific" ("Today we have a scientific lesson about friction") (M = 0.04, SD = 0.2). There was strong agreement between the language coders (Cohen's kappa = 0.88 for interrater reliability), and all discrepancies were decided upon by a third senior researcher.

Also, to explore possible implications of differences in the training videos across conditions with regard to the examples provided by the model teacher for the various components of the lesson, the research assistants also coded the transcripts for whether the classroom teachers introduced and explained the concept of friction and the concepts presented as part of the scientific process: observing, predicting, and checking. There was almost perfect agreement between the coders (Cohen's kappa = 0.94). There were no differences by condition in how often teachers introduced and explained the concepts of friction, OR = 0.94, CI = 0.46, 1.89, z = -0.38, P = 0.70 (teachers provided an average of one explanation of friction in both conditions), but teachers in the experimental condition were more likely to introduce and explain the components of the scientific process emphasized in the lesson (observing, predicting, and checking), OR = 3.21, CI = 2.47, 4.15, z = 8.77, P < 0.001 (Experimental, M = 4.03, CI = 3.55, 4.58; Control, M = 1.26, CI = 1.0, 1.58). Neither of these indicators of the content of the teaching (teachers' descriptions of friction or of the processes of observing, predicting, and checking) related to children's behavior on the persistence task, their science interest, or any of the other measures of child outcomes (in contrast to our key findings, as reported earlier, that science behavior and interest were related to the form of teachers' language). We acknowledge, however, that the lessons across the two conditions could have differed in other ways not captured by this coding scheme.

Approximately 2–4 d after the target lesson, on either the Thursday or Friday of the same week of implementation, researchers visited each classroom to assess the children's science engagement and interest in science. To do this, researchers presented the students with touchscreen tablet computers that contained a video-game version of the friction lesson their teachers had taught earlier that week. This "friction game" had two separate sets of tasks that were presented to individual children randomly—one game contained a persistence task (a measure of behavioral engagement), the other game presented children with a series of tasks that measured science interest, science self-efficacy, science knowledge, and science exclusivity beliefs (see *Measures*). After they completed all of the study activities, teachers were asked to fill out a brief survey to provide feedback on their experiences, including how well they thought the lesson went. We did not find any condition-differences on any of the teachers' postlesson measures.

Measures. All dependent measures were presented to students on touchscreen tablet computers using animations so that when children wore headphones, they could hear all of the instructions and click on the touchscreen to give their responses. In this way, children self-administered the study measures (with researchers supervising to start the games and



Fig. 5. The setup of the target lesson, taken from the training video.

make sure that each child only participated once). The tablets were set up on a table within children's classrooms, and all children in participating classrooms were invited to visit the table and do the measures as part of a typical classroom morning in which children circulate through various table-based activities. We made the decision to present the measures in a self-administered manner and as part of regular classroom activities, instead of by interviewing children one-by-one outside the room (as is commonly done in experimental studies with children this young) in order to maximize participation. Indeed, with this method, we were able to test over 1,100 children in a single school year. If we had interviewed children one-by-one, then we would only have been able to work with children whose families were comfortable with this more intensive form of research (which would have required missed class time, individual interaction with a research assistant they did not know, and providing their child's name to researchers). We expect that if we had taken this approach, our sample would have been less representative of the populations served by the schools. However, the choice to do this classroom-based, independent, and anonymous administration of study measures with such young children required certain trade-offs in the design of our measures. For example, children had less help and support completing the measures and understanding the response scales than is common in studies using similar items with children this young (for example, researchers interviewing children individually often provide individualized feedback and training in how to use the scales). Given these constraints, the limited time available for assessment, and the desire to measure a range of constructs, we measured many of our self-report beliefs and attitude variables with single items and relatively limited response scales (e.g., 4-item instead of 6-item). Thus, while our measure of behavioral persistence was set up quite similarly to previous work (and is itself a more psychologically concrete task) our measures of children's self-reported beliefs and attitudes were likely less sensitive than measures of these constructs used in other related research (25, 27, 28).

Measures were split into two different "friction games"—one of which was presented at random to each individual child. One "friction game" assessed behavioral engagement using a persistence task, modeled after the previously successful structure of games (14). The other "friction game" assessed science interest, science self-efficacy, content learned from the lesson, and science exclusivity beliefs. The two friction games can be viewed on Open Science Framework at <https://osf.io/pe7k5/>.

