Sesame Street Report [Draft]

Armelle, Sara, Ibrohim

2024-02-05

Table of contents

1 Project Description

In this project, we analyze the results of an observational study designed to assess the impact of *Sesame Street* viewership on children's learning outcomes; More specifically, we are interested in letters, numbers, and forms. We begin by exploring data collected from children in five different sites across the United States. Variables observed include, but are not limited to, viewing frequency, setting, encouragement to watch, and pretest scores of vocabulary maturity (Peabody Picture Vocabulary Test). We aim to determine the show's effectiveness in educational content delivery and identify areas for improvement. These results will be used to enhance *Sesame Street*'s educational focus and impact, as per the client's request for an upcoming board meeting presentation. We begin by posing the following research questions.

1.1 Research Questions

Question 1: Does our programming improve children's knowledge of letters, numbers, and forms?

Question 2: What, if any, area should we focus on for improvement? E.g. are we better at teaching letters than we are at numbers?

1.2 Variables

We considered a variety of variables in our preliminary analysis of the Sesame Street study. The dataset contained several explanatory variables, but we chose to focus on viewing frequency (Viewcat) as our main explanatory variable; and site, sex, age, setting, and encouragement as possible confounding variables. The dataset contained pre- and post-test scores for six different domains, but we were only interested in scores for letters, numbers, and forms. We also defined two possible response variables: percent increase and percent achievable gain (PAG). The first, percent increase, is simply the difference between the posttest score and pretest score, as percentages. The second, PAG, is the child's improvement (post-test score minus pre-test score) divided by total possible improvement (maximum test score minus pretest score). The goal with this measure is to capture improvement while accounting for the fact that advanced students cannot improve as much as those who were not as advanced to begin with. All variables used in our analysis are summarized in the table below. Although there were other variables included in the dataset, we decided that these were the most important ones to answer our client's research questions.

Table 1: Summary of variables used in analysis

Name	Type	Notes
ID	Numerical	Identifying numeric sequence
Site sampling sites (Explanatory)	Categorical	Five different
Sex (Explanatory)	Categorical	Male or Female
Age (Explanatory)	Numerical	Age in months
Viewcat show child watched (Explanatory)	Categorical	Categorical 1-4 encoding amount of
Setting (Explanatory)	Categorical	Home or School

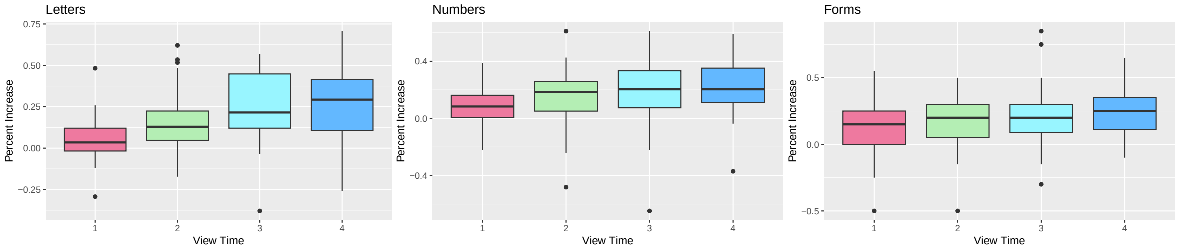
Table 1: Summary of variables used in analysis

Name	Type	Notes
Viewenc view show (Explanatory)	Categorical	Whether or not child was encouraged to
PAG (Response)	Numerical	Percent Acheivable Gain
Percent Increase pre-score percent (Response)	Numerical	The difference of post-score percent and

2 Exploratory Data Analysis (EDA)

Our exploratory data analysis of the Sesame Street study revealed promising leads for modeling, as well as concerns about possible data issues and confounders we will have to watch out for as we move forward. As a whole, it appeared that Sesame Street viewership was positively correlated with greater improvement, as measured by PAG and percent increase across all tests (letters, forms, and numbers). The comparative improvement across tests did not appear to have a clear trend measured by both PAG and percent increase.

While PAG is has nice theoretical properties, its use as a response variable in this study may not be practical; it introduced more nonuniform variability in the outcomes. In addition, exploratory analysis revealed that variance may not be equal across all groups (particularly with regards to site), and this could cause possible problems in modeling. More in-depth exploratory analysis is reported below. First, we look at visualizations to answer the two primary research questions; then, we investigate possible confounders.



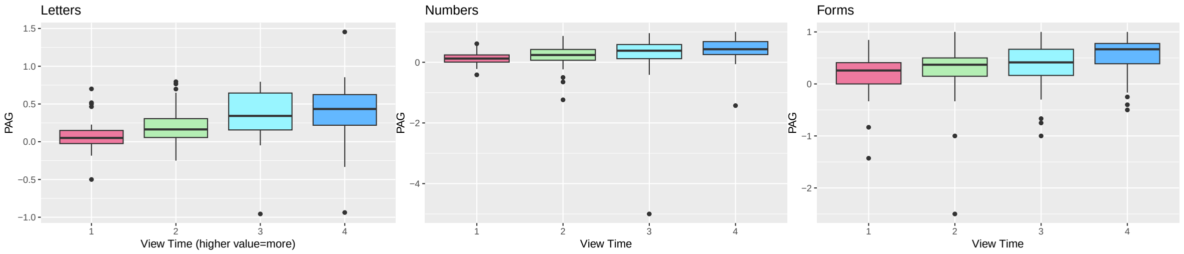
(a) Letters

(b) Numbers

(c) Forms

Figure 1: The Relationship Between Viewing Time of Sesame Street Percent Score Increase for Letters, Numbers, and Forms. Note that viewing time is categorical with higher values corresponding to more time viewing Sesame Street

All three plots in Figure 1 appear to indicate a positive linear relationship between viewing time (categorized as levels 1-4) and percent score increase. Although the variability is not completely consistent across viewing levels, overall the data are relatively well-behaved.



