# 

# 

# 

# **Analysing How Third Graders Develop Scientific Conceptions and Performance in the Evolving Minds Curriculum**

Client: Dr. Deborah Kelemen, Aarti Bodas

Team Members: Truc Minh Nguyen, Haochen Li, Chenran Zhang,Ziheng Li

**Table of Contents**

**1. Background and Objectives 2**

**2. Data collection 2**

**3. Experiment Procedure: 2**

**4. Objectives 3**

**5. Data Preprocessing: 3**

**6.Exploratory Data Analysis (EDA).......................................................................4**

**7. Appendix 27**

# 

**Background and Objectives**

The client is Aarti Bodas, a PhD student in the Department of Psychology and Brain Sciences at Boston University. The objective of this project is to investigate how third-grader students develop scientific understanding and performance within the Evolving Minds curriculum.

### **Data collection**

Data were collected from 459 third-grade students (8–9 years old) in eight public elementary schools in the Boston area. After applying inclusion criteria (completeness of survey/test responses), 387 students remain in the analysis. Each student was randomly or administratively assigned to either a 7-lesson or 12-lesson version of the Evolving Minds curriculum.

# **Experiment Procedure:**

### **Pre-Test**

1. Concept Inventory of Evolution Readiness (CIER):
   * Students individually completed the CIER test, a multiple-choice assessment focused on evolution concepts.
   * Each student’s pre-test performance is computed as: Number of correct answers/ Number of questions answered.
2. Mindset & Conceptions Survey:
   * Students answered a set of fixed mindset questions (e.g., “How smart you are is something you can’t change,” each on a 1–4 scale: strongly disagree → strongly agree).
   * A mindset composite (mindset\_comp\_pre) was formed by averaging these item ratings.
   * Students also rated nature of science conceptions on a 1–4 scale:
     + scievidence\_pre: “Science is looking for evidence…”
     + scitools\_pre: “Science is using scientific tools…”
     + sciexplain\_pre: “Science is explaining animals, objects, and environment…”
   * Science identity questions included items like “I am a person who enjoys doing science” (selfsciid\_pre).
3. Curriculum Assignment:
   * Students then participated in one of two versions of the Evolving Minds curriculum:
     + A 7-lesson adaptation. (Short version)
     + A 12-lesson “Speciation” version. (Long version)

### **Post-Test**

After completing the assigned curriculum (short vs. long), students repeated:

1. **CIER Test:**
   * Same multiple-choice instrument, yielding a post CIER score: Number of correct answers/ Number of questions answered.
   * In some analyses, this is treated as percentage\_post.
2. **Mindset & Conceptions Survey:**
   * The same 1–4 scale items for fixed mindset (yielding mindset\_comp\_post)
   * The same 1–4 scale items for nature of science conceptions (scievidence\_post, scitools\_post, sciexplain\_post)
   * The same science identity question (selfsciid\_post)

## **Objectives**

1. Do kids' pre-existing conceptions of the nature of science (e.g. science is using microscopes) positively or negatively predict their post-test performance on the longer and shorter version of a challenging science curriculum?
2. Whether kids' participation in a longer vs. shorter version of a challenging curriculum changes three general conceptions of science?
3. Whether pre-test science conceptions are stronger predictors of a final test grade than pre-test science identity or beliefs about the fixedness of intelligence (fixed mindset)

# **Data Preprocessing:**

Data preprocessing began by labeling and merging various datasets into a single file that contained both baseline (pre) and follow-up (post) measurements for each student. This process involved creating a binary time variable, in which “0” corresponded to the pre-test period and “1” corresponded to the post-test period. An additional indicator called “session(curriculum)” distinguished the two versions of the curriculum, with “0” indicating the shorter (7-lesson) Adaptation and “1” representing the longer (12-lesson) Speciation version. These steps ensured that all relevant survey items and test scores were properly aligned by student ID and session type.

