Brandon Rhodes
12/09/2020
STAT 34900 — Data Analysis Project
Professor Peter McCullagh
Project Final Report

1 Executive summary

This final project is a reanalysis of a study on a field experiment carried out on a set of preschools in New York City by researchers at New York University. For the experiment, teachers received either a video that implicitly emphasized use of action-based language or a video which did not emphasize such language when the teachers were preparing for a lesson they would teach on friction. These teachers later taught the lesson demonstrated in their video to their students, all of whom were ages three to four, and the researchers made measurements on a student's willingness to continue playing a video game related to friction. The researchers hypothesized that students who had a teacher in the treatment condition would exhibit more willingness to continue playing the friction game. This was their hypothesis because they thought the treatment videos would bias teachers away from using identity-presupposing language: this is the type of language which requires a student to assume the role of a particular figure to study a concept concept. In this case, the figure would be the scientist and the concept would be science (or more specifically, friction). The authors analyzed data pertaining to student persistence as related to both the treatment and linguistic behavior of the teachers, and they analyzed data pertaining to the effect the treatment videos had on teacher language. This report will consist largely of a re-examination of the same data, making changes to the analysis, and then comparing the conclusions to those of the authors.

In this report, we use a different model than the authors to examine the treatment effect on student persistence, where the model here explicitly accounts for the censoring which occurs in the data. We conclude that there is not sufficient evidence to believe there is a treatment effect; or, being optimistic, there exists an small effect in the direction of students being more persistent when taught by a teacher who was in the treatment condition. This effect is magnitudes smaller in comparison to the main effects of the model (which are intercepts for the hazard function assumed). In general, students were likely to attempt no additional trials in the video game on which their persistence was measured. There was a contingent of students, though, who were censored, suggesting that this is in fact an important aspect of an analysis. The main conclusion about no treatment effect is in opposition to what the authors claim. I also consider a reinterpretation of 'persistence', but the conclusions remain the same.

For the effect of the treatment on teacher's linguistic behavior, we arrive at similar conclusions to those of the authors in this report, but the details of these conclusions do vary in important ways. There is evidence that teachers in the treatment condition did produce more action-based language, but there is not evidence that these teachers produced less identity-presupposing language. This is in contrast with the conclusions from the authors. The main differences between the analysis here and that of the authors are that (i) the analysis here addresses information derived from the categorization of the words *science* and *scientist* and (ii) the analysis here addresses an important aspect of teachers with zero counts for both categories of either of these words. The authors made no such distinction in their analysis for these teachers with zero counts for one of the categories and zero counts for both of the categories of *science* and *scientist*. In agreement with the authors, teachers in the treatment condition did seem to emphasize more the steps of doing science in their lesson on friction.

We consider last the relationship between a teacher's linguistic behavior and a student's persistence. We find that there is no evidence to suggest that there is any relationship. We argue for the need to transform the linguistic variables to account for important structure in the units, and upon transforming the variables there is not strong evidence to suggest the presence of a relation between either action-based or identity-presupposing language and student persistence. This is also in opposition to the claims of the authors. They claim that identity-presupposing language is associated with less student persistence. The sensitivity of the conclusions to transformations of the variables suggests the need for at least a more careful analysis of this data.

2 Brief description of the study

Researchers from New York University collaborated with the New York City Department of Education to complete a field experiment to evaluate the effect a specific training regimen has on preschool students' engagement and interest in science. The authors wanted to see if a training regimen that encouraged subtle differences in language usage by teachers could promote specific student behaviors, such as increased persistence in learning a science concept and more positive attitudes towards science and towards their ability to do science. The linguistic differences pertained to whether or not a teacher's language implied that a student must be a scientist for them to study science. For this study, the science lesson was on friction. The authors claimed that the standard language a child hears is language which implies they must assume the identity of a scientist to study friction; an example of this kind of language would be, 'Let's be scientists today and study friction!'. In an alternative case, a student would hear language which would not imply they need to be a scientist to study friction; an example of this language would be, 'Let's do science today! We are going to study friction!' These two types of language were called identity-presupposing language and action-based language respectively by the researchers.

The treatment was the distribution of training videos to teachers: teachers in the control group received a training video with no emphasis on using action-based language, and teachers in the treatment group received a training video where there was implicit emphasis on the use of action-based language. Approximately two days after the lesson was given to the students, a researcher visited the classes and made measurements on each student. Two types of measurements were made, but only one set was made on a given student, and this was decided at random. The first type of measurement was persistence in learning about friction. The second type of measurement consisted of a series of questions designed to ascertain a given student's excitement about science and their self-belief with respect to their ability to do science. I will only focus on the first type of measurement, the persistence one.

Teachers also recorded (just audio) the lesson that they gave to the students, and the authors had research assistants, blind to the treatment, code the language from each teacher's session. Research assistants counted the number of times a teacher used identity-presupposing language and how many times they used action-based language, by looking at specific instances of the words *scientist* and *science* and categorizing them. There is more detail on this in sections §5.3 and section §6 of the report.

3 Information on the experiment and measurements recorded

3.1 Timeline of study and randomization scheme

Prior to the school year, NYU and the NYC Department of Education agreed to the study, and they sampled 11 school districts from New York City. At this time, demographic information was obtained from teachers who were willing to provide that data. Then, within each district, schools were randomly assigned to the control or treatment group. Treatment was assigned at the school level so that there would not be any contamination of the treatment; if it were assigned at the class / teacher level, contamination could occur via teachers talking to one another. In other words, this prevented interference. At some point after randomized treatment assignment, the classes were assembled. For the lesson on friction, the training video was distributed to all teachers on the Thursday of the week before the lesson was taught. On Monday or Tuesday of the next week, each teacher gave the lesson. On Thursday or Friday of the lesson week, the researchers recorded their measurements.

3.2 Information on the treatment

Teachers in the control group received a training video which contained instructions on how to set up and implement the lesson on friction, but no examples of a teacher demonstrating it to students was provided. Teachers in the treatment group received the same video as the control group, but their video was augmented with specific examples of teachers implementing the lesson with preschool students, using action-based language. The teachers were blind to treatment, and they were not explicitly told to use action-based language (or not to use identity-presupposing language). All teachers were given a set of questions to answer

after watching the video to ensure comprehension. For the actual lesson, all teachers were given the same physical materials. Teachers were also given a written lesson plan, controlled for treatment.

3.3 Information on the measurements

The persistence measurements were obtained by recording a student's willingness to continue playing a video game on a tablet, where the video game mirrored the activity the teachers used to present the concept of friction. The point of the game was to guess how far a car would roll down a ramp, given the material of the ramp (e.g. a wooden ramp versus a carpet ramp versus a ramp with a paper surface). The student was given a tablet and headphones to play the game. The game was rigged so that any answer the student gave on their first iteration of the game was seen as correct and then any answer the student gave on their second iteration was seen as incorrect. After the second iteration, a narrator in the game asked the student, 'Do you want to keep playing, or do something else?' Researchers recorded how many additional iterations beyond the first two that a student would play the game, capping the amount of additional iterations at six. No feedback was given on any of the additional iterations, only the first two. The narrator asked the same question about whether or not the student wanted to continue after each iteration played, except after the sixth additional iteration.

The teachers' language production was measured by counting the instances of certain words in the transcript of the recording of their lesson. Research assistants, blind to the treatment, coded each teacher's lesson, and for the words *science* and *scientist*, they categorized each into two distinct groups. This is addressed in sections §5.3 and section §6.

4 A diagram of the study set up and components for analysis

A diagram of the experimental setup is telling with regard to how we may approach analyzing this data, which is the bulk of this final project report. It makes up all of section §5. The response of primary interest for the researchers is students' persistence; however, it is important to note the presence of intermediate responses, which are related to the linguistic behavior of the teachers. Below we see this represented as the space in between the treatment and the primary response.

$$\underset{}{\text{Treatment}} \rightarrow \underset{\text{INTERMEDIATE RESPONSE}}{\text{Teacher language}} \rightarrow \underset{\text{PRIMARY RESPONSE}}{\text{Persistence}}$$

It is important to note that in the analysis of a treatment effect on the primary response, we must not include the intermediate responses as covariates. They would be confounded with the treatment effect, and it would make inference on the treatment effect less reliable. Conclusions drawn from such an analysis would warrant immediate skepticism. With this observation in mind, section §5.1 is concerned with an analysis of a treatment effect on student persistence, and the strategy for this analysis can be visualized in the diagram below, which makes note of the observation. We do not consider a teacher's linguistic behavior during their lesson when analyzing the effect of treatment.

Analysis of treatment on student persistence (experiment)

$$\underset{}{\text{Treatment}} \rightarrow \underset{\text{INTERMEDIATE RESPONSE}}{\text{Teacher language}} \rightarrow \underset{\text{PRIMARY RESPONSE}}{\text{Persistence}}$$

Continuing with outcomes of the experiment, we can look at the initial link in the diagram as representing an analysis which seeks to examine the effect of the treatment on a teacher's linguistic behavior. This can be thought of as assessing the efficacy of the training, and section §5.3 will go into detail on how exactly the treatment may have biased a teacher's language. In this analysis, we make no reference to the student persistence data, as it is largely irrelevant for these purposes.

