Does asking young children to 'do science' instead of 'be scientists' increase science engagement in a randomized field experiment? — Let's see the re-analysis!

Brandon Rhodes

November 30, 2020

Roadmap

1. Recap on experiment and measurements
2. Analysis of linguistic behavior on persistence (observational study)
3. Closing thoughts
4. Analysis of treatment on persistence (experiment)

Recap on experiment and measurements

Questions

- Does use of action-language — such as ‘let’s do science’ — versus identity-presupposing language — such as ‘let’s be scientists’ — increase preschool students’ engagement in learning about friction?

Basic facts

- Experiment carried out by New York University in collaboration with New York City Department of Education
- All school were public pre-kindergarten ones, coming from 11 different districts.
- Randomization was done within each district.
  (Done at district level to prevent interference.)

The experiment

- Students were learning about friction.
- Teachers were either trained with videos that have an implicit emphasis on action-based language or videos where there is no such emphasis.
- Treatment videos contained specific examples of action-based language through an example of a teacher giving the lesson to a preschool class.

A picture is worth many words . . .

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

The measurements

- **Persistence**: measurement on how long students continued to play after receiving negative feedback in a video game resembling their friction lesson.

PERSISTENCE

- **Teacher linguistic behavior**: count data — number of tokens for the words *science*, *scientist*, *observe*, *predict* and *check* in the transcript.

TRAINING EFFICACY

Persistence game scheme

Trial 1: Student plays, rigged to be correct.

Trial 2: Student plays, rigged to be wrong.

NARRATOR: DO YOU WANT TO KEEP PLAYING, OR DO SOMETHING ELSE? Student chooses Y/N

Trial 3: If student answered yes, student plays. No feedback provided.

NARRATOR: DO YOU WANT TO KEEP PLAYING, OR DO SOMETHING ELSE? Student chooses Y/N

⋮

Trial 6: If student answered yes, student plays. No feedback provided.

Measurements on the teachers

- Teachers had their lessons recorded (just audio) and then transcribed.
- Two research assistants, blind to the treatment, recorded the number of tokens of *science* and *scientist* and categorized them.
- Four categories
  - Scientists study friction. — GENERIC USE OF *scientist*
  - Let's be scientists. — IDENTITY-PRESUPPOSING USE OF *scientist*
  - Let's do science. — ACTION USE OF *science*
  - It's science time. — NOUN USE OF *science*
- Research assistants also recorded the number of tokens of *observe*, *predict* and *check* in the transcript of a teacher's lesson.

A diagram of the experimental setup

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

Strategy of the analyses in terms of the diagram

Analysis of treatment on persistence (experiment)

Treatment → Teacher language → Persistence

INTERMEDIATE RESPONSE PRIMARY RESPONSE

Analysis of treatment on linguistic behavior (experiment)

Treatment → Teacher language → Persistence

INTERMEDIATE RESPONSE PRIMARY RESPONSE

Analysis of linguistic behavior on persistence (observational study)

Treatment → Teacher language → Persistence

INTERMEDIATE RESPONSE PRIMARY RESPONSE

Analysis of linguistic behavior on persistence (observational study)

Treatment $\rightarrow$ Teacher language $\rightarrow$ Persistence

INTERMEDIATE RESPONSE PRIMARY RESPONSE

Structure in the units

- **Block factors**: District, school and teacher / class NESTED
- **Available covariates**: QUANTITATIVE VARIABLES
  - Number of tokens for *scientist*, where each token is classified as either GENERIC or IDENTITY
  - Number of tokens for *science*, where each token is classified as either NOUN or ACTION
  - Number of tokens for *observe*, *predict* and *check*

Modeling strategy

I chose to treat the data as that of the discrete-time survival sort, leading to a modeling of the hazard function.