(a) Letters

(b) Numbers

(c) Forms

Figure 2: The Relationship Between Viewing Time of Sesame Street PAG for Letters, Numbers, and Forms. Note that viewing time is categorical with higher values corresponding to more time viewing Sesame Street

Similar to Figure 1, all three plots in Figure 2 seem to indicate a positive relationship between viewing time and percent score increase. However, by comparison to the plots using percent increase as the response variable, the data have a much wider spread with many more outliers. This could pose some problems in future modeling.

The plot in Figure 3 doesn't indicate a clear trend with regards to which subject Sesame Street teaches the most effectively. The highest median is forms, followed by numbers and then letters. However, all three boxplots have significant overlap with each other, which means that statistical analysis will likely not yield usable differences.

Once again, it appears that PAG introduces a lot of additional variance compared to the same plot using percent increase as the response variable. In this case, there is clearly significant left-skew in the data with lots of outliers. Despite this, the plot in Figure 4 does appear to show similar results as the plot in Figure 3. When put on the same scale, the values appear to be similar around the middle of the boxplot.

For simplicity, all investigation of possible confounders (below), was done with only percent increase as response variable.

The plots in Figure 5 have somewhat unclear results. Encouragement may have an impact on percent score increase, particularly for letters, but there is enough overlap of the boxplots in all three plots that more analysis is needed. This may be an important confounder to consider in modeling.

The plot in Figure 6 seems to show little to no relationship between sex and percent increase in score in any of the tests. It does not appear that sex will be an important confounder to consider in modeling.

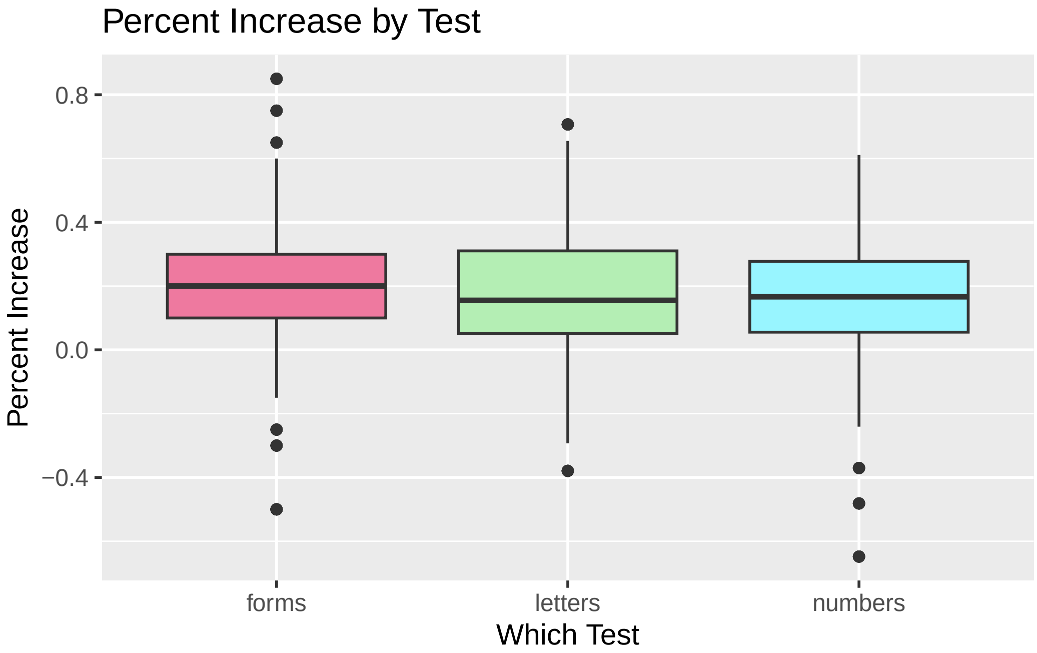


Figure 3: Comparing Percent Increase Across Three Subjects: Letters, Numbers, and Forms