Analysis of treatment on linguistic behavior (experiment)

$$\underset{}{\text{Treatment}} \rightarrow \underset{\text{INTERMEDIATE RESPONSE}}{\text{Teacher language}} \rightarrow \underset{\text{PRIMARY RESPONSE}}{\text{Persistence}}$$

Last, we can examine the relationship between the intermediate responses and the primary response. This is the focus of section §6. In this case, we regard the intermediate responses as covariates, but we note that this is now a study of the observational type. It is not an experiment. We do not consider the treatment factor in this analysis. Though the conclusions drawn from this analysis should be looked at more closely, it is still informative to look at this link in the diagram.

Analysis of linguistic behavior on persistence (observational study)

Treatment → Teacher language → Persistence
INTERMEDIATE RESPONSE PRIMARY RESPONSE

5 Analysis of the outcomes of the experiment

5.1 Analysis of the effect of treatment on students' persistence

The first part of the project is concerned with the persistence data. The goal of the researchers was to see if increased usage of action-based language by an instructor is associated with higher persistence outcomes for students. A student's persistence was considered to be the number of additional trials of a video game they played after they received negative feedback on the second of two initial trials playing the video game. The authors wanted to look for the presence of an effect of the treatment, which were the training videos distributed at the school level. With this in mind, the first analysis here is one of an experiment with randomized treatment assignment at the school level. I will also consider a student's persistence to be the number of additional trials of a video game they played, but I will consider a reinterpretation of the term 'persistence' in section §5.2.

5.1.1 Structure in the units (students) for persistence data

The observational units are child, trial pairs (i, t), where i indexes the child and $t \in \{0, 1, 2, 3, 4, 5\}$. The response Y is in the set $\{0, 1\}$; it is an indication of whether or not child i quit playing the video game after completing trial t. The child was not asked after the sixth trial whether or not they wanted to continue playing, so children who played the game six times I will consider to be censored. There are several block factors, which are nested. The block factors are the following: district, school and teacher. There are several covariates which provide further structure on the observational units. For the experiment, the covariates of interest are the following in Table 1.

Name	Description	Variable type
Gender	Gender of the child (two levels: girl and boy)	Classification factor
Class size	Size of the child's class	Quantitative variable
Nonwhite	The proportion of students at the child's school which are not white (i.e. Asian, Black, Hispanic or other)	Quantitative variable
Trial	Trial at which the child has a decision to stop playing the game (six levels: 0, 1, 2, 3, 4 and 5)	Classification factor

Table 1: Covariates available for the analysis of the experiment.

There is also the treatment factor called 'Condition', with two levels (control and experimental). The use of the Nonwhite variable is perhaps suspect; or, at the very least, it is one likely to be questioned. I am considering it not as much of an indication of any relationship between persistence and race, as we do not have race data on the child level, but more so as a rough indicator of the socio-economic status of the child. As it turns out, we will see there is no evidence for a meaningful effect due to this covariate anyways.

5.1.2 Model specification for persistence data

As suggested by the definition of the observational unit and response above, I consider these data to be of the survival sort in discrete-time. I assume the response Y_{it} is distributed as Bernoulli when conditioned on the random effects to account for the correlation of districts, schools and classes; these random effects are assumed to be normally distributed with mean zero and with variance being the same within each level of the hierarchy. All the random effects are independent of one another. Conditioned on these random effects, the responses are independent. The conditional mean is the hazard function $h(t;\mathbf{x}_i) = \mathbb{P}\big(Y_{i,t} = 1 \big| Y_{i,t-1} = 0; \mathbf{x}_i\big) = \pi_{it}$. Recall that $Y_{it} = 1$ indicates a child quit after completing trial t. Specifically, I will assume the logits of the hazard function are linear in the covariates. An implicit specification of the model is shown below. The functions d, s and c just indicate which district, school and class to which a child i belongs. All models considered are estimated by maximum likelihood, where the log-likelihood is computed using the Laplace approximation.

$$\eta_{it} = \beta^T \mathbf{x}_{it} + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)} \qquad \pi_{it} = \frac{e^{\eta_{it}}}{1+e^{\eta_{it}}} \qquad Y_{it}\Big|\tau_{d(i)}, \tau_{s(i)}, \tau_{c(i)} \sim \text{Bern}(\pi_{it})$$

$$\tau_{d(i)} = \begin{cases} \tau_{d_1} \sim N(0, \sigma_0^2) & \text{if } d(i) = 1 \\ \vdots & \\ \tau_{d_{n_d}} \sim N(0, \sigma_0^2) & \text{if } d(i) = n_d \end{cases} \qquad \tau_{s(i)} = \begin{cases} \tau_{s_1} \sim N(0, \sigma_1^2) & \text{if } s(i) = 1 \\ \vdots & \\ \tau_{s_{n_s}} \sim N(0, \sigma_1^2) & \text{if } s(i) = n_s \end{cases}$$

$$\tau_{c(i)} = \begin{cases} \tau_{c_1} \sim N(0, \sigma_2^2) & \text{if } c(i) = 1 \\ \vdots & \\ \tau_{c_{n_c}} \sim N(0, \sigma_2^2) & \text{if } c(i) = n_c \end{cases} \qquad \text{All random effects are independent of one another.}$$

As one informal justification of the family of models selected, we can look at a histogram displaying the counts of students who played a certain amount of additional trials of the video game. This is shown in Figure 1. The survival nature of the data is evident in the roughly monotonic decreasing trend in the data. In addition, we see an accumulation of data at the sixth additional trial, reinforcing the idea that we should take into account the censoring.

5.1.3 Conclusions for the persistence data

In Table 2 we see the parameter estimates and standard errors for the final model. It can be seen that there is no parameter included for gender nor for class size. Starting with a proportional odds model, I tried including an intercept for the gender classification factor and a parameter estimate for quantitative variable of class size, but neither improved the fit. To test these hypotheses, I considered the log-likelihood ratio between the model whose linear predictor of the logits for the hazard function is shown below and the model which included a term for gender or class size. The difference in deviances for considering gender was 1.2 and that considering class size was 1.3, both of which are not particularly convincing given the asymptotic null χ_1^2 distribution of these statistics. The most simple, yet appropriate model is seen below. The function $treat(i)$ is an indicator function for whether or not the student was in the treatment group.

$$\eta_{it} = \alpha_t + \beta_0 \cdot treat(i) + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)} \qquad \pi_{it} = \frac{e^{\eta_{it}}}{1+e^{\eta_{it}}} \qquad Y_{it}\Big|\tau_{d(i)}, \tau_{s(i)}, \tau_{c(i)} \sim \text{Bern}(\pi_{it})$$

$$\begin{aligned} \tau_{d(i)} &\sim N(0, \sigma_0^2) \\ \tau_{s(i)} &\sim N(0, \sigma_1^2) \qquad \leftarrow \text{All random effects independent of one another.} \\ \tau_{c(i)} &\sim N(0, \sigma_2^2) \end{aligned}$$

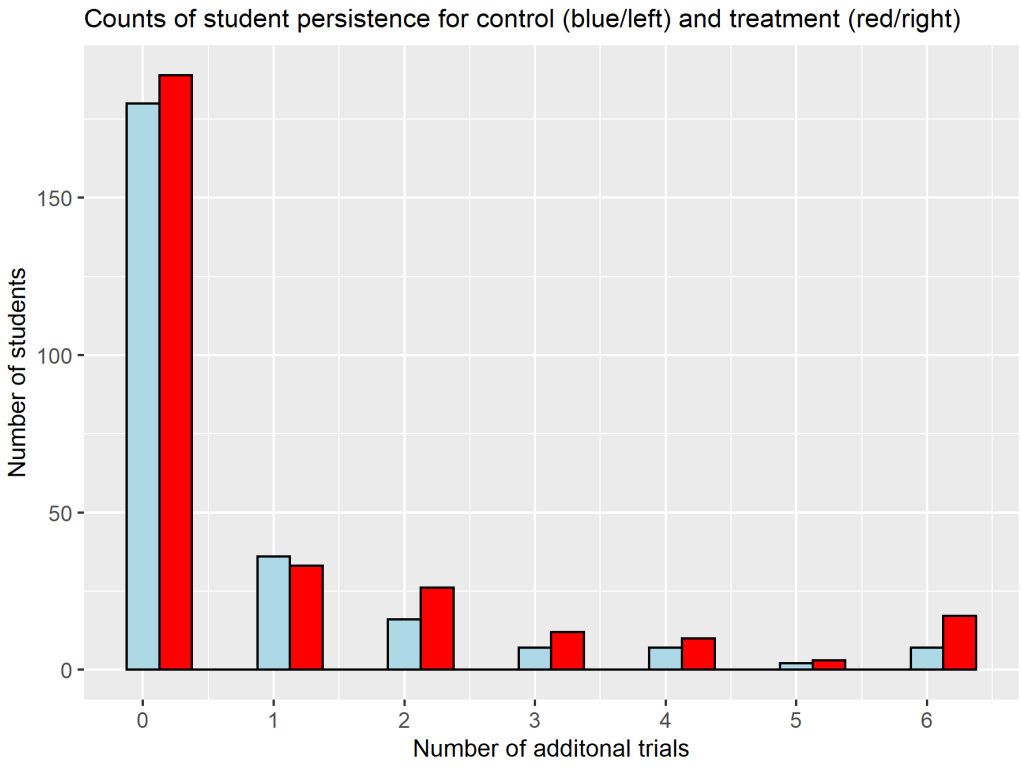

Figure 1: The counts of students who played the video game a certain number of additional trials beyond their negative feedback on the second of two initial trials.