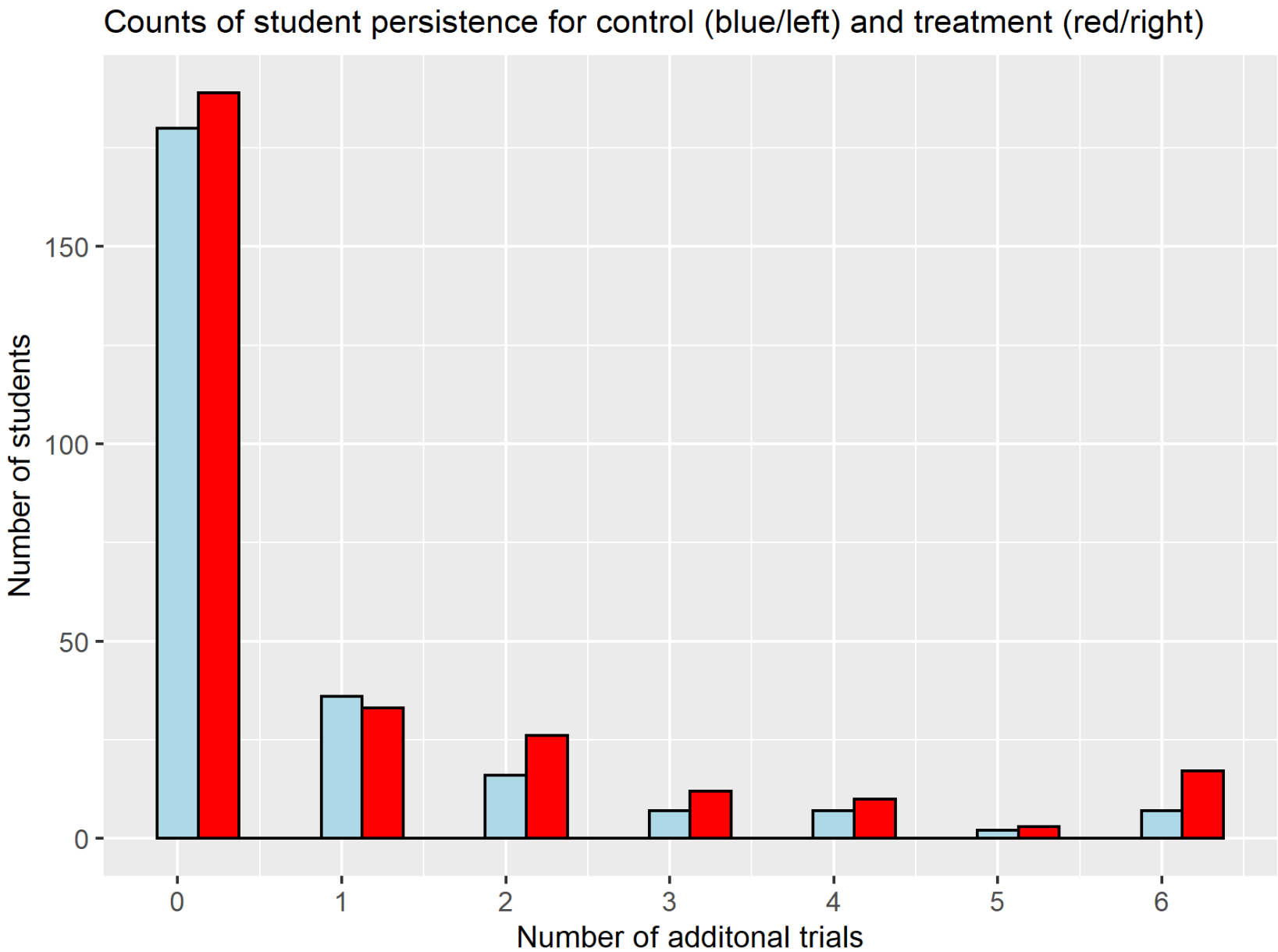
- **Observational unit**: child, trial pairs (i, t)
- **Response**: Binary variable $Y_{it} \in \{0, 1\}$ indicating whether or not a child i stopped playing the video game $(Y_{it} = 1)$ after trial t
- **Hazard function**: $h(t; \mathrm{x}_i) = \mathbb{P}(Y_{i,t} = 1 | Y_{i,t-1} = 0)$

Note that the probability a child i quits after trial t is the following:

$$\mathbb{P}(\text{child } i \text{ plays } t \text{ additional trials}) =$$

$$[1 - h(0; \mathrm{x}_i)] \cdots [1 - h(t-1; \mathrm{x}_i)] \cdot h(t; \mathrm{x}_i)$$

Informal justification of this strategy



Model specification

An implicit specification of the set of probability distributions under consideration. The functions d, s, c return the district, school and class for a child i.