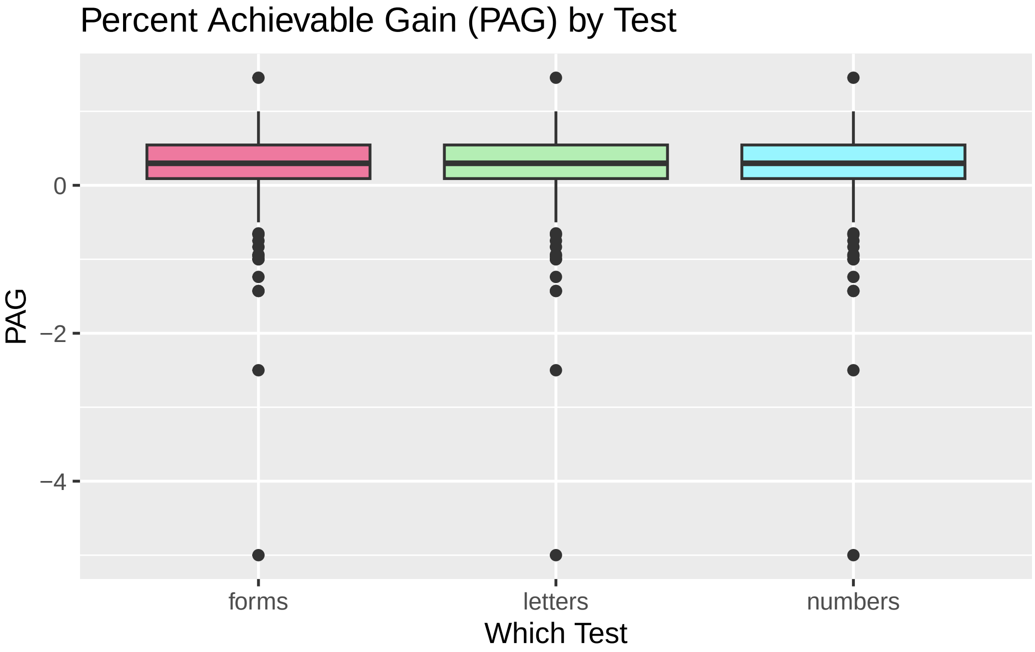
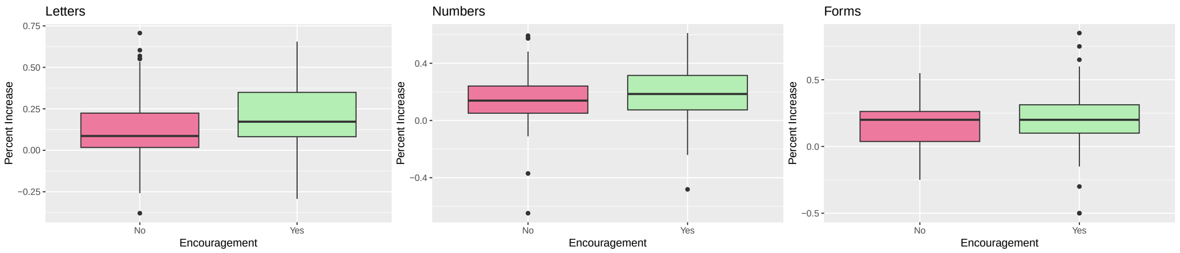


Figure 4: Comparing PAG Across Three Subjects: Letters, Numbers, and Forms



(a) Letters

(b) Numbers

(c) Forms

Figure 5: Comparing Percent Increase Across Three Subjects (Letters, Numbers, and Forms) depending on Encouragement

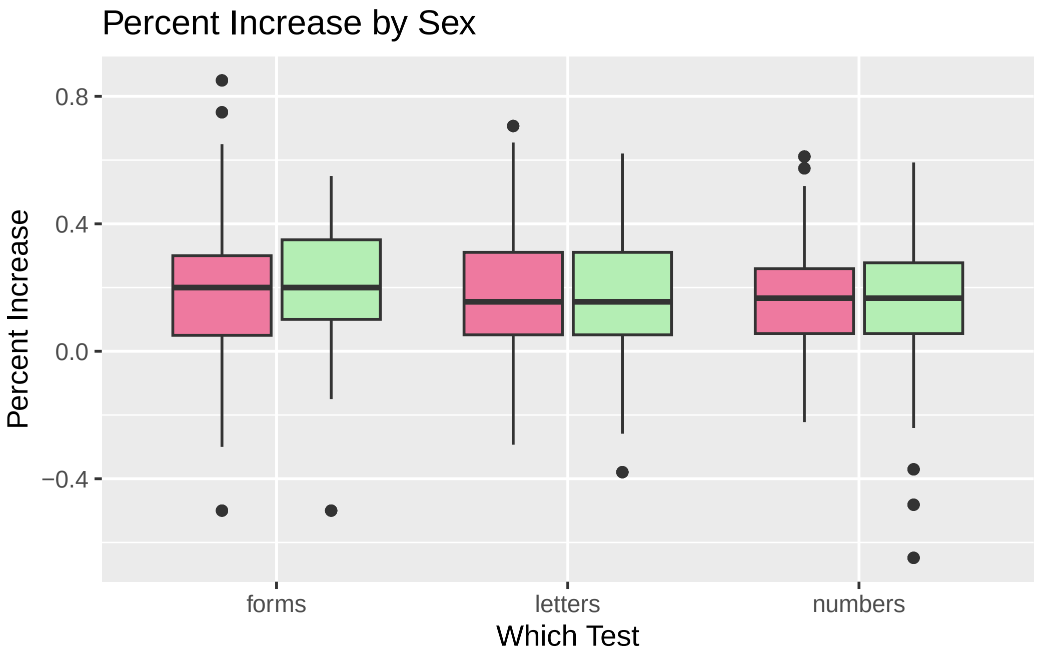


Figure 6: Comparing Percent Increase Across Three Subjects (Letters, Numbers, and Forms) depending on Sex