Interestingly, the variance components are not large. Both the district and class level variance components, σ_0^2 and σ_2^2, is estimated to be zero. The school level variance component is estimated to be $\sigma_1^2 = 0.045$. This is small compared to the estimates for the parameters. The parameter estimates for the logits of the hazard function at each addition trial are 0.91, -0.25, -0.21, -0.64, -0.27 and -1.29 with standard errors in the range 0.13–0.51. The standard errors for the logit estimates increase monotonically in additional trial number; this makes sense given there is less data with each additional trial. We see that most prominent estimates are the one for the zero-th additional trial and the one for the fifth additional trial, with opposing signs. As we can see from the histogram in Figure 1 and the estimate here, most kids did not want to continue playing the game after the second initial trial where the game was rigged to give them negative feedback. On the other hand, children who made it through at least the fifth trial were more likely to want to continue playing, as seen by the negative log-odds ratio. Apparently, they were absorbed by the game.[1] The estimated logits are 1.29, which is 2.5 times the standard error. At all other time points, the estimates suggest it is a coin flip whether or not the child continues to play.

The question of most interest, however, is the one concerning the treatment factor. The analysis done here is not very conclusive with regard to the existence of a treatment effect; however, if it does exist, it is very mild in this experiment. The estimated treatment effect is -0.29 with a standard error of 0.16. The two sided p-value for this estimate is 6%. The log-likelihood ratio of the model with a treatment factor and the model without a treatment factor is 3.1, which is at the 94th percentile of the asymptotic χ_1^2 distribution. Neither of these are convincing pieces of evidence, but they may be suggestive.

This is close to the conclusion the authors reached, though they are more forthright with their declaration of the existence of a treatment effect. They consider the response to be distributed as negative binomial, and they report a two-sided p-value of 4% for the parameter estimate for the treatment factor. They do not report any likelihood ratio statistics. The parameter estimate they report is 1.44 with a standard error

[1] Or, endearingly, they were just nerdy...

Name	Parameter	Estimate	Standard error
$h(t=0)$ logits	α_0	0.91	0.13
$h(t=1)$ logits	α_1	-0.25	0.18
$h(t=2)$ logits	α_2	-0.21	0.22
$h(t=3)$ logits	α_3	-0.64	0.30
$h(t=4)$ logits	α_4	-0.27	0.33
$h(t=5)$ logits	α_5	-1.29	0.51
Treatment	β_0	-0.29	0.16
Level	Variance component	Estimate	
District	σ_0^2	0	
School	σ_1^2	0.045	
Class	σ_2^2	0	

Table 2: The parameter estimates for the model fitted to look at the treatment effect on student persistence.

of 0.51. Although the parameter they estimate has different sign, it is in the same direction as the effect estimated here with regard to the interpretation. Overall, the evidence is not clear in favor of there being an effect of treatment; however, as the authors also note, any treatment effect given the experiment design may be difficult to detect at the child level since the randomization was done at the school level within each district. From this, they claim the estimate for the effect will be conservative, and that is probably why they feel comfortable concluding there is an effect due to treatment. However, the evidence here suggests they should perhaps reconsider their conclusions, especially since the log-likelihood ratio statistic is known to be anti-conservative.

It is worth noting that the proportional odds model fit better than one where this assumption was relaxed. I tested the relaxing of this assumption formally by first fitting a model with an interaction term for the treatment factor and trial number and then looking at the log-likelihood ratio between that model and the proportional odds model. This statistic was 1.7, and the asymptotic distribution is χ^2_6, so there is no evidence to suggest we use the more complex model in lieu of the simpler proportional odds model.

5.2 Another perspective on the persistence data

The preceding analysis in §5.1 is not entirely conclusive with regard to the presence of a treatment effect. However, we may want to reevaluate the question being asked. The question the researchers are concerned with is the following: Does training teachers with videos with implicit emphasis on action-based language increase students' persistence? The analysis above and in the paper seem to get at an answer to that question, but I am not sure that the set up of the experiment measures student persistence the way the researchers may actually want. In effect, the researchers want to see if the treatment leads to students playing more trials of the game. They call this persistence. This seems to be a bit of a looser interpretation of persistence. A tightened interpretation has to do with a student's willingness to continue immediately after they have faced some adversity. The adversity in the experiment was on the second initial trial, the one which was rigged to always give the students negative feedback. However, note that the students did not receive any feedback on any of the additional trials, which were the ones used as a measurement. This means that perhaps a more interesting, or even better, measure of a student's persistence is whether or not the student decided to play **any** additional trials after receiving negative feedback on the second initial trial. I believe this is a measurement of persistence when its interpretation is not as loose.

The hypothesis for the treatment effect in the model above and the one suggested in the paper is that the treatment has some effect on the total number of steps played after receiving one instance of negative feedback; the hypothesis under consideration now is that the treatment had some effect on whether or not students decide to play **any** additional trials after receiving negative feedback. There is a subtle difference in the question asked, but I think different question is worth distinguishing and worth asking.

5.2.1 Model specification for persistence data — a reinterpretation

The structure in the units is the same as before, as mentioned in 5.1.1. This leads us to the same assumptions for the random effects, namely that they are nested and normally distributed with mean zero and have the same variance within each level of the hierarchy. The probability distribution for the response is different, though. The response $Y \in \{0, 1\}$, when conditioned on the random effects, is Bernoulli distributed with the mean being the probability of a child playing beyond the two initial trials, where they received negative feedback after the second. Any child i who played one to six additional trials is considered to be $Y_i = 1$ and any child who decided to stop playing the game at the first opportunity is considered to be $Y_i = 0$. This leads us to a similar model specification as before, which can be seen below. The linear predictor is no longer a function of the additional trials t. As before, this model is estimated by maximum likelihood and the Laplace approximation is used to compute the likelihood.

$$\eta_i = \beta^T \mathbf{x}_i + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)} \qquad \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \qquad Y_i \Big| \tau_{d(i)}, \tau_{s(i)}, \tau_{c(i)} \sim \text{Bern}(\pi_i)$$

$$\tau_{d(i)} = \begin{cases} \tau_{d_1} \sim N(0, \sigma_0^2) & \text{if } d(i) = 1 \\ \vdots & \\ \tau_{d_{n_d}} \sim N(0, \sigma_0^2) & \text{if } d(i) = n_d \end{cases} \qquad \tau_{s(i)} = \begin{cases} \tau_{s_1} \sim N(0, \sigma_1^2) & \text{if } s(i) = 1 \\ \vdots & \\ \tau_{s_{n_s}} \sim N(0, \sigma_1^2) & \text{if } s(i) = n_s \end{cases}$$

$$\tau_{c(i)} = \begin{cases} \tau_{c_1} \sim N(0, \sigma_2^2) & \text{if } c(i) = 1 \\ \vdots & \\ \tau_{c_{n_c}} \sim N(0, \sigma_2^2) & \text{if } c(i) = n_c \end{cases} \qquad \text{All random effects are independent of one another.}$$

5.2.2 Conclusions for the persistence data — a reinterpretation

In Table 3, we see the estimates for the parameters and their standard errors. The final fitted model includes a term for gender but not for nonwhite. This is a result of residual analysis done. The residuals were binned by covariate class, averaged and then plotted. The computations were done by the function `binnedplot` in the R package `arm`. Although neither of these predictors substantially increases the fit in terms of a change in log-likelihood, as the differences in deviances (from the model with only an intercept and treatment factor in the linear predictor) are 1.5 and 1.6 for gender and nonwhite respectively, there was a linear trend in the residuals from the model fit without at least one of these covariates. To remove the linear trend in the residuals, I only include an intercept for gender. I am not aware of a principled reason to prefer gender over nonwhite, but I chose to do this because there are 45 missing observations for the nonwhite and only one for gender. That is a notable amount of observations missing for nonwhite but not for gender. Further analysis of the missingness for nonwhite values would have to occur before I feel comfortable including it in a model. The final fitted model is shown below. The functions $treat, girl$ are indicator functions for treatment and girl gender respectively.

$$\eta_i = \mu + \beta_0 \cdot treat(i) + \beta_1 \cdot girl(i) + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)} \qquad \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \qquad Y_i \Big| \tau_{d(i)}, \tau_{s(i)}, \tau_{c(i)} \sim \text{Bern}(\pi_i)$$

$$\begin{aligned} \tau_{d(i)} &\sim N(0, \sigma_0^2) \\ \tau_{s(i)} &\sim N(0, \sigma_1^2) \quad \leftarrow \text{All random effects are independent of one another.} \\ \tau_{c(i)} &\sim N(0, \sigma_2^2) \end{aligned}$$

The variance components in this analysis for σ_0^2, σ_1^2 and σ_2^2 are estimated to be 0.05, 0.004 and 0 respectively. Again, there is not a substantial amount of variance, but most of variation which is estimated occurs at the district level. Though the intercept is typically not interesting, in this case I think it is worth

noting. It is estimated to essentially be -1 with a standard error of 0.18, which means children are roughly 2.7 times more likely to not continue playing the game after the second initial trial where they received negative feedback. It is unclear whether or not that has more to do with the attention span of a four-year old or discouragement from the negative feedback. As expected based on the goodness of fit statistic of the log-likelihood ratio of the model with and without the gender term, the estimate is close to zero at 0.23 with a standard error of 0.19. Likewise, the treatment effect is small; it is 0.25 with a standard error of 0.19. This estimate is only slightly larger than its standard error, which means that the treatment effect is entirely unconvincing in this case. The log-likelihood ratio statistic between the model whose linear predictor include the intercept and gender term and then the one whose linear predictor includes those two terms as well as a treatment term is 1.6. This is not strong evidence given the χ^2_1 asymptotic distribution of the statistic.