Science persistence. Persistence: Children played a video game that focused on a target concept of friction. Children saw a ramp covered in a particular material and heard a description of the texture (e.g., as smooth, rough, or bumpy, and so on). Children were then asked to guess how far they thought the car would go down the ramp covered in that material. The game was rigged so that they automatically made one correct guess followed by one incorrect guess. Then, we tested how long children chose to keep playing the friction game (with a possibility of six additional trials) after making this incorrect guess, as a measure of their task persistence (a critical component of behavioral engagement, previously found to be sensitive to linguistic cues in laboratory studies of children in this age group). No feedback as to the accuracy of their guess was given on the six test trials. After each trial, children were asked (by the narrator in the game), "Do you want to keep playing the science game, or do something else?" Scores ranged from 0 to 6 additional trials. The primary hypothesis in our preregistered analysis plan was that children in schools assigned to the experimental condition would show more persistence on this task.

Interest/content game. Interest in Science: Children were shown a scale (15) containing four faces and asked: "How much do you like science? Do you not like science? Do you like science a little? Do you sort of like science? Or do you like science a lot?" Responses were scored from 1 to 4. Responses to this item were high across the board (median = 4) and did not vary by condition in the intention-to-treat analysis (Control, $M = 3.54$, $CI = 3.39$, 3.68 ; Experimental, $M = 3.49$, $CI = 3.35$, 3.64), but did vary by teacher language production, as reported above.

Science Self-Efficacy: Using the same scale, children were asked: "How good are you at science? Are you: Not good at science? A little good at science? Sort of good at science? Or really good at science?" Responses were scored from 1 to 4. Responses to this item were high across the board (median = 4) and did not vary by condition in the intention-to-treat analysis (Control, $M = 3.38$, $CI = 3.24$, 3.52 ; Experimental, $M = 3.47$, $CI = 3.33$, 3.62).

Science Exclusivity Beliefs/Prevalence: Children next completed a measure of whether they think that doing science or being a scientist is common in their community (15). Children were shown a scale that presented different-sized groups of people and were asked: "Think of all the parents of all the kids at your school. Of all the parents of all the kids at your school, how many do you think do science/are scientists? Only one does science/is a

scientist? Just a few do science/are scientists? Some do science/are scientists? Or a lot do science/are scientists?" Children completed questions both about "being a scientist" or "doing science" in counter-balanced order across participants. Responses were scored from 1 to 4. As with the other scale measures, responses were high across the board (median = 4) and did not vary by condition (Control, $M = 3.31$, $CI = 3.21$, 3.41 ; Experimental, $M = 3.37$, $CI = 3.26$, 3.47).

Science Knowledge/Content Measure: To measure what children had learned from the target lesson, children were presented with seven questions. These included four questions about friction: 1) "Let's pretend you want to race a toy car down a ramp and need to choose the material to put down. You want it to go as far as possible! Should you race the car on wrapping paper or on a bath towel?" 2) The question was asked again, but with the answer choices of leaves or wood, 3) "Now let's pretend you want the car to not go very far. If you do not want the car to go very far, should you put down blanket or aluminum foil?" 4) "You noticed that the car goes further on some materials and not as far on other materials. Does this happen because of friction or electricity?" Children were then asked three questions about the scientific method as it was presented during the target lesson: 1) children were shown a picture of a child smelling a flower and a picture of a messy room full of toys and were asked to select which one showed, "observing", 2) children were shown a picture of a child guessing how tall a plant will grow and a picture of two children sharing an apple and were asked to select which one showed, "predicting," and 3) children were shown a picture of a child playing jump rope and a picture of a child testing how far his paper airplane would fly and were asked to select which one showed, "checking." Each question was scored as correct ("1") or incorrect ("0"). The proportion of accurate responses did not differ by condition for the friction items (Control, $M = 0.62$; Experimental, $M = 0.63$) or the questions about the scientific method (Control, $M = 0.53$; Experimental, $M = 0.47$).

Data Analysis. We tested whether the number of times that teachers produced action-based descriptions of science or identity-based references to scientists varied by condition in separate mixed effects negative binomial models. Using the "glmer.nb" function in the MASS package (29) in R version 3.6.1, these models included random intercepts for teachers, schools, and districts and tested for condition as a fixed effect. We then used the summary function to perform a Wald test of the parameter estimate associated with the fixed effect. We transformed the coefficients to Odds Ratios for reporting purposes and report the test statistics and P values from the Wald tests. For the random effects, all models revealed variance associated with the level of the individual teacher, but not for districts or schools. We adopted a similar approach to modeling for all analyses, with the following changes.

We tested for the effect of teachers' action-focused and identity-cuing language on children's persistence in separate mixed effects negative binomial models. These models included random intercepts for participants, teachers, schools, and districts and tested for the fixed effect of teacher language (continuous predictor variables were centered for analyses).