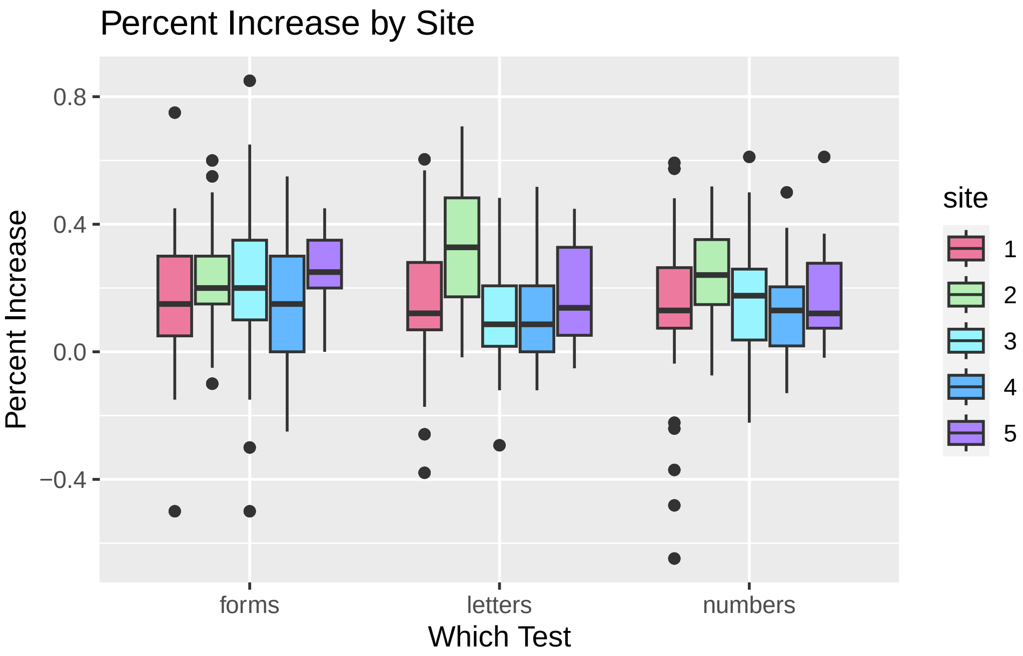


Figure 7: Comparing Percent Increase Across Three Subjects (Letters, Numbers, and Forms) depending on Site

The plot in Figure 7 shows some variation in score improvement across the different sites for all three tests. Not only is median percent increase in score different depending on the site, but the variability is not uniform across groups either. This variable seems to be an important confounder, and we will have to look out for issues with the non-uniform varibility.

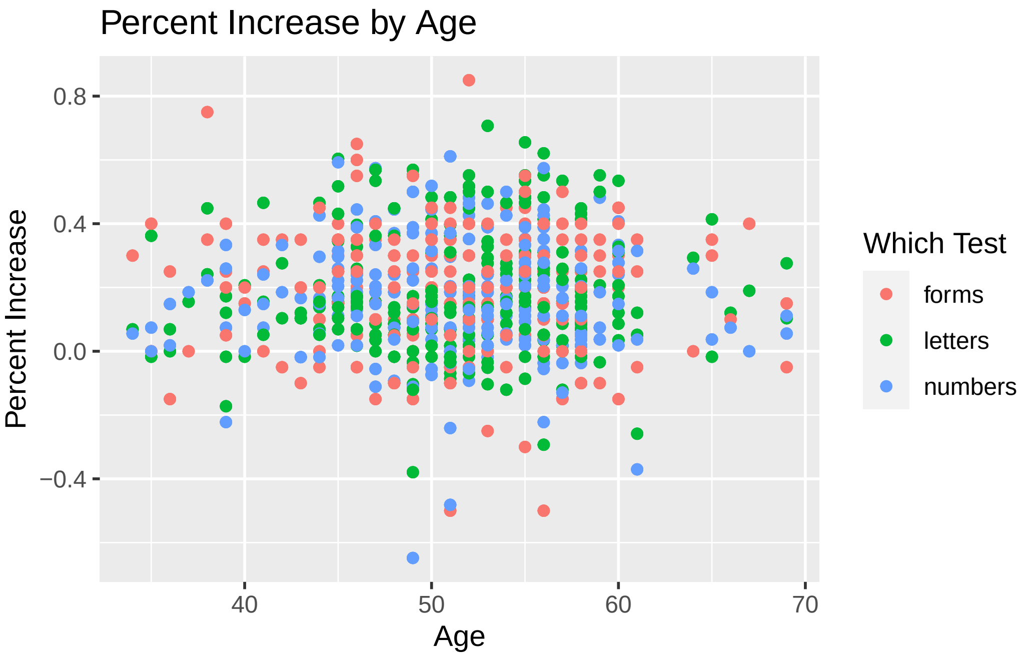


Figure 8: Comparing Percent Increase Across Three Subjects (Letters, Numbers, and Forms) depending on Age

The plot in Figure 8 seems to show little to no relationship between age and percent increase in score in any of the tests. The points appear to be scattered more or less at random. It does not appear that age will be an important confounder to consider in modeling.