Once the data were combined, the next step was reshaping from wide to long format. In the wide format, each row corresponded to a student’s overall pre-test and post-test data, whereas in the long format, each student appeared in two rows, one for the pre-test and one for the post-test. Reshaping was essential for a repeated-measures approach, allowing the model to treat time as a factor in order to capture how students’ conceptions and performance might evolve throughout the study.

Because the client wanted to capture different ways students approached the tests, they introduced two metrics: CIER\_raw\_score, which calculates the fraction of correct items out of six total, and “percentage,” which divides the number of correct items by the number answered, excluding any unattempted items. Both metrics were tested to examine robustness, since students might skip questions or respond inconsistently.

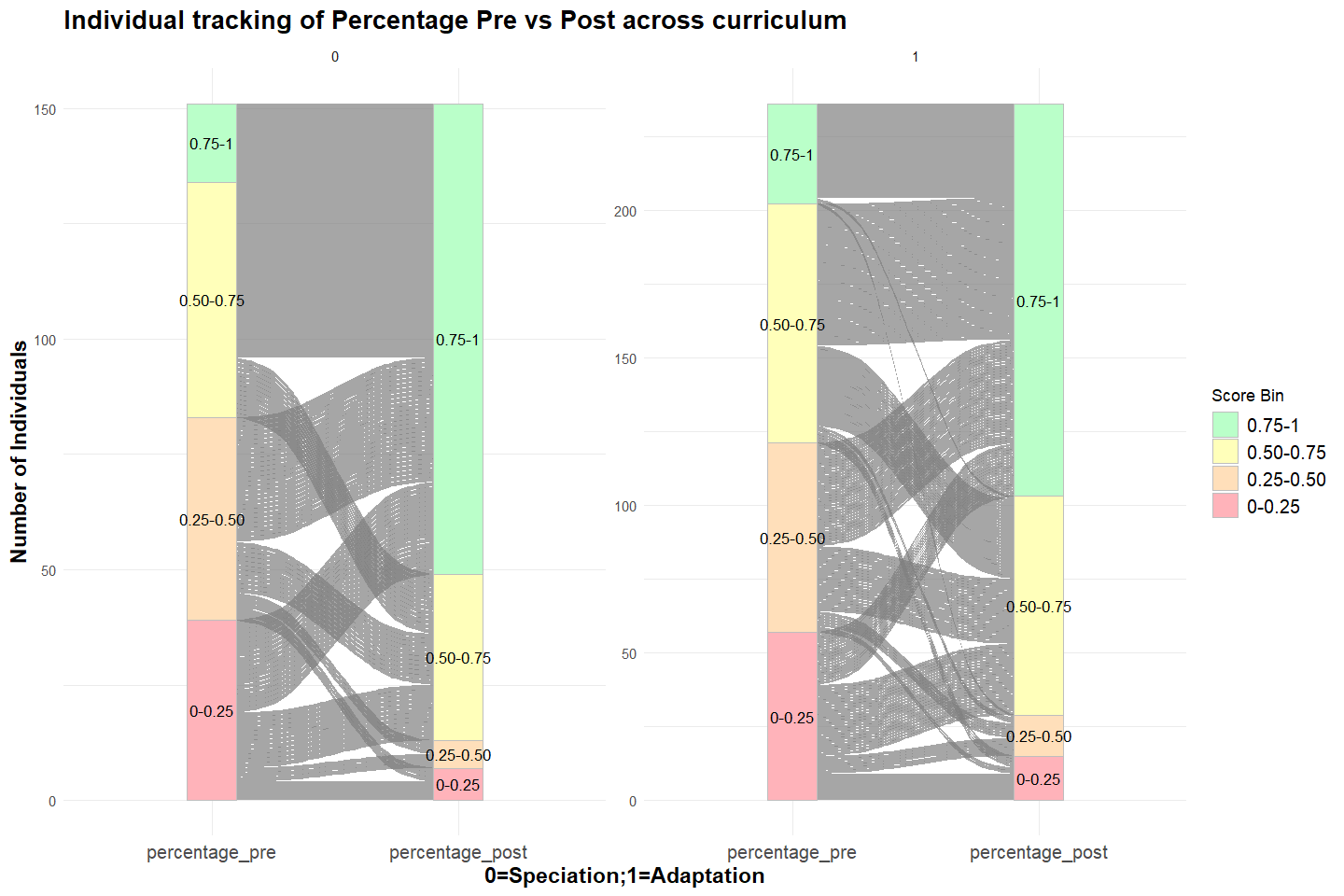
To ensure model robustness, we compared two approaches for predicting student performance: percentage-based scoring (percentage\_post) and raw percentage scoring (CIERrawpercent\_post). Both models included key predictors such as school, session, pre-test scores, and science-related factors, with the aim of determining whether percentage\_post is a valid substitute when CIERrawpercent\_post has missing data.