$$\text{logits}[h(t; x_i)] = \eta_{it} = \beta^T x_{it} + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)}$$

$$\pi_{it} = \frac{e^{\eta_{it}}}{1 + e^{\eta_{it}}}$$

$$\begin{aligned} \tau_{d(i)} &\sim N(0, \sigma_0^2) \\ \tau_{s(i)} &\sim N(0, \sigma_1^2) \\ \tau_{c(i)} &\sim N(0, \sigma_2^2) \end{aligned} \qquad \text{all random effects iid}$$

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

Transformations of linguistic covariates

- **Use rates**: The linguistic covariates increase as a function of lesson length. Divide by lesson length to get the rate. (Informal motivation: the range for scienceAction is 0–68.)

  e.g. for teacher i, $\quad \text{sayPredict}_i = (\#\text{ of } \textit{predict} \text{ tokens in lesson})_i / \text{ lesson length}_i$

- **Sum** *observe*, *predict* **and** *check*: For each teacher, take the sum of these numbers. It leads to a simpler model which is easier to interpret.

  $\text{saySum} = \text{sayObserve} + \text{sayPredict} + \text{sayCheck}$

- **Difference of** *science*–ACTION **and** *science*–NOUN: Recall that tokens of *science* were classified as either ACTION or NOUN. Use this information by taking difference.

  $\text{Action} = \text{sayScienceAction} - \text{sayScienceNoun}$

- **Difference of** *scientist*–IDENTITY **and** *science*–GENERIC:

  $\text{Identity} = \text{sayScientistIdentity} - \text{sayScientistGeneric}$

Additional notes on the transformations

I did not consider ratios for the different categories of *science* and *scientist*: there were many zero counts, leading to many undefined ratios.

Note that saySum and Action are positively correlated (≈ 0.7), so I did not include both in a model at once.

Model specification

Parameters were estimated by maximum likelihood and the likelihood was computed using the Laplace approximation.

(`glmer` in R was used)

Model 1: Action-based versus Identity-presupposing language

$$\text{logits}[h(t;\mathrm{x}_i)] = \eta_{it} = \alpha_t + \beta_0 \cdot \text{Action} + \beta_1 \cdot \text{Identity} + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)}$$

Model 2: Specific (?) language

$$\text{logits}[h(t;\mathrm{x}_i)] = \eta_{it} = \alpha_t + \beta_0 \cdot \text{sumSay} + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)}$$

Action-based versus Identity-presupposing language

Name	Parameter	Estimate	Standard error
$h(t=0)$ logits	α_0	0.80	0.11
$h(t=1)$ logits	α_1	-0.31	0.17
$h(t=2)$ logits	α_2	-0.23	0.22
$h(t=3)$ logits	α_3	-0.91	0.32
$h(t=4)$ logits	α_4	-0.49	0.34
$h(t=5)$ logits	α_5	-1.27	0.50
Action	β_0	-0.39	0.13
Identity	β_1	0.25	0.81
Level	**Variance component**	**Estimate**	
District	σ_0^2	0	
School	σ_1^2	0	
Class	σ_2^2	0	

Specific (?) language

Name	Parameter	Estimate	Standard error
$h(t=0)$ logits	α_0	0.76	0.17
$h(t=1)$ logits	α_1	-0.34	0.22
$h(t=2)$ logits	α_2	-0.27	0.26
$h(t=3)$ logits	α_3	-0.93	0.34
$h(t=4)$ logits	α_4	-0.46	0.36
$h(t=5)$ logits	α_5	-1.26	0.52
sumSay	β_0	0.0009	0.09
Level	**Variance component**	**Estimate**	
District	σ_0^2	0	
School	σ_1^2	0.08	
Class	σ_2^2	0	

Overarching conclusions

- **Discouragement (or boredom)**: Students were more likely (conditional odds 2:1 — 3:1) to *not* play any additional trials of the video game.

- **Die-hards**: Students who made it through the last additional trial were more likely (conditional odds 2:1 – 6:1) to continue playing the video game.