The plot in Figure 9 seems to show little to no relationship between setting and percent increase in score in any of the tests. It does not appear that setting will be an important confounder to consider in modeling.

2.1 3. Statistical Analysis

2.1.1 Research Question 1: Setup

To address the first research question, we fit three multiple linear regression models: one for each subject. The models predicted improvement using prescore, age, viewcat, and site. The

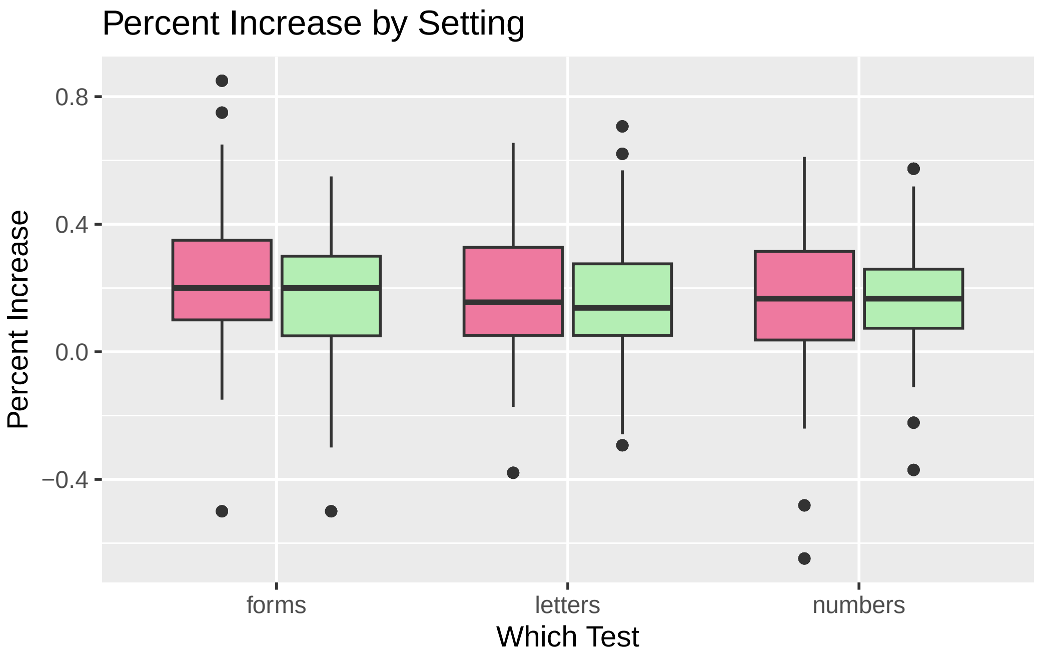


Figure 9: Comparing Percent Increase Across Three Subjects (Letters, Numbers, and Forms) depending on Setting

models were all of the form:

$$\text{SubjectImprovement} = \beta_0 + \beta_1\text{Prescore} + \beta_2\text{Age} + \beta_{3\ldots5}I_{Viewcat} + \beta_{6\ldots9}I_{Site} + \epsilon,$$

$$\epsilon \sim N(0, \sigma_\epsilon).$$

In this model, β_1 and β_2 are attached to the two quantitative variables, and they signify the change in predicted improvement given a unit increase in prescore and age, respectively. The other β's (3 through 9) are attached to the two categorical variables, viewcat and site. The interpretation of β_3 is the change in predicted improvement for viewcat 2 when compared to viewcat 1 (the baseline), and this interpretation extends through to β_5. Similarly, the interpretation of β_6 is the change in predicted improvement for site 2 when compared to site 1 (the baseline), and this interpretation extends through to β_9. Finally, β_0 is the intercept, and its interpretation is not important in this context. A (type II) ANOVA table for each model is included in Table 2, Table 3, and Table 4.

3 Statistical Analysis

3.1 Research Question 1

Table 2: ANOVA table for multiple linear regression model to predict improvement in letters using viewcat, prescore, age, and site

	Sum Sq	Df	F value	Pr(>F)
site	1.3563322	4	15.290852	0.0000000
age	0.1705559	1	7.691167	0.0060044
viewcat	1.4749363	3	22.170609	0.0000000
prelet	0.5921747	1	26.703946	0.0000005
Residuals	5.1003764	230	NA	NA

In this table the values that we are most interested in are the p-values on the right-hand column. Although some values appear to be zero, they are not, but are so small that they were rounded to zero. We will consider any p-value less than 0.05 to be "significant," meaning we can infer that what we observe is a product of the true process and not random chance. For the quantitative variables of age and prescore, the small p-values indicate that each of these variables is important for predicting improvement. For the categorical variables of viewcat and site, the small p-values indicate that at least one of the viewing categories and at least one of the sites resulted in a different improvement than another. The same interpretations can be applied to the two ANOVA tables below, which are analogous to this one but for the forms and numbers models.