The coefficient comparison plot (Appendix 0.1, 0.2, 0.3) reveals that predictors produce similar effects across both models, indicating consistent results regardless of which scoring method is used. Pre-test scores (percentage\_pre and CIERrawpercent\_pre) emerge as the most significant predictors of post-test performance in both models, underscoring the importance of baseline knowledge. The results also show that school and session effects are stable across scoring methods: students from "SCHOOLUnknown (missing school information)" perform significantly worse in both models, and "SCHOOLSB" exhibits a negative impact. Meanwhile, session has a somewhat stronger negative effect under CIERrawpercent\_post, possibly reflecting test fatigue or instructional differences. Finally, science-related predictors and mindset variables show minimal influence in both models, suggesting that these factors contribute less to performance once other variables are accounted for.

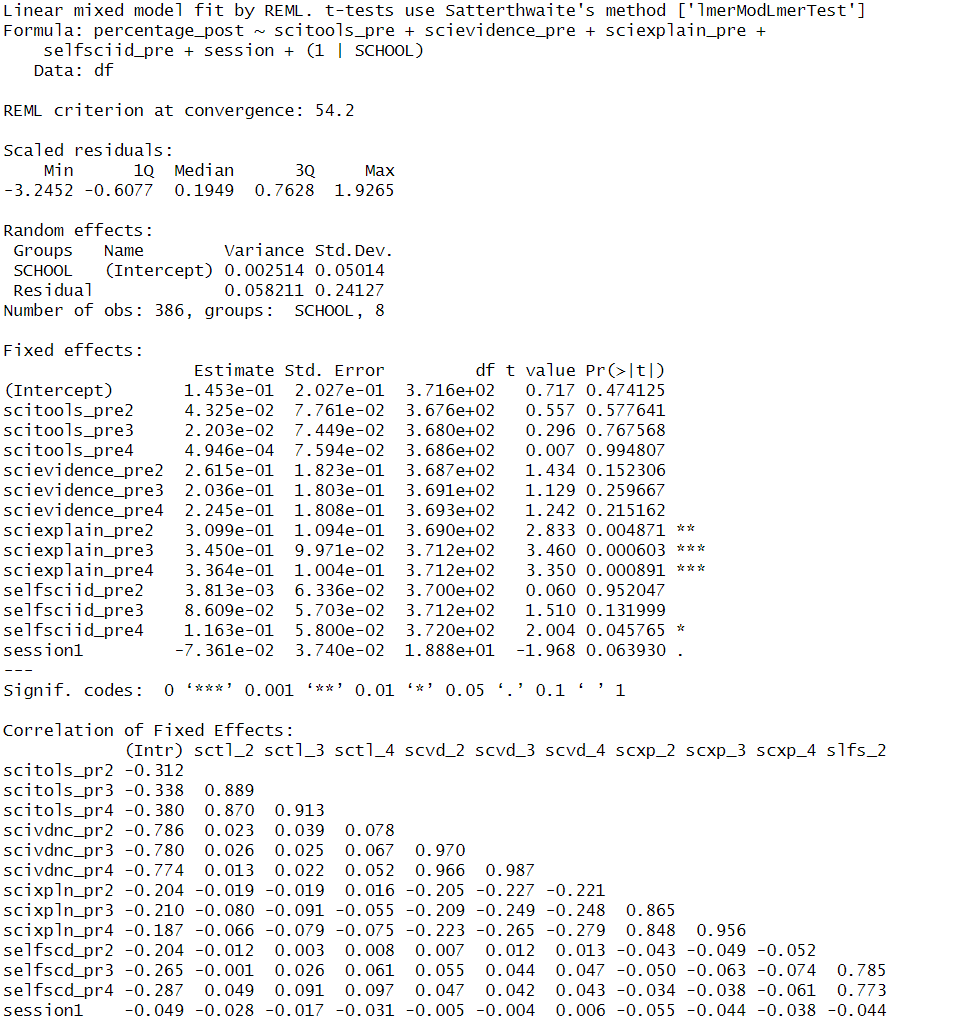