Targeted conclusions

- **Action-based, but not necessarily specific**: Action-based language may increase students' persistence, but an emphasis on the specific steps in the scientific method may not.
  (LLR of 7.1 for Action; LLR of 0.1 for Identity; asymptotic null distribution is χ^2_1)

- **No — weak effect of identity-presupposing language**: There is not much evidence for an effect due to identity-presupposing language ('Let's be scientists!').
  (LLR of 0 for sumSay (!); asymptotic null distribution is χ^2_1)

These conclusions are at odds with those of the study. The authors of the study report no effect due to action-based language and a negative effect of identity-presupposing language.

Additional notes on the conclusions – 1

Recall: Action and Identity are defined (respectively) as

$$(\#\text{ of } \textit{science}\text{-ACTION tokens} - \#\text{ of } \textit{science}\text{-NOUN tokens in lesson})_i / \text{lesson length}_i$$

$$(\#\text{ of } \textit{scientist}\text{-IDENTITY tokens} - \#\text{ of } \textit{scientist}\text{-GENERIC tokens in lesson})_i / \text{lesson length}_i$$

I reach the same conclusions if I do not divide the differences by the length of a teacher's lesson.

Additional notes on the conclusions – 2

Recall: Action and Identity are defined (respectively) as

$$(\#\text{ of } \textit{science}\text{-ACTION tokens} - \#\text{ of } \textit{science}\text{-NOUN tokens in lesson})_i / \text{lesson length}_i$$

$$(\#\text{ of } \textit{scientist}\text{-IDENTITY tokens} - \#\text{ of } \textit{scientist}\text{-GENERIC tokens in lesson})_i / \text{lesson length}_i$$

Further, I reach the same conclusions if I do not take differences but divide by the length of a teacher's lesson.

Additional notes on the conclusions — 3

Recall: scienceAction and scientistIdentity are defined (respectively) as

$$\frac{\#\text{ of } \textit{science}\text{-ACTION tokens}}{}$$

$$\frac{\#\text{ of } \textit{scientist}\text{-IDENTITY tokens}}{}$$

I reach different conclusions — but the same as the authors' conclusions — if I only consider the number of tokens for each with no differencing from an opposing category or dividing by lesson length.

Closing thoughts

Closing thoughts

My intuition is that the effects are too small to detect reliably, so I am not sure any exist.

Choosing whether or not to transform the linguistic covariates is a somewhat important decision, but the effects are quite small relative to the baseline odds. That means even if the conclusions change slightly, we should not overstate anything.

Although not shown here, training was effective. Teachers who received treatment videos did produce more action-based and less identity-presupposing language. So, perhaps more targeted training would be beneficial.

Closing thoughts

I am considering a reinterpretation of the *persistence*. Since the students only receive negative feedback on the second of the initial trials (which is not recorded), they do not face any 'adversity' on the measured trials.

Another, perhaps more appropriate, measure of persistence could just be those students who decided to play *any* additional trials.

There is a contingent of the student population who are 'die-hards'. We can use this persistence data to examine how student interest relates to teacher language or how it was affected by the treatment.

Thank you for your time.

Table to summarize conclusions — action-based languaage

Recall:

scienceAction = # of *science*-ACTION

Action = (# of *science*-ACTION − # of *science*-NOUN)$_i$/ lesson length$_i$

ActionDiff = # of *science*-ACTION − # of *science*-NOUN

ActionRate = (# of *science*-ACTION)$_i$/ lesson length$_i$

Estimates for action-based language. All on the log-scale (log conditional-odds for BR and log rates-ratio for MR):

Author	Name	Est.	S.E.	Conclusion
BR	Action	−0.29	0.16	Slight decrease in quitting
BR	ActionDiff	−0.02	0.01	No — slight decrease in quitting
BR	ActionRate	−0.30	0.14	Slight decrease in quitting
BR	scienceAction	−0.01	0.01	No effect
MR	scienceAction	0.02	0.01	No effect

Table to summarize conclusions — identity-presupposing language

Recall:

scientistIdentity = # of *scientist*-IDENTITY

Identity = (# of *scientist*-IDENTITY − # of *scientist*-GENERIC)$_i$/ lesson length$_i$

IdentityDiff = # of *scientist*-IDENTITY − # of *scientist*-GENERIC

IdentityRate = (# of *scientist*-IDENTITY)$_i$/ lesson length$_i$

Estimates for identity-presupposing language. All on the log-scale (log conditional-odds for BR and log rates-ratio for MR):