Name	Parameter	Estimate	Standard error
Intercept	μ	-1.00	0.18
Treatment	β_0	0.25	0.19
Girl	β_1	0.23	0.19
Level	Variance component	Estimate	
District	σ^2_0	0.05	
School	σ^2_1	0.004	
Class	σ^2_2	0	

Table 3: The parameter estimates for the model fitted in the reinterpretation of the question.

This analysis suggests that there is not a convincing effect on children's persistence which is due to the distribution of different training videos to schools. This is in opposition to the claims the authors make but consistent with the analysis in §5.1. The conclusions from that section are a bit tentative, but the conclusions from this analysis suggest it is maybe justified to be a bit hesitant with regard to claiming the existence of a treatment effect on children's persistence. Overall, students did not care to play the video game any additional trials, and as I alluded to above, it is unclear whether or not this is from boredom or from receiving negative feedback. It is not clear how harsh this feedback was, and I do not have the expertise to decide with any confidence which I think it is. It is hard to be convinced of a treatment effect on persistence at this point.

5.3 Analysis of the effect of treatment on teachers' language production

The intermediate responses in the study at hand relate to the linguistic behavior of the teachers involved. The goal of the treatment was to bias teacher language behavior away from using identity-presupposing language related to science by implicitly emphasizing action-based language use in the training video. As we see above in section §5.1, there is not much evidence to suggest the treatment had an effect on student persistence; if there was any effect, it was small. With this in mind, it is instructive to question whether or not the treatment had an effect on a teacher's linguistic behavior, as it is this to which a given student was exposed. As a quick summary, the treatment appears to have had the following effects: (i) increased usage of *science* as an action versus that of a noun; (ii) no (or very slight) decrease in usage of *scientist* in an identity-presupposing way versus a generic way; (iii) increased usage of action-oriented language; and (iv) no effect on the length of a teacher's lesson. I will not present an analysis regarding the absence of evidence for an effect of the treatment on a teacher's lesson length: in all models I fitted to this data, the evidence firmly supported the claims of the authors that there was no effect from the treatment. Albeit an important sanity check on the implementation of the experiment, analyzing the effect of the treatment on lesson length is not of much interest here. In the remainder of the section, I only address the other three responses mentioned.

5.3.1 Structure in the units (teachers)

The observational units in all of the following analyses are teachers, and no longer students, but note that the experimental units are still schools. Unfortunately, there was not much information recorded on the teachers at baseline, so the structure to use in any model in the following analyses is limited. The researchers had given

the participating teachers an opportunity to provide demographic information, but there were approximately 40% values missing, so for the analysis below I do not consider demographic information. Though it could be of potential interest to examine whether or not there exist differences in how demographics respond to the treatment, this requires a much more careful analysis of the missing data, which will not be pursued below.[2] For this analysis, the only available structure in the observational units are the school and district block factors.

5.3.2 Model specification for use of *science*

As one proxy for measuring a teacher's usage of action-based language versus identity-presupposing language, the researchers counted the number of times a teacher said *science* during their lesson and they categorized each instance as either 'action' or 'noun'. An example of *science* used in an action way is *Let's do science!*; an example of it used as a noun is *It's science time!* They obtained these counts from the transcripts of each teacher's lesson, as the audio for each lesson was recorded. Each of these counts are regarded as a response; however, note one instance of *science* is **either** action **or** noun, so they are not independent. This suggests instead to consider the response of interest to be the number of action counts for *science* or the number of noun counts of *science* as Y_i, and we can make assumptions on its distribution conditioned on the total number of counts N_i for *science* during the lesson for a given teacher. A more complete analysis considers the treatment effect on the number of times a teacher used *science* in general, and that is pursued in section §5.3.3.

The response formally is notated as Y_i where i indexes the teacher, and Y represents the number of counts of *science* classified as action in the transcript of a teacher's lesson. I assume independent random effects for each district and school which follow a normal distribution with zero mean and variance being equal within each level of the hierarchy. Conditional on these random effects and the total number of counts N_i of *science*, Y_i follows a binomial distribution with probability π_i and N_i trials.[3] I will not use any covariates and only include an intercept and the treatment effect in the parameterization of the logits for the mean π_i. The model specification is shown below. The model is estimated by maximum likelihood, and the likelihood is computed using the Laplace approximation.

$$\eta_i = \mu + \beta_0 \cdot treat(i) + \tau_{d(i)} + \tau_{s(i)} \qquad \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \qquad Y_i \Big| \tau_{d(i)}, \tau_{s(i)}, N_i \sim \text{Binom}(\pi_i, N_i)$$

$$\tau_{d(i)} = \begin{cases} \tau_{d_1} \sim N(0, \sigma_0^2) & \text{if } d(i) = 1 \\ \vdots & \\ \tau_{d_{n_d}} \sim N(0, \sigma_0^2) & \text{if } d(i) = n_d \end{cases} \qquad \tau_{s(i)} = \begin{cases} \tau_{s_1} \sim N(0, \sigma_1^2) & \text{if } s(i) = 1 \\ \vdots & \\ \tau_{s_{n_s}} \sim N(0, \sigma_1^2) & \text{if } s(i) = n_s \end{cases}$$

All random effects are independent of one another.

5.3.3 Conclusions on the effect of treatment on *science*

Below in Table 4, we have the parameter estimates. Here the variance components are interesting: the variance components σ_0^2 and σ_1^2 are estimated as 1.31 and 2.04 respectively. Recall that these estimates correspond to random effects in the parameterization of the logits, so differences between schools and between districts is quite large. In the face of this variability, we must take the parameter estimate for the treatment effect with some grain of salt; however, there is still something to be said since the estimate for this parameter is large at 5.02 with standard error 0.75. The intercept is estimated to be -2.83 with standard error 0.69,

[2] Another difficulty, which is less problematic but admittedly more annoying, is that the researchers did not give informative names to the levels of the classification factors for demographic information. So, even when trying to sort this data out, it was difficult to get any grasp on what was going on.

[3] Though each instance of *science* can be argued to not be independent of the others, I will pursue an analysis which makes this independence assumption in the model, just as an attempt to understand what is going on.

suggesting the default for those teachers not in the treatment condition was to not use *science* in an action way.

Name	Parameter	Estimate	Standard error
Intercept	μ	-2.83	0.69
Treatment	β_0	5.02	0.75
Level	Variance component	Estimate	
District	σ_0^2	1.31	
School	σ_1^2	2.04	

Table 4: The parameter estimates for the model fitted for the counts of the usage of *science* as an action versus a noun.

The main conclusion is that the treatment did seem to bias teachers' linguistic behavior in a way the researchers intended even though there is a bit of variability estimated at both levels of the hierarchy. The difference of deviances between the model with a treatment effect and one without is 36, which yields a near zero p-value for the asymptotic χ_1^2 distribution for the null hypothesis.

5.3.4 Model specification for use of *scientist*

Next, we consider the responses related to a teacher's usage of the word *scientist* during the lesson. As with the previous contrast between *science* as an action or it as a noun, the researchers contrasted uses of *scientist* as a proxy for identity-presupposing language versus action-based language. The researchers classified each instance of *scientist* in the transcript of a teacher's lesson as either 'identity' or 'generic'. An example of an identity instance is *Today, we are going to be scientists!*; an example of a generic instance is *Scientists work hard to solve problems.* I analyze this data in a completely analogous way as the *science* data in the previous section (§5.3). The model is implicitly specified below, except now that the response Y_i corresponds to the number of times *scientist* was classified as 'identity' in the transcript of teacher i's lesson.

$$\eta_i = \mu + \beta_0 \cdot treat(i) + \tau_{d(i)} + \tau_{s(i)} \qquad \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \qquad Y_i \Big| \tau_{d(i)}, \tau_{s(i)}, N_i \sim \text{Binom}(\pi_i, N_i)$$

$$\tau_{d(i)} = \begin{cases} \tau_{d_1} \sim N(0, \sigma_0^2) & \text{if } d(i) = 1 \\ \vdots & \\ \tau_{d_{n_d}} \sim N(0, \sigma_0^2) & \text{if } d(i) = n_d \end{cases} \qquad \tau_{s(i)} = \begin{cases} \tau_{s_1} \sim N(0, \sigma_1^2) & \text{if } s(i) = 1 \\ \vdots & \\ \tau_{s_{n_s}} \sim N(0, \sigma_1^2) & \text{if } s(i) = n_s \end{cases}$$

All random effects are independent of one another.

5.3.5 Conclusions on the effect of treatment on *scientist*

In Table 5, we have the parameter estimates. Again, there is a fairly large estimated variance component. The variance component estimates for σ_0^2 and σ_1^2 are 0 and 1.26 respectively. One very important thing to keep in mind is that this analysis is done on only a fraction of the data, as only 39 of the 116 teachers used *scientist* during their lesson. This fact is somewhat counter to what the study implies; at the very least, the authors do not make explicit the fact that the main source of identity-presupposing language, which is through the word *scientist*, is not common in the first place. Even more so with this part of the study, we need a careful analysis of the missing data, but that will be left for future work. The parameter estimates for the intercept and treatment effect are 1.90 and -1.72 with standard errors 0.55 and 0.74 respectively.