We conducted an intention-to-treat analysis, which tested for the causal effect of condition-assignment (at the level of the children's school) on children's level of persistence using a similar approach but with condition-assignment, instead of teacher language, as the fixed effect. Note that this provides a conservative estimate of the efficacy of the intervention because it does not account for variation in implementation or across the experiences of individual children.

Similar analyses were conducted on all of the scale assessments (including science interest, persistence, and prevalence) but with models with linear rather than negative binomial distributions. Similar analyses were also conducted on children's learning data, but with binomial models (as these were composed of a series of 0–1 responses across trials).

Changes from the Preregistered Analysis Plan. In the analysis plan posted prior to analyses (we posted this plan after data collection had begun but before any data had been processed at all), we had said that we would decide on the best models for analyzing the various count data collected in this project based on the distributions of the collected data. At the start of data analyses, we determined that negative binomial models would be appropriate for both the persistence task and language data, so decided to use that approach throughout.

At the start of the project, we had hoped to examine how the effectiveness of the intervention varied by children's demographic factors; however, because children completed study measures anonymously, we had limited access to this information. Children were asked to self-report their gender; we

are not sure of the accuracy of these reports (it seemed possible in some cases that children misunderstood this question), but based on the collected data, effects did not vary by participant gender. We had also thought to examine how effects change with children's (self-reported) age, but because we included only children who had already turned four and did not have birthdate information, we were unable to examine age-related changes. We did not have information on individual children's racial or ethnic backgrounds, but obtained the percentages of children from various racial and ethnic groups at each school from public records, as described earlier. While this is useful for describing our sample, it does not provide the level of information needed to examine how children's race or ethnicity predicted their response to this intervention. We did run analyses testing for whether any of the effects of the intervention varied by the proportion of children at the schools from various racial and ethnic backgrounds and found no evidence that this was the case. Also, we originally expected 12 school districts to participate, but one was never able to fit the project into their schedule. Due to this change and changes from enrollment estimates obtained prior to the start of the year, the ultimate sample was smaller than we had initially anticipated when drafting our first analysis plan.

In our plan, we noted that we would run models testing whether effects of the intervention were moderated by whether teachers watch the training video. Because the vast majority did so, we did not run these analyses and instead report the more conservative intention-to-treat analyses throughout. We also described a plan to code the teacher language data that was slightly different from what we ultimately report. In particular, we noted in the plan that we would analyze how often teachers said various target words in the lesson, including "observe," "predict," "check," and "friction," in addition to the specific language that teachers used to refer to science (identity-focused or action-focused cues). In doing so, we found that teachers in the experimental condition were more likely to say, "observe," "predict," and "check," and teachers in the control condition were more likely to say, "friction." However,

we realized in retrospect that because these target words were modeled more often in the experimental training video, this analysis is not very informative. Thus, we did not include these analyses in the paper (although note that, as reported above, the number of times that teachers explained these concepts did not relate to any outcome measure) and instead focus only on the more relevant and informative analyses of teachers' use of specific words to describe science (action-focused or identity-cuing language).

Finally, although we planned in advance to examine the effects of condition and linguistic production together (to test, for example, if the effect of condition is mediated by language), we realized upon beginning data analysis that these models with interaction terms were not tractable (given the complicated, nested structure of the data), and therefore, we report the effects of condition and of language in separate analyses.

Data and Materials Availability. All data for this study are available on the Open Science Framework: <https://osf.io/pe7k5/>.

ACKNOWLEDGMENTS. We are grateful to all of the teachers and children who participated in our research, to Aneesha Jacko for her support of this project, and to the New York City Department of Education. We thank Danielle Schuler, Maisy Rohrer, and members of the Conceptual Development and Social Cognition Lab for their assistance with data collection and preparation of materials. We thank Kelsey Moty for her help with development of programming for the dependent variables for this project and Dr. Amy Yamashiro for her contribution to data processing and analyses. Research reported in this publication was supported by a McDonnell Scholars Award as well as the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the NIH under Award Number R01HD087672. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the McDonnell Foundation.