Author	Name	Est.	S.E.	Conclusion
BR	Identity	0.25	0.81	No effect
BR	IdentityDiff	0.03	0.05	No effect
BR	IdentityRate	1.27	0.87	No effect
BR	scientistIdentity	0.09	0.04	Slight — no increase in quitting
MR	scientistIdentity	−0.08	0.04	Slight decrease in persisting

Analysis of treatment on persistence (experiment)

Treatment $\rightarrow$ Teacher language $\rightarrow$ Persistence

INTERMEDIATE RESPONSE PRIMARY RESPONSE

Structure in the units

- **Block factors**: District, school and teacher / class — NESTED

- **Available covariates**:
  child gender $\in \{\text{boy, girl}\}$ — CLASSIFICATION FACTORS
  trial $\in \{0, 1, 2, 3, 4, 5\}$

  class size $\in \mathbb{Z}_+$ — QUANTITATIVE VARIABLES
  nonwhite $\in [0, 1]$
  proportion of nonwhite students at school

I did not use teacher demographic information because many observations were missing.

Modeling strategy

I chose to treat the data as that of the discrete-time survival sort, leading to a modeling of the hazard function.

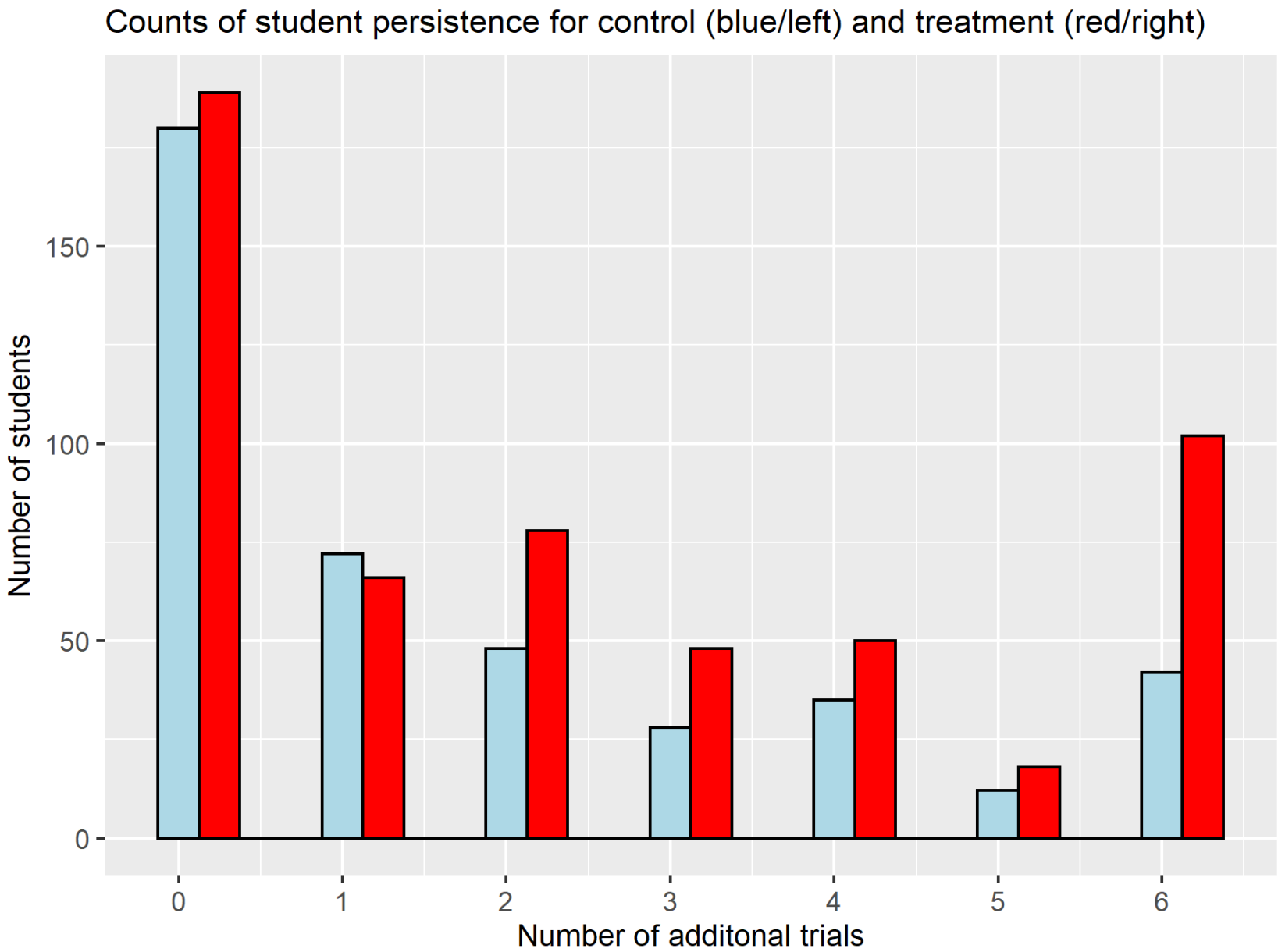
- **Observational unit**: child, trial pairs (i, t)
- **Response**: Binary variable $Y_{it} \in \{0, 1\}$ indicating whether or not a child i stopped playing the video game $(Y_{it} = 1)$ after trial t
- **Hazard function**: $h(t; \mathrm{x}_i) = \mathbb{P}(Y_{i,t} = 1 | Y_{i,t-1} = 0)$