The estimated treatment effect is close in size to the estimate of the variance component and just over two times the standard error. These facts suggest that if there is any effect of the treatment, it is quite small, unlike in the previous section where the estimated effect is quite large compared to the variance components

Name	Parameter	Estimate	Standard error
Intercept	μ	1.90	0.55
Treatment	β_0	-1.72	0.74
Level	Variance component	Estimate	
District	σ_0^2	0	
School	σ_1^2	1.26	

Table 5: The parameter estimates for the model fitted for the counts of the usage of *scientist* in an identity-presupposing way versus a generic way.

and standard errors. The difference in deviances between the model with and without a treatment fact is 4.9, which gives a p-value of 3% for the asymptotic χ_1^2 distribution. The two-sided p-value for the parameter estimate of the treatment effect is 2%. We are in a situation with only 39 observations, so we must be careful with conclusions drawn from these approximations. What is notable about these p-values is that they are much larger than the ones reported in the study. The authors reported a two-sided p-value of 0.1%. However, they considered only the response of the counts of *scientist* as identity, and did not consider those of *scientist* used in a generic way. They assumed the response followed a negative binomial distribution. Their analysis did not acknowledge the fact that *scientist* was categorized as **one** of identity **or** generic, and their analysis included **all** the teachers who had zero counts for *scientist* in the identity sense in the transcript of their lesson. This second part is problematic because it does not acknowledge the fact that 76 of these teachers **also** did not have any counts of *scientist* in the generic form, leaving only 39 teachers who had nonzero counts for *scientist* used in an identity-presupposing way or a generic way. With that in mind, I am not sure the teachers with zero counts for both ways of using *scientist* should be used as evidence for a treatment effect, given their assumptions on the exclusivity of the categories for instances of *scientist*. In short, teachers did not use the word *scientist* much in the first place, and not noting this fact led to the authors' conclusions which may be a bit dubious. In the re-analysis here, we find that there may be a treatment effect in the same direction as the authors claim, but the analysis suggests that if it exists, it is fairly small, which is counter to what the authors claim.

5.3.6 Model specification for use of *observe, predict* and *check*

This part of the section concerns the usage of the words *observe, predict* and *check*. These words were of interest to the authors because they were implicitly emphasized as the steps of doing science in the treatment videos they distributed to the schools. These words were not emphasized in the control videos, and this fact suggests that looking at the counts of these words can serve as some measure of the treatment's efficacy. As with the other linguistic variables, these variables are counts: the researchers counted the number of instances of *observe, predict* and *check* in a transcript of a teacher's lesson. The main difference is that these words were not further classified by the researchers. With this knowledge and with not strong reason at first to believe the treatment would effect each one of these different, I am interested in the sum of these responses as the response of interest. Although the authors do not explicitly model this response, they allude to the counts of these words together being an indicator of treatment efficacy. The response is then Y_i corresponding to the total number of instances of *observe, predict* and *check* in the transcript of a teacher i's lesson. Again, I assume random effects that are all independent of one another and follow a normal distribution with zero mean and variance equal within each level of the hierarchy. Conditional on the random effects, I assume the response follows a Poisson distribution.[4] Note that I will include here the intermediate response of time as a covariate in addition to the intercept and the treatment factor as predictors in the logarithm parameterization of the response mean λ_i. The parameters in the model are estimated by maximum likelihood, and the Laplace approximation is used to compute this likelihood. The model specification is summarized below.

[4]I did consider using a linear model with a log-transformed response, but there was a teacher who did not use any of these words, giving a zero count. Though omitting just one observation is not something to be too concerned about when there are 116 observations, there is also not a convincing reason to believe the response should always be strictly positive. The conclusions, though not presented, are the same, so in this situation, it does not matter.

$$\eta_i = \mu + \beta_0 \cdot length(i) + \beta_1 \cdot treat(i) + \tau_{d(i)} + \tau_{s(i)} \qquad \lambda_i = e^{\eta_i} \qquad Y_i \Big| \tau_{d(i)}, \tau_{s(i)} \sim \text{Pois}(\lambda_i)$$

$$\tau_{d(i)} = \begin{cases} \tau_{d_1} \sim N(0, \sigma_0^2) & \text{if } d(i) = 1 \\ \vdots & \\ \tau_{d_{n_d}} \sim N(0, \sigma_0^2) & \text{if } d(i) = n_d \end{cases} \qquad \tau_{s(i)} = \begin{cases} \tau_{s_1} \sim N(0, \sigma_1^2) & \text{if } s(i) = 1 \\ \vdots & \\ \tau_{s_{n_s}} \sim N(0, \sigma_1^2) & \text{if } s(i) = n_s \end{cases}$$

All random effects are independent of one another.

The choice to include length in the model is due to positive correlation between a teacher's lesson and overall count of the words *observe, predict* and *check* — i.e. teachers who had given a longer lesson typically used more words. Though I do not present a formal analysis of this, we can see the plot below in Figure 2, which is one where we plot the logarithm of the total counts of the science words *observe, predict, check, science* and *scientist* against lesson length in minutes. Since we do not have access to the actual transcripts of a teacher's lesson, this total count serves as a rough approximation for the total count of words a teacher used in their lesson. The acceptability of including length in the linear predictor is due to the fact that there was not evidence to suggest the treatment had any effect on the lesson length, so it presumably will not serve as a confounding variable, allowing for a more nuanced, yet still legitimate analysis of the treatment effect. I did consider an analysis where it was not included in the model, and the conclusions remain unaffected.

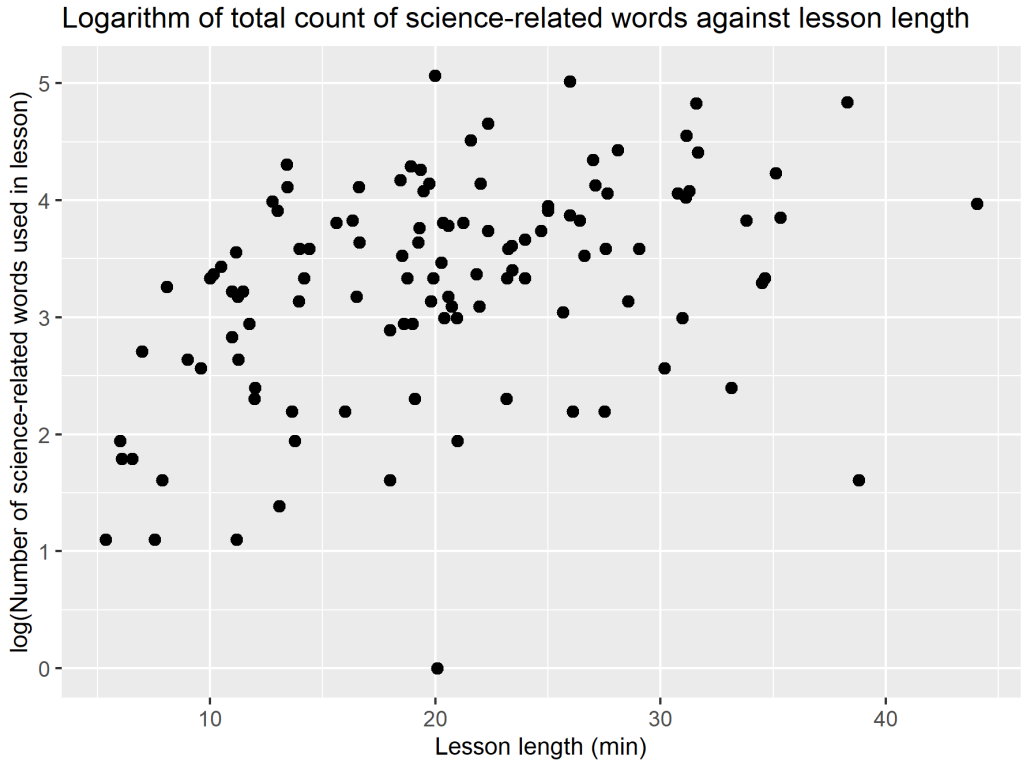

Figure 2: The total counts of science-related words used by teachers is positively correlated with the lesson length. Here the logarithm of the count is plotted to highlight this multiplicative relationship.

5.3.7 Conclusions on the treatment effect on *observe, predict* and *check*

Below in Table 6 are the parameter estimates for the model fitted. The estimates for the variance components σ_0^2 and σ_1^2 are quite small at 0.01 and 0.18. What is worth noting here is that the variation at the school level is substantially larger than that at the district level. The estimated effect of the treatment is 0.68 with a standard error of 0.14; the estimate for the effect is much larger than the effect of lesson length, which has an estimate of 0.04 and a standard error of 0.003. The inclusion of lesson length reduces slightly the estimate for the treatment effect, which is estimated to be 0.70 with a standard error of 0.13 when length is excluded from the model. In either case, the estimated treatment effect is much larger than any estimated variance component and the estimate for length, and it is positive, giving evidence in support of the efficacy of the treatment. The difference in deviances between the model with and without the treatment effect is 19.8, which is far out in the right tail of the asymptotic χ_1^2 null distribution.