# **Exploratory Data Analysis (EDA):**

**Objective #1:** Do kids' pre-existing conceptions of the nature of science (e.g. science is using microscopes) positively or negatively predict their post-test performance on the longer and shorter version of a challenging science curriculum?

To answer this question, we decided to perform a linear mixed model, which treats scitools\_pre, scievidence\_pre, and sciexplain\_pre as categorical (unordered) factors, assesses whether children’s different conceptions of science predict performance (percentage\_post) on a challenging curriculum.



*Figure 1.1 The alluvial diagram tracking each student’s percentage-based pre- and post-test score distribution by session (speciation = 0, adaptation = 1), highlighting how individuals transition among four score bins (0–0.25, 0.25–0.50, 0.50–0.75, 0.75–1.0).*

****

*Table 1.1 Result of the linear mixed-effects model (LMM) fit by REML.*

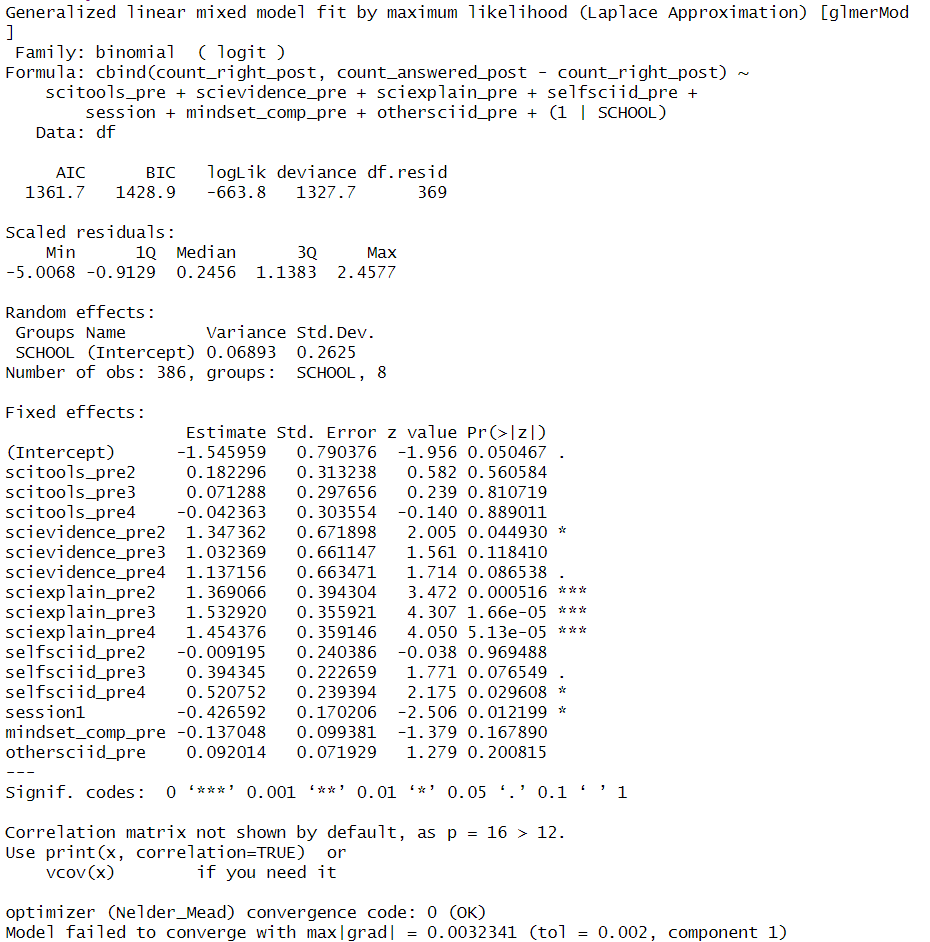
The result of REML (*Table 1.1*), with SCHOOL set as a random intercept, estimates indicate minimal between-school variation: the intercept variance is 0.002514 (SD = 0.05014), whereas the residual variance is 0.058211 (SD = 0.24127). These figures suggest that most outcome variability stems from individual differences rather than institution-level factors.