Table 3: ANOVA table for multiple linear regression model to predict improvement in forms using viewcat, prescore, age, and site

	Sum Sq	Df	F value	Pr(>F)
site	0.3762414	4	4.141300	0.0029334
age	0.1730415	1	7.618692	0.0062423
viewcat	0.9975679	3	14.640345	0.0000000
preform	2.6819266	1	118.080169	0.0000000
Residuals	5.2239349	230	NA	NA

Table 4: ANOVA table for multiple linear regression model to predict improvement in numbers using viewcat, prescore, age, and site

	Sum Sq	Df	F value	Pr(>F)
site	0.4052706	4	3.943926	0.0040742
age	0.1522685	1	5.927256	0.0156704
viewcat	0.7486290	3	9.713796	0.0000046
prenumb	0.9852310	1	38.351439	0.0000000
Residuals	5.9085953	230	NA	NA

In Figure 10, we can observe the confidence intervals for the mean improvement in letters stratified by category after accounting for pretest score, age, and site. In this plot, if the red arrows for two confidence intervals don't overlap, than we can infer that those two categories are statistically different from each other. In this case, we can see that viewing category 4 was associated with greater improvements than categories 1 and 2, but similar amounts of improvement as category 3 given a particular site, pre-score, and age. Similarly, given a particular site, pre-score, and age, viewing category 3 was associated with greater improvements than categories 1 and 2 and viewing category 2 was associated with greater improvements than category 1.

In Figure 11, we can observe the confidence intervals for the mean improvement in forms stratified by category after accounting for pretest score, age, and site. This plot can be interpreted in the same way as the one for letters, and we can observe similar trends.

Likewise, in Figure 12, we can observe the confidence intervals for the mean improvement in numbers stratified by category after accounting for pretest score, age, and site. Once again, the same interpretations can be applied and the trend is similar to the other two.

Effect size is a metric which tells us whether the statistical differences we find are large enough to be practically important. We calculated effect sizes with respect to viewing category for each model, and found that in all models, differences between viewing category were associated with a moderate or large effect sizes.

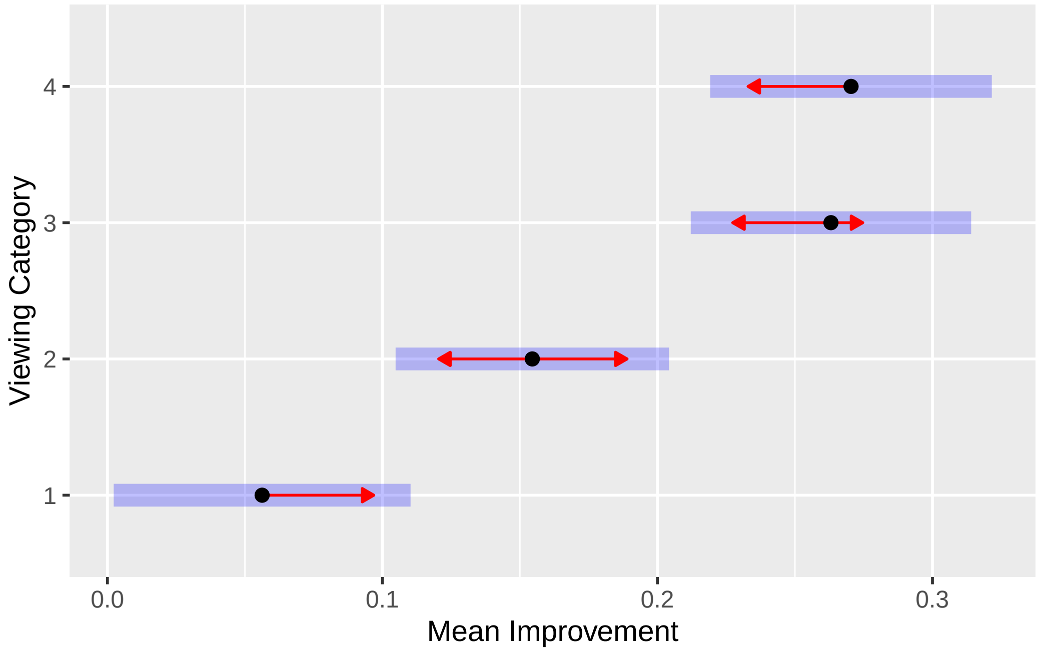


Figure 10: Mean Improvement in Letters by Viewing Category after accounting for site, age, and prescore