Name	Parameter	Estimate	Standard error
Intercept	μ	1.99	0.12
Length	β_0	0.04	0.003
Treatment	β_1	0.68	0.14
Level	Variance component	Estimate	
District	σ_0^2	0.01	
School	σ_1^2	0.18	

Table 6: The parameter estimates for the model fitted for the counts of the usage of *observe, predict* and *check*.

5.3.8 Model specifications for treatment effect on presence of *science* and *scientist*

As we have just seen above, there were a number of teachers who did not use *science* or *scientist* at all during their lesson. This leads to the question of whether or not the treatment had any effect on the presence of these words more generally; in other words, it is of interest to examine whether or not teachers exposed to the different treatment condition were more or less likely to use *science* or *scientist* at all. The analysis below relies on the assumption that all occurrences of *science* and *scientific* were classified as belonging to one of the categories previous mentioned, namely those of noun and action for *science* and identity-presupposing and generic for *scientist*. By all indications from the documentation of the study online, this seems to be the case, but we must keep in mind this assumption may turn out to be false.

Although I will fit two separate models for *science* and *scientist*, idea for each model is analogous, so I present just one. Just for the sake of concreteness, I consider the situation with *science*. The observational units are still the teachers, and the experimental units are still the schools. The response in these analyses are $Y_i \in \{0, 1\}$, where $Y_i = 1$ indicates that a teacher i used *science* at least once during their lesson. I assume the presence of random effects for district and school; these random effects are independent of one another and follow a normal distribution with mean zero and variance equal with each level of the hierarchy. Conditioned on these random effects, the response Y_i follows a Bernoulli distribution with mean π_i. The logits of this mean π_i are modeled linearly with an intercept and the treatment factor. All parameter estimates are obtained by maximum likelihood, where the likelihood is computed using the Laplace approximation. Below you can see an implicit specification of the model.

The model for *scientist* is exactly the same; only the referent of the response in the real world is different. I decided not to use length as a variable in the predictor because although there is no evidence to suggest it is confounded with treatment, there was also no evidence that it improved the fit of the models. The differences in deviances for the models with and without length for both *science* and *scientist* were 0.8, which is not strong evidence given the asymptotic χ_1^2 null distribution.

$$\eta_i = \mu + \beta_0 \cdot treat(i) + \tau_{d(i)} + \tau_{s(i)} \qquad \pi_i = \frac{e^{\pi_i}}{1 + e^{\pi_i}} \qquad Y_i \Big| \tau_{d(i)}, \tau_{s(i)} \sim \text{Bern}(\pi_i)$$

$$\tau_{d(i)} = \begin{cases} \tau_{d_1} \sim N(0, \sigma_0^2) & \text{if } d(i) = 1 \\ \vdots & \\ \tau_{d_{n_d}} \sim N(0, \sigma_0^2) & \text{if } d(i) = n_d \end{cases} \qquad \tau_{s(i)} = \begin{cases} \tau_{s_1} \sim N(0, \sigma_1^2) & \text{if } s(i) = 1 \\ \vdots & \\ \tau_{s_{n_s}} \sim N(0, \sigma_1^2) & \text{if } s(i) = n_s \end{cases}$$

All random effects are independent of one another.

5.3.9 Conclusions on the treatment effect on the use of *science* and *scientist* more generally

First, we consider the parameter estimates for scenario involving usage of *science*. The parameter estimates can be seen below in Table 7. Consistent with our observations above, at least one of the variance components in the model is quite large: the estimate for σ_0^2 is 0.32 and σ_1^2 is 1.76. This suggests substantial variability at the school level. The variability at the school level is almost as large as the estimated effect for treatment itself, which is 2.44 with a standard error of 0.92. Although the estimated treatment effect is large, its standard error is not negligible. With that said, the fit does appear to improve in a meaningful way when including the treatment factor, as the difference in deviances between the model with and without the treatment factor is 8.2. This gives a p-value of 0.4% which I would suggest is strong enough evidence to support the claim that the treatment did increase the odds of a teacher using the word *science* in general despite the large variability occurring at the school level.

Name	Parameter	Estimate	Standard error
Intercept	μ	1.18	0.57
Treatment	β_0	2.44	0.92
Level	Variance component	Estimate	
District	σ_0^2	0.32	
School	σ_1^2	1.76	

Table 7: The parameter estimates for the model fitted concerning the presence of any tokens of *science* in the transcript for a teacher's lesson.

Now, we consider the estimates in the other situation, the one concerning the presence of *scientist*. The parameter estimates are listed below in Table 8. Interestingly, this is the first situation with the linguistic responses where the variance components are estimated to be zero. The intercept estimate is actually of interest in this situation as well, as it is estimated to be 0.07 with a standard error of 0.26. Essentially, teachers in the control condition have a 50% chance of using *scientist* at all during their lesson on science, which is something the authors do not make clear in the motivation for their study. As we saw above, the use of *scientist* is not that common. However, there does appear to have been an effect of treatment on whether or not a teacher used *scientist* at all during their lesson. The estimate for the treatment effect is -1.36 with a standard error of 0.43. This goes in the opposite direction as the above effect, which is consistent with what the authors claim. The effect does not appear to be as large for *scientist* as it is for *science*. It does still appear to be real, though, as the difference in deviances between the models with and without the treatment effect is 10.8, which yields a p-value of 0.1% given the asymptotic χ_1^2 null distribution.

5.3.10 Model specification for the use of *science* versus *scientist*

The authors consistently contrasted *science* and *scientist*, suggesting that the latter is a word which is, but not necessarily, more likely to encourage a presupposing of the identity of a scientist on the part of a student. However, they never explicitly analyzed the data concerning these words from the transcripts jointly. It makes sense that they did not do this because although they are semantically related, the set of

Name	Parameter	Estimate	Standard error
Intercept	μ	-0.07	0.26
Treatment	β_0	-1.36	0.43
Level	Variance component	Estimate	
District	σ_0^2	0	
School	σ_1^2	1×10^{-15}	

Table 8: The parameter estimates for the model fitted concerning the presence of any tokens of *scientist* in the transcript for a teacher's lesson.

particular sentences one would expect *scientist* to show up is different than those one would expect *science* to show up. By way of example, we would not see the first sentence in the pair below (where semantic anomaly is traditionally denoted by # in linguistics), but we would see the second. The word *science* can not serve as the subject of the sentence, but *scientist* can.

Science do / does many experiments each year throughout the world.
Scientists do many experiments each year throughout the world.

Likewise, *scientist* would not be able to show up as the object of prepositional phrase headed by *in* for the sentence frame below.

Observation, prediction and checking are involved in science.
Observation, prediction and checking are involved in scientist(s).

However, if we take a step of towards the more abstract setting and consider the linguistic strategy for teaching a lesson, we could possibly argue these words to be distinct, exclusive outcomes of some process. More explicitly, a teacher can choose whether or not to introduce a particular concept, such as *prediction*, through the perspective of a scientist or through the perspective of a step involved in science, and these perspectives may at least roughly coincide with usage of *scientist* and *science* respectively. Admittedly, this is a tenuous assumption, but I at least think it is an interesting one and one which is worth entertaining, even for just this section.

If we assume that *science* and *scientist* are exclusive outcomes of some more abstract underlying process, we can analyze the data in the following way. First, we assume that the response Y_i is the number of *scientist* tokens in the transcript for a teacher i's lesson, given the total number of instances N_i of *science* and *scientist* combined. We assume the same random effect structure as we have throughout the preceding analyses. Then, conditioned on the random effects, the response is distributed as binomial with mean π_i and number of trials N_i. The logits of the mean is assumed to be linear in a intercept and the treatment effect. Parameters are estimated by maximum likelihood and this likelihood is computed using the Laplace approximation. The specification of the model is summarized below.

$$\eta_i = \mu + \beta_0 \cdot treat(i) + \tau_{d(i)} + \tau_{s(i)} \qquad \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \qquad Y_i \Big| \tau_{d(i)}, \tau_{s(i)}, N_i \sim \text{Binom}(\pi_i, N_i)$$

$$\tau_{d(i)} = \begin{cases} \tau_{d_1} \sim N(0, \sigma_0^2) & \text{if } d(i) = 1 \\ \vdots & \\ \tau_{d_{n_d}} \sim N(0, \sigma_0^2) & \text{if } d(i) = n_d \end{cases} \qquad \tau_{s(i)} = \begin{cases} \tau_{s_1} \sim N(0, \sigma_1^2) & \text{if } s(i) = 1 \\ \vdots & \\ \tau_{s_{n_s}} \sim N(0, \sigma_1^2) & \text{if } s(i) = n_s \end{cases}$$

All random effects are independent of one another.

5.3.11 Conclusions on the treatment effect on the use of *science* versus *scientist*

A table of the parameters in shown below. The estimates for the variance components for the district and school levels, σ_0^2 and σ_1^2, are 0 and 4.76 respectively. The estimated variance component for schools has a very large estimate, much larger than the estimate for the treatment effect which is -3.59 with standard error 0.84. Although there is this variation at the school level, the treatment effect does to have had some effect on the usage of *science* versus *scientist*, where teachers in the treatment condition were less likely to use *scientist*. The magnitude of this effect is hard to evaluate given the standard errors and variation, but the treatment does appear to have had the desired effect; that is if one accepts the premises for fitting this model, which are not innocent.