Note that the probability a child i quits at trial t is the following:

$$\mathbb{P}(\text{child } i \text{ plays } t \text{ additional trials}) =$$

$$[1 - h(0; \mathrm{x}_i)] \cdots [1 - h(t-1; \mathrm{x}_i)] \cdot h(t; \mathrm{x}_i)$$

Informal justification of this strategy



Model specification

An implicit specification of the set of probability distributions under consideration. The functions d, s, c return the district, school and class for a child i.

$$\text{logits}[h(t; x_i)] = \eta_{it} = \beta^T x_{it} + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)}$$

$$\pi_{it} = \frac{e^{\eta_{it}}}{1 + e^{\eta_{it}}}$$

$$\begin{aligned} \tau_{d(i)} &\sim N(0, \sigma_0^2) \\ \tau_{s(i)} &\sim N(0, \sigma_1^2) \\ \tau_{c(i)} &\sim N(0, \sigma_2^2) \end{aligned} \qquad \text{all random effects iid}$$

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

Model selected

Parameters were estimated by maximum likelihood and the likelihood was computed using the Laplace approximation.
(`glmer` in R was used)

The proportional hazards model below fit best — interaction between baseline hazard function and treatment did not improve fit.
(Test statistic: LLR < 6; null distribution asymptotically χ^2_6)

$$\text{logits}[h(t; \mathrm{x}_i)] = \eta_{it} = \alpha_t + \beta_0 \cdot \mathit{treat}(\mathrm{x}_i) + \tau_{d(i)} + \tau_{s(i)} + \tau_{c(i)}$$

Class size, gender and nonwhite did not substantially improve the fit, so they were left out of the model.
(Test statistic: LLR ≤ 1.3 for each; null distribution asymptotically χ^2_1)

Parameter estimates and standard errors

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	

Conclusions

- **Discouragement (or boredom)**: Students were more likely (conditional odds 2:1 – 3:1) to *not* play any additional trials of the video game.

- **Die-hards**: Students who made it through the last additional trial were more likely (conditional odds 2:1 – 6:1) to continue playing the video game.

- **No — weak evidence for treatment**: There is not strong evidence in favor of a treatment effect — two-sided p-value of 6% for the estimate and a LLR of 3.1, which is at the 94th percentile of the asymptotic χ^2_1 null distribution.
  (Conditional odds of continuing to play the game increased by 13–57% for students in treatment group.)

Comparison to the study

Authors assumed the response was distributed as negative binomial (over-dispersed Poisson) and made the same random effect assumptions as I did. They only had a treatment factor in the systematic component of the model fitted.

The authors reported a treatment effect in the same direction as mine, and they were confident about the existence of an effect. They reported a two-sided p-value of 4%.

Authors suggest estimate for treatment effect is conservative, given randomization was done at school level. However, note that LLR tends to be anti-conservative.

Table to summarize

Estimates for treatment effect. Log conditional-odds for Brandon Rhodes (BR) and log rates-ratio for Marjorie Rhodes et al. (MR).

Author	Estimate	Standard Error	Conclusion
BR	-0.29	0.16	No — slight decrease in quitting
MR	0.36	0.18	Slight increase in persisting