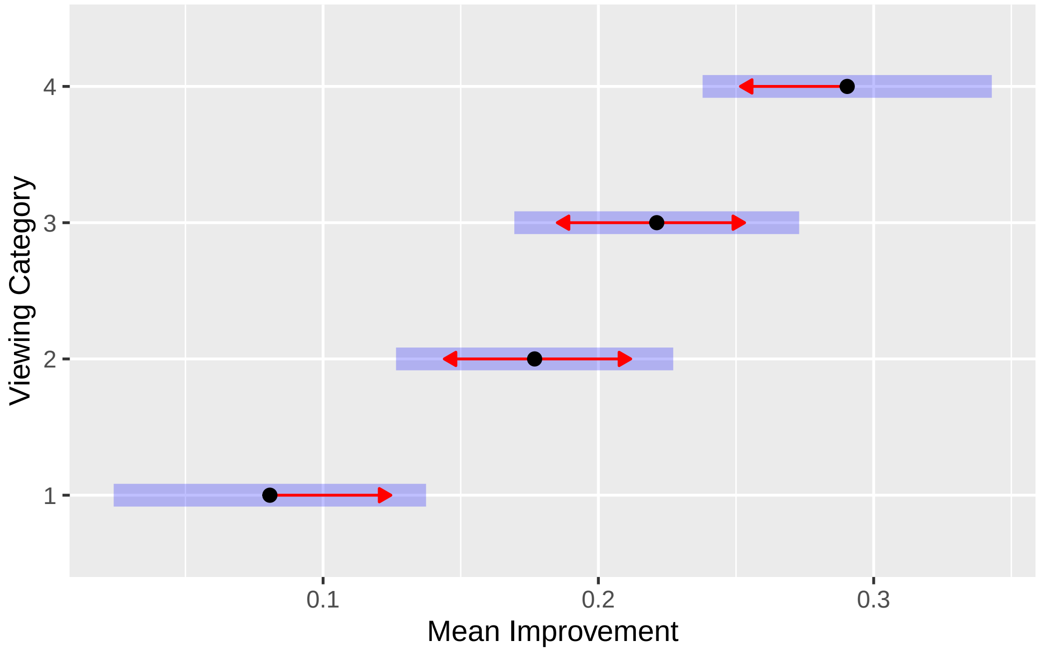


Figure 11: Mean Improvement in Forms by Viewing Category after accounting for site, age, and prescore

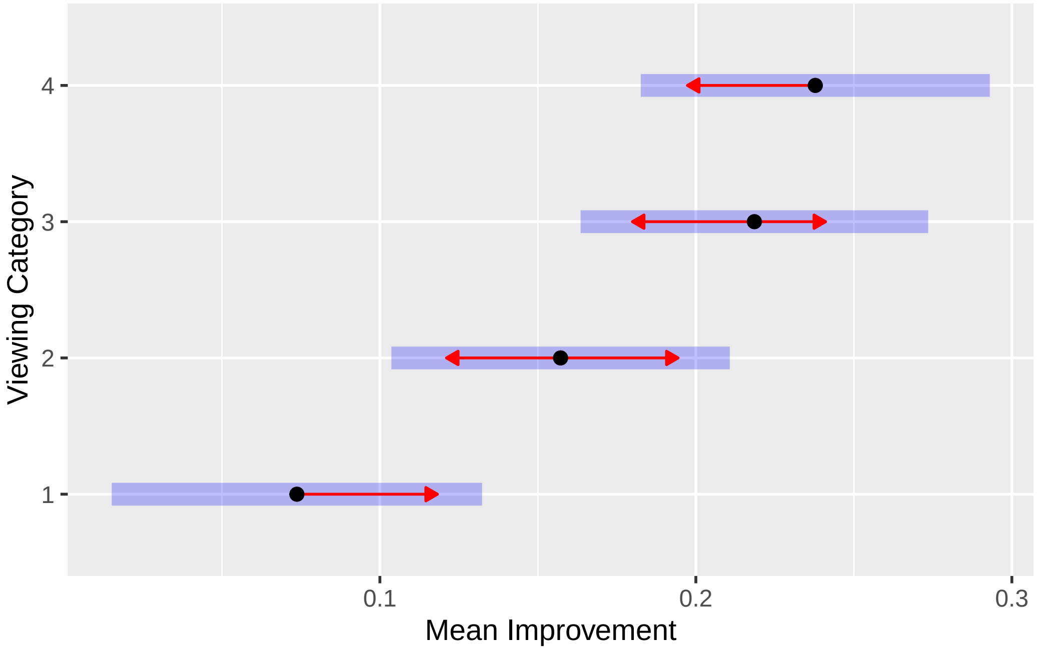


Figure 12: Mean Improvement in Numbers by Viewing Category after accounting for site, age, and prescore

3.2 Research Question 2

To address the second research question, we fit a mixed-effects linear model predicting improvement with subject, age, viewcat, site, and an interaction term between subject and viewcat while accounting for the random effects of the individual child (in the model, we call the random effect ID). The mathematical form of this model is:

$$\text{Improvement} = \beta_0+\beta_1\text{Age}+\beta_{2,3}I_{Subject}+\beta_{4\ldots6}I_{Viewcat}+\beta_{7\ldots10}I_{Site}+\beta_{11\ldots16}I_{Viewcat}I_{Site}+\gamma+\epsilon,$$

$$\epsilon \sim N(0,\sigma_\epsilon), \gamma \sim N(0,\sigma_\gamma)$$

The interpretations of β_0 to β_{10} work the same as they did in the previous model. However, in this model we also include an interaction term and a random effects term. The interaction term considers the effect of viewcat (amount of time watched) in the presence of another variable, which_test, denoting which subject is being tested. This way, we are considering if watching more sesame street has a unique benefit on a particular subject. The purpose of the random effects term is to account for the correlation between scores in different subjects for any particular child. That is, a child's score in letters is likely correlated with that child's score in forms and numbers. A (type II) ANOVA table for each model is included in Table 5.