Name	Parameter	Estimate	Standard error
Intercept	μ	-0.44	0.53
Treatment	β_0	-3.59	0.84
Level	Variance component	Estimate	
District	σ_0^2	2.9×10^{-9}	
School	σ_1^2	4.76	

Table 9: The parameter estimates for the model fitted concerning the usage of *science* versus *scientist* in the transcript for a teacher's lesson.

5.4 Conclusions for the experiment

From the analysis done here, there is not overwhelming evidence that the treatment had any effect on student persistence. If it had any effect, it was small. The authors had similar results in terms of the two-sided p-value of the estimate for the treatment effect, where theirs was 4% and the one we see here is 6%, but they were much more confident in the existence of the treatment effect. Perhaps that is somewhat due to the fact that their reported p-value is less than the typical 5% threshold for Type I error, but they expressed their confidence in terms of the experimental design. They said that the fact their p-value was near the threshold at all was convincing evidence of a treatment effect given that the experimental design makes it hard to detect any effect at the level of a individual child. Here we have seen that a slightly more complex model, which explicitly accounts for the censoring in the data, shows the instability of their conclusions. Further, the p-value of 6% is obtained from a likelihood ratio test statistic, which is known to be anti-conservative, meaning that even this estimate could be generous. For these reasons, I believe there is not sufficient evidence to claim a treatment effect that is not due merely to random variation. At the very least, I think a follow-up experiment should be carried out to further study it.

The analysis here does largely agree with the authors' in terms of the treatment efficacy. The authors reported that the treatment had the effect of increasing a teacher's usage of *science* in an action way as well as increasing the overall usage of the steps of science, which they considered to be *observe, predict* and *check*. The analysis here agrees with both of those claims, as the evidence seems to be fairly strong for an effect. This is not entirely surprising, though, because although the treatment videos did not explicitly coach teachers on the language to use in the lesson, it is very evident that there is an implicit emphasis on exactly those words after having watched the videos myself. The part of the authors' analysis which is different than the analysis here is the strength, or even existence, of a treatment effect on the usage of *scientist* in an identity-presupposing way versus a generic way. The authors claimed that the treatment had a large effect by decreasing use of *scientist* in an identity-presupposing way, but I do not think the effect is as large as they suggest. They were led to their conclusions by including all of the data, but I argue that not all of the data is actually informative on this fact. When we look at the subset of data which is most informative, we see that this effect is very small, if it even exists. Although this is a small detail in the overall set of things one can analyze in this experiment, it is an important detail because their hypothesis is more or less pegged mostly to the idea that identity-presupposing language is problematic, and there is not convincing evidence in my opinion that the treatment had an effect on this. Given that this language appears to be rare anyways, which is contrary to what the authors imply, it may not be too surprising that the effect does not exist or is small.

The treatment did have a more general effect on the use of *scientist*, though. Although the analysis here could not detect a meaningful difference in the different categories for *scientist*, we are able to reasonably conclude that the treatment biased teachers toward producing less tokens of *scientist*. We also find evidence to support the claim that the treatment did increase the chances of a teacher producing *science* during their lesson as well. Perhaps most interestingly, the analysis here indicates that the most prominent effect the treatment had was the contrast on the presence of *science* versus *scientist* in a teacher's lesson. As noted above, these conclusions rely on the assumption that *science* and *scientist* are the two distinct outcomes of some more semantically abstract process, but it is at least interesting to think about. If we accept this assumption, it appears that teachers in the treatment condition were less likely to use the archetype of a scientist as a vehicle for their instruction, instead using the idea or philosophy behind the scientific process to teach. Although the authors' were interested in the the particular type of reference to *scientist*, namely the identity ones, I think this bit is of interest.

Overall, I recommend that we reconsider the conclusions from the experiment, and I would suggest that there is not evidence for any effect of the treatment on the behavior of students. There is a good amount of evidence for a treatment effect on the linguistic behavior of teachers, but not exactly in the way the researchers had stated in their paper. It is hard to evaluate exactly how large the treatment effects are because the standard error on the logits scale is quite large, and there are also large estimates for the variance components, especially at the level of the schools. To get the desired linguistic outcomes, it may be beneficial for the researchers to have a treatment video which more explicitly coaches teachers on the use of certain science-related words, especially *scientist*.

6 Analysis of persistence from the perspective of an observational study

The authors were also interested in how a teacher's language usage related to students' persistence outcomes. Questions along these lines are considered in this section. The important distinction to make here is that study of teacher linguistic behavior and student persistence outcomes is observational. The treatment was the training video a teacher received; although this training video implicitly emphasized particular language, there is no notion of language being assigned to teachers. We see in section §5 that this training was effective in biasing the linguistic behavior of teachers, but this linguistic behavior was a response in the context of the experiment. Here, we consider these intermediate variables to be covariates, as they play a role in parameterizing the probability distributions which we fit to the data. This change can be seen as fixing a different point in time to be the baseline, namely any time after the student had the lesson on friction and prior to the time which a student had completed the video game activity to measure their persistence.

6.1 Structure in the units (students)

For the observational study, we include additional structure on the units — the language use of a child's teacher. This is an intermediate response that we do not include the analysis of the experiment, as those effects would be confounded with that of the treatment. The covariates for the observational study analysis are in Table 10.

Unfortunately, we do not have access to the transcripts, and this raises some slight concerns regarding these measurements; however, this data should still be informative on some level. We just need to exercise caution with conclusions. As mentioned above, the classification of *science* and *scientist* was exclusive: *science* was either classified as used in an action way or a noun way; *scientist* was classified as either identity-presupposing or generic. This structure turns out to be important, as it leads to different conclusions than those found by the authors.

6.2 Transformations of the covariates

Using the linguistic variables separately does not make important use of the structure in the units. First, and most importantly, it does not capture the fact that uses of *science* and *scientist* had two exclusive classes. Second, it does not account for the fact that the counts are positively correlated with lesson length, as the plot in Figure 2 suggests. It seems fair to say that we would not want to treat a student who heard *scientist*

Name	Description	Variable type
scienAction	Number of times the child's teacher used *science* in an action construction e.g. *Let's do* **science***!*, *Doing* **science** *is fun!*	Quantitative variable
scienNoun	Number of times the child's teacher used *science* in a noun construction e.g. *It's* **science** *time!*, *This is a* **science** *question.*	Quantitative variable
scienGeneric	Number of times the child's teacher used *scientist* in a generic way e.g. **Scientists** *work hard to solve problems.*, **Scientists** *observe.*	Quantitative variable
scienIdentity	Number of times the child's teacher used *scientist* in an identity-presupposing way e.g. *Today, we are going to be* **scientists***!*, *Let's put our* **scientist** *hats on!*	Quantitative variable
sayObserve	Number of times the child's teacher said *observe*	Quantitative variable
sayPredict	Number of times the child's teacher said *predict*	Quantitative variable
sayCheck	Number of times the child's teacher said *check*	Quantitative variable

Table 10: Additional covariates available for the analysis of the observational study. These are responses of the experiment. The first two variables and the third and fourth variables are grouped together because coders were instructed to choose between scienAction and scienNoun, as well as between scienGeneric and scienIdentity.

ten times in a ten minute lesson the same as a student who heard *scientist* ten times in a forty minute lesson. In short, transformations should at least be considered to address this important structure in the units.

I consider the following transformations: (i) differencing the counts of *science* and *scientist*; and (ii) dividing all of these counts by the lesson length. These are not gold standard transformations; they are just an approximation given the data we have. One may think to take ratios rather than differences. I have chosen differences because there are zero counts for all of the linguistic variables, so for those which are paired it is not feasible to take the ratio. One may also be skeptical about the effective weighting of these variables by lesson length, as lesson length may not be the best way to weight the counts. I am very sympathetic to this argument; however, since we do not have access to the actual transcripts for each teacher's lesson, this appears to be one of the only ways to do some weighting, which as I have argued above is desirable in an analysis. Interestingly, we find that the type of transformation considered here, whether it be differencing, weighting or both, does not matter; it is just important that we do **some** type of transformation. The authors did not consider transformations of these variables, and it turns out that this could have led them to conclusions which may not be tenable. Examples of the mentioned transformations are shown below with the names I use to reference them. The notation is that i indicates a student who received a lesson from teacher j. I only write it in the first example below so it does not become distracting for the reader, and I suppress the subscript j in the notation in the next section, but keep in mind these covariates do make reference to a teacher as well as the student.

$$\text{Science}_{ij} = \frac{\text{scienceAction}_{ij} - \text{scienceNoun}_{ij}}{\text{length of lesson}_{ij}} \qquad \text{Scientist}_{ij} = \frac{\text{scientistIdentity}_{ij} - \text{scientistGeneric}_{ij}}{\text{length of lesson}_{ij}}$$

$$\text{scienceDiff} = \text{scienceAction} - \text{scienceNoun} \qquad \text{scientistDiff} = \text{scientistIdentity} - \text{scientistGeneric}$$

$$\text{actionDiv} = \frac{\text{scienAction}}{\text{length of lesson}} \qquad \text{identityDiv} = \frac{\text{scientistIdentity}}{\text{length of lesson}}$$

6.3 Model specification for student persistence in the context of an observational study

We can again consider the data in this setting to be of the discrete-time survival sort. Therefore, I assume the same random effect structure and the same idea for the distribution of the response Y_{it} when conditioned on these random effects. Finer details on this part of the model can be gleaned from notation below and can be seen in prose in section §5.1. The major change in the setting of the observational study occurs in the parameterization of the logits of the hazard function $h(t; \mathbf{x}_i)$, the mean π_{it} of the Bernoulli distribution of Y_{it} conditional on the random effects. We use the linguistic covariates in the setting of an observational study and we omit the treatment effect. For simplicity, we consider models where we only look at one of these transformed variables at a time, but this does not make a difference in the conclusions reached.