1. S. R. Waxman, Names will never hurt me? Naming and the development of racial and gender categories in preschool-aged children. *Eur. J. Soc. Psychol.* **40**, 593–610 (2010).
2. S. R. Waxman, *Weaving a Lexicon* (MIT Press, 2004).
3. S. R. Waxman, D. B. Markow, Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognit. Psychol.* **29**, 257–302 (1995).
4. S. A. Gelman, E. A. Ware, F. Kleinberg, Effects of generic language on category content and structure. *Cognit. Psychol.* **61**, 273–301 (2010).
5. S. A. Gelman, G. D. Heyman, Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychol. Sci.* **10**, 489–493 (1999).
6. M. Rhodes, S.-J. Leslie, C. M. Tworek, Cultural transmission of social essentialism. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13526–13531 (2012).
7. M. Rhodes, S.-J. Leslie, Children's media communicates essentialist views of science and scientists. [10.31234/osf.io/s6fcb](https://doi.org/10.31234/osf.io/s6fcb) (4 October 2017).
8. S. A. Graham, C. S. Kilbreath, A. N. Welder, Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Dev.* **75**, 409–427 (2004).
9. G. Segall, D. Birnbaum, I. Deeb, G. Diesendruck, The intergenerational transmission of ethnic essentialism: How parents talk counts the most. *Dev. Sci.* **18**, 543–555 (2015).
10. S. A. Gelman, S. O. Roberts, How language shapes the cultural inheritance of categories. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 7900–7907 (2017).
11. E. Foster-Hanson, S.-J. Leslie, M. Rhodes, Speaking of kinds: How generic language shapes the development of category representations. [10.31234/osf.io/28qf7](https://doi.org/10.31234/osf.io/28qf7) (20 January 2019).
12. S. A. Gelman, *The Essential Child: Origins of Essentialism in Everyday Thought* (Oxford University Press, 2003).
13. M. Rhodes, T. M. Mandalaywala, The development and developmental consequences of social essentialism. *Wiley Interdiscip. Rev. Cogn. Sci.*, [10.1002/wics.1437](https://doi.org/10.1002/wics.1437) (2017).
14. M. Rhodes, S.-J. Leslie, K. M. Yee, K. Saunders, Subtle linguistic cues increase girls' engagement in science. *Psychol. Sci.* **30**, 455–466 (2019).
15. R. F. Lei, E. R. Green, S.-J. Leslie, M. Rhodes, Children lose confidence in their potential to "be scientists," but not in their capacity to "do science". *Dev. Sci.* **22**, e12837 (2019).
16. M. Rhodes, L. Bushara, "Learning about science and self: A partnership between the Children's Museum of Manhattan and the psychology department at New York University" in *Cognitive Development in Museum Settings*, D. M. Sobel, J. L. Jipson, Eds. (Routledge, 2015), pp. 103–119.
17. Final School Level Data by Grade 2018-2019, District Enrollment – Economically Disadvantaged. <http://www.p12.nysed.gov/irs/statistics/enroll-n-staff/District2019EconDisadv.xlsx>. Accessed 15 January 2020.
18. J. A. Fredricks, P. C. Blumenfeld, A. H. Paris, School engagement: Potential of the concept, state of the evidence. *Rev. Educ. Res.* **74**, 59–109 (2004).
19. D. Mahatmya, B. J. Lohman, J. L. Matjasko, A. F. Farb, "Engagement across developmental periods" in *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, C. Wylie, Eds. (Springer, 2012), pp. 45–63.
20. J. E. Jacobs, S. Lanza, D. W. Osgood, J. S. Eccles, A. Wigfield, Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Dev.* **73**, 509–527 (2002).
21. D. Stipek, D. Mac Iver, Developmental change in children's assessment of intellectual competence. *Child Dev.* **60**, 521–538 (1989).
22. C. R. Barman, Students' views about scientists and school science: Engaging K-8 teachers in a national study. *J. Sci. Teach. Educ.* **10**, 43–54 (1999).
23. E. McPherson, B. Park, T. A. Ito, The role of prototype matching in science pursuits: Perceptions of scientists that are inaccurate and diverge from self-perceptions predict reduced interest in a science career. *Pers. Soc. Psychol. Bull.* **44**, 881–898 (2018).
24. A. Cimpian, The impact of generic language about ability on children's achievement motivation. *Dev. Psychol.* **46**, 1333–1340 (2010).
25. A. Cimpian, H. M. Arce, E. M. Markman, C. S. Dweck, Subtle linguistic cues affect children's motivation. *Psychol. Sci.* **18**, 314–316 (2007).
26. E. Foster-Hanson, A. Cimpian, R. A. Leshin, M. Rhodes, Asking children to "be helpers" can backfire after setbacks. *Child Dev.* **91**, 236–248 (2020).
27. Y. Dunham, A. S. Baron, S. Carey, Consequences of "minimal" group affiliations in children. *Child Dev.* **82**, 793–811 (2011).
28. A. Master, S. Cheryan, A. N. Meltzoff, Social group membership increases STEM engagement among preschoolers. *Dev. Psychol.* **53**, 201–209 (2017).
29. B. Ripley et al., "Support Functions and Datasets for Venables and Ripley's MASS" in *Modern Applied Statistics with S* (Springer-Verlag, New York, ed. 4, 2019).