The Analysis of Variance table provides the significance of each fixed effect and interaction term. Notably, the interaction term between which_test and viewcat is highly significant ($p < 2.2e\text{-}16$), indicating that the amount of time spent watching Sesame Street has a differential impact on the improvement scores across the different test subjects. The random effects component shows variability in the intercepts across individual children (ID), indicating that there is significant individual difference in improvement scores that is not explained by the fixed effects in the model. This is a good sign that a random effects model is necessary in this case. To further justify using random effects in our model, we conducted a likelihood ratio test comparing a model with and without random effects. The test resulted in a very low p-value ($p < 2.2e\text{-}16$), and so we confirmed that the random effects were necessary in this model.

Table 5: ANOVA table for mixed effects model to predict improvement in all subjects using viewcat, which_test (subject), age, and site

	Chisq	Df	Pr(>Chisq)
which_test	12.27101	2	0.0021646
viewcat	33.86941	3	0.0000002
site	14.86611	4	0.0049871
which_test:viewcat	101.10677	6	0.0000000

In Figure 13, we can observe the confidence intervals for the mean improvement stratified by which_test (subject) and viewcat after accounting for age and site. The reason that this plot is stratified by both variables is because our model has an interaction term, and so these are conditional means rather than marginal means. This plot helps us to identify which area has the lowest estimated means, signaling the most room for improvement. Visually, we see that this trend is not the same across all viewing cateogories. In viewcat 1, letters has the lowest mean improvement. In viewcat 2, the trend is not so clear since the red arrows for numbers and letters overlap with each other, so either subject could plausibly have the lowest mean improvement. In viewcat's 3 and 4, we see that either of numbers or forms have the lowest mean improvement. Since Sesame Street cares most about the effect of watching their show, the important result here is for viewing categories 3 and 4, since children who watched less of the show will presumably have been less impacted by it.

We calculated effect sizes using Pearson's r statistic for this model with respect to the differences between the subjects, and we found that the differences between subjects had moderate effect sizes.

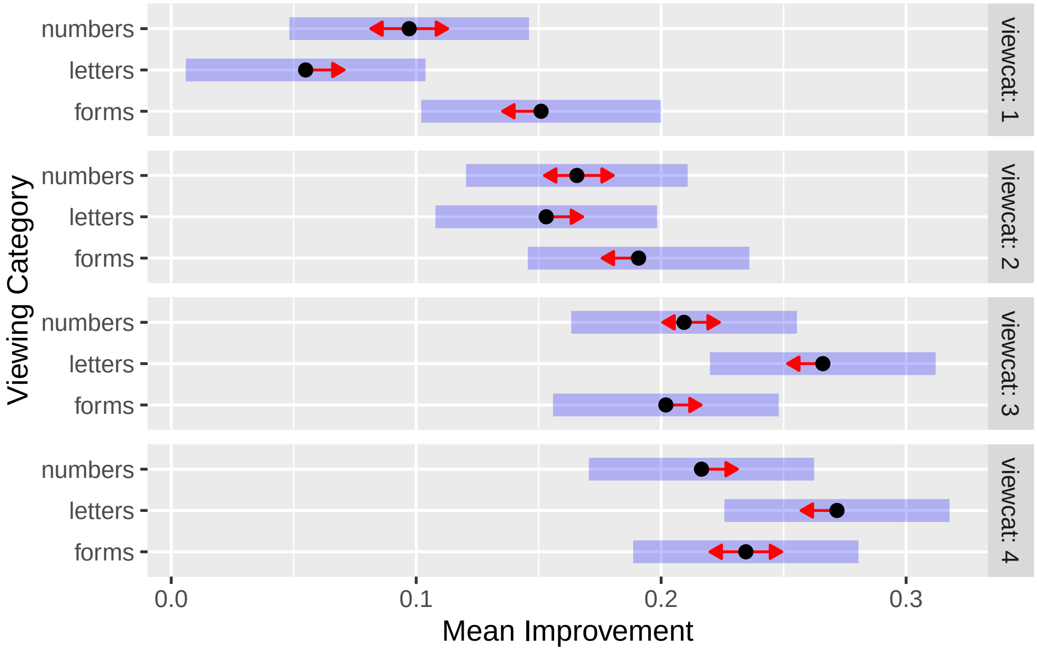


Figure 13: Improvement by Subject given viewcat after accounting for site and age

4 Recommendations

Question 1: Does our programming improve children's knowledge of letters, numbers, and forms?

More Sesame Street watch-time was associated with greater improvements in letters, numbers, and forms after accounting for differences in pre-score, age, and site.

Question 2: What, if any, area should we focus on for improvement? E.g. are we better at teaching letters than we are at numbers?

The relationship between subject depended on both site and viewing category. For children who watched more Sesame street (viewcat 3 or 4), their smallest improvements were in numbers and forms, while their greatest improvements were in letters, after accounting for age and site.

5 Appendix