One last, but important, note is that we include an intercept 'noScience' which indicates whether or not a student's teacher did not say *science* at all during their lesson; similarly, the intercept 'noScientist' is defined. Only when we consider the transformed variables concerned with usage of *science* is the intercept noScience included; we do the same for noScientist in the case of transformed variables concerning the usage of *scientist*. These intercepts are important to account for a particularly interesting detail of the structure in the units: zero counts for both categories of *science* or *scientist* is presumably not the same as equal counts of for each category of *science* and *scientist*. When considering differences, these situations coincide, so I think it is important to distinguish between the two. Parameters for all the models are estimated by maximum likelihood and the likelihood is computed using the Laplace approximation. Below I show a 'prototypical' model, whereby scienceVariable is a placeholder for any of the transformed variables concerning *science* and similarly for scientistVariable.

$$\eta_{it} = \alpha_t + \beta_0 \cdot noScience + \beta_1 \cdot scienceVariable + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)} \qquad \pi_{it} = \frac{e^{\eta_{it}}}{1 + e^{\eta_{it}}}$$

or

$$\eta_{it} = \alpha_t + \beta_0 \cdot noScientist + \beta_1 \cdot scientistVariable + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)}$$

$$\begin{aligned} \tau_{d(i)} &\sim N(0, \sigma_0^2) \\ \tau_{s(i)} &\sim N(0, \sigma_1^2) \qquad \leftarrow \text{All random effects independent of one another.} \\ \tau_{c(i)} &\sim N(0, \sigma_2^2) \end{aligned}$$

$$Y_{it} \Big| \tau_{d(i)}, \tau_{s(i)}, \tau_{c(i)} \sim \text{Bern}(\pi_{it})$$

Other covariates such as gender and nonwhite do not improve the fit, so they are omitted from any of the models.

6.4 Conclusions on student persistence in the context of an observational study

There are three tables of parameter estimates below. In Table 11 we have the estimates for the hazard function in the model where there are no linguistic covariates included. I include these just to serve as a

point of reference for interpretation. I also included the maximum estimated variance component among the models. All of the estimates were small, so reporting the maximum may not be the canonical thing to do, but I think the little information this glosses over is worth the possibility of all the near zero estimates distracting the reader. The largest variance component estimate occurs at the school level in all models fitted, with a maximum estimate of σ_1^2 being 0.07. The estimates for the hazard function exhibit the same trends as before, which is not surprising, and these estimates do not change appreciably with each model fit.

Name	Parameter	Estimate	Standard error
$h(t=0)$ logits	α_0	0.74	0.11
$h(t=1)$ logits	α_1	-0.35	0.18
$h(t=2)$ logits	α_2	-0.27	0.22
$h(t=3)$ logits	α_3	-0.93	0.32
$h(t=4)$ logits	α_4	-0.47	0.35
$h(t=5)$ logits	α_5	-1.27	0.51
Level	Variance component	Max estimate	
District	σ_0^2	6.5×10^{-8}	
School	σ_1^2	0.07	
Class	σ_2^2	2.3×10^{-7}	

Table 11: The parameter estimates for the model with no linguistic covariates. The estimates of the hazard function are similar for all models fitted. There is not much variance at the district or school level.

More interesting for this part of the report are the p-values of the likelihood ratio statistic. This ratio is between each of models which consist of the transformed variables and the corresponding noScience or noScientist intercept and the model which just contains intercepts for the hazard function. The p-values are perhaps of more interest than the parameter estimates themselves at this particular moment because their interpretation does not depend on transformations.

What is important to note about these log-likelihood ratio statistics is that their asymptotic null distribution is χ_2^2 instead of χ_1^2. It arguably makes most sense to consider the intercept and each transformed variable as a pair: this is because the intercept is serving as a way to explicitly model the difference between the non-existence of *science* or *scientist* counts for both categories and the existence of these counts in each category but the canceling nature of the transformations. Both of the intercepts are estimated at various points near zero in each model fit to the data with a standard error approximately 0.2 in all models for both intercepts.

Table 12 shows the estimates for the parameters of the transformed linguistic variables. The estimates themselves are not straightforward to interpret, but they are still reported. What is of interest is whether or not they improve the fit in a meaningful way. In all cases, there is not strong evidence to suggest that any of these linguistic variables have a relationship to student persistence. Perhaps the usage of *science* relates to student persistence in some way, but it is not clear. There is essentially no evidence for a relationship between usage of *scientist* in an identity-presupposing way and student persistence, which is contrary to what the authors report. They reported a meaningful relationship whereby decreased usage of *scientist* in an identity-presupposing way is related to greater student persistence. When transforming the variables, there is no such evidence for that relationship.

Intercept used	Name	Estimate	Standard error	p-value of LLR (χ_2^2)
noScience	Science	-0.39	0.14	0.03
	ScienceDiff	-0.02	0.007	0.08
	ScienceDiv	-0.31	0.15	0.13
noScientist	Scientist	-0.28	1.08	0.52
	ScientistDiff	0.01	0.05	0.53
	ScientistDiv	1.19	1.31	0.36

Table 12: The parameter estimates for the linguistic covariates (transformed) in the observational study. All were fit with the intercept noScience or noScientist.

To make a more direct comparison to the original paper, we can consider the non-transformed linguistic variables. The authors did not transform the linguistic variables, and it is of interest whether or not the model family or the choice of transformations in parameterizing a chosen family of distributions is the reason for the misalignment of conclusions. In Table 13, we see estimates which help us make sense of these differences in conclusions. I consider two scenarios: one where we use just scienAction and just scienIdentity as a covariate, just as the authors did; the other where we use these paired with the noScience and noScientist intercepts respectively, which I suggest may be more appropriate. The asymptotic null distributions when comparing the the model with just intercepts for the hazard function are different. In the first situation, this is a chi-squared distribution on one degree of freedom, and in the second it is a chi-squared distribution on two degrees of freedom, as we have the intercept and the quantitative variable paired. In the most direct comparison to the study, which is the top section of Table 13, we see the conclusions are similar to those of the study, but much more tempered. This model would suggest that there is no relationship between usage of *science* in an action way and student persistence, and it would suggest that there is **also** no relationship between usage of *scientist* in an identity-presupposing way and student persistence. It is this second part which is contrary to the conclusions of the authors. They reported that usage of *scientist* in an identity-presupposing way was associated with less student persistence; they reported a two-sided p-value of 1% for the parameter estimate. We have reason to believe considering scienIdentity on its own without scienGeneric is potentially problematic, or at least not sufficient. With the argument for transformation of this variable in mind alongside the argument for a model which at least takes into account the censoring, we see that the authors conclusions may be a bit optimistic. Further, we can see that when including the noScientist intercept the authors' conclusion becomes even more suspect. Inclusion of the intercept noScience does not change any conclusion regarding action-based usages of *science* and its relationship with student persistence. The last thing worth noting in this comparison is that the inclusion of the intercepts noScience or noScientist do not improve the fit in a meaningful way over using just scienAction or scienIdentity. Evidence comes from likelihood ratio statistics, as they are 0 and 2.8 respectively. This suggests that although it may seem important in principle to include these intercepts, there is not strong evidence in this data they matter too much.

Name	Estimate	Standard error	p-value of LLR	Asymptotic null of LLR
scienAction	-0.01	0.008	0.13	χ^2_1
scientistIdentity	0.09	0.04	0.05	
scienAction with noScience	-0.02	0.007	0.08	χ^2_2
scientistIdentity with noScientist	0.11	0.07	0.13	

Table 13: The parameter estimates for the linguistic covariates (non- transformed) in the observational study, comparing the analysis here and the analysis in the original study.

Overall, there does not seem to be any strong evidence to support claims that there is a relationship between the linguistic variables under consideration by the researchers and student persistence. The form of the model slightly changes the conclusions from the authors, but transforming the variables drastically changes the conclusions. To better understand the relationship between a teacher's language and student persistence, a more nuanced analysis must be done, but even the analysis here suggests that this relationship may be weaker than the researchers thought. We may be fairly confident in the absence of evidence in this experiment given that we reach essentially the same conclusions with the set of transformations considered here. Perhaps a more insightful transformation could help provide clarity on the relationship if it is there.

7 Accessing the article and data

- **Article**: Asking young children to 'do science' instead of 'be scientists' increases science engagement in a randomized field experiment
- **Article link**: Found here https://www.pnas.org/content/117/18/9808.
- **Data availability**: Data is freely available online. Found here https://osf.io/pe7k5/.

References

[1] Marjorie Rhodes, Amanda Cardarelli, and Sarah-Jane Leslie. Asking young children to "do science" instead of "be scientists" increases science engagement in a randomized field experiment. *Proceedings of the National Academy of Sciences*, 117(18):9808–9814, 2020.