In the fixed effects, three main conceptions of science—tools-centric (scitools\_pre), evidence-centric (scievidence\_pre), and explanation-centric (sciexplain\_pre)—are represented as categorical factors with multiple levels. The coefficients for scitools\_pre and scievidence\_pre at their higher levels are small and generally not significant, implying that believing science is primarily about using scientific tools or gathering evidence does not clearly boost or lower post-test scores in this dataset. By contrast, sciexplain\_pre at levels 2 through 4 shows positive and significant coefficients (t-values of approximately 2.83 to 3.46), indicating that children who view science as an explanatory process tend to achieve higher percentage\_post outcomes. This finding aligns with the notion that emphasizing science as a means of explaining phenomena may foster deeper engagement and understanding, thereby improving performance.

Additionally, selfsciid\_pre4 exhibits a borderline significant positive effect (t ≈ 2.004), suggesting that students who identify strongly as “science people” may earn slightly higher post-test scores. The session (adaptation = 1) variable, with a t-value near −2.0, is marginally negative, pointing to potential performance decrements for kids’ participated in the adaptation curriculum.

To ensure the robustness of the model, we performed model estimation, diagnostic checks, including residual plots (Appendix 1.1) and simulation-based assessment via DHARMa (Appendix 1.4), were conducted to evaluate the adequacy of the linear assumptions. The Pearson residual versus fitted plot (Appendix 1.1) exhibits a downward trend at higher predicted values, and Q–Q plots (Appendix 1.2) reveal notable deviations from normality in the tails. The Kolmogorov–Smirnov (KS) test (Appendix 1.4) within DHARMa further confirms significant departures from the model’s simulated reference distribution. Although the dispersion test is not indicative of over- or under-dispersion (Appendix 1.4) , the observed clustering of scores near 1.0 (a possible ceiling effect) and the systematic residual patterns suggest that a normal-based approach may not fully capture the behavior of these proportion-based data.

Considering the property of the dependent variable which represents the proportion of correct responses out of total answered, a binomial mixed model was deemed more appropriate than a standard linear model. This approach directly models correct versus incorrect answers and handles bounded (0–1) data more naturally. The binomial mixed model uses 386 student records from 8 schools, treating each school as a random intercept to account for institution-level variation. Although a minor convergence warning arose (max|grad| ≈ 0.0032, tolerance = 0.002), the model converged sufficiently for interpretive purposes. The random-effect variance for SCHOOL (0.06893, SD = 0.2625) indicates modest differences among schools, but the predominant variability emerges from individual student factors rather than institutional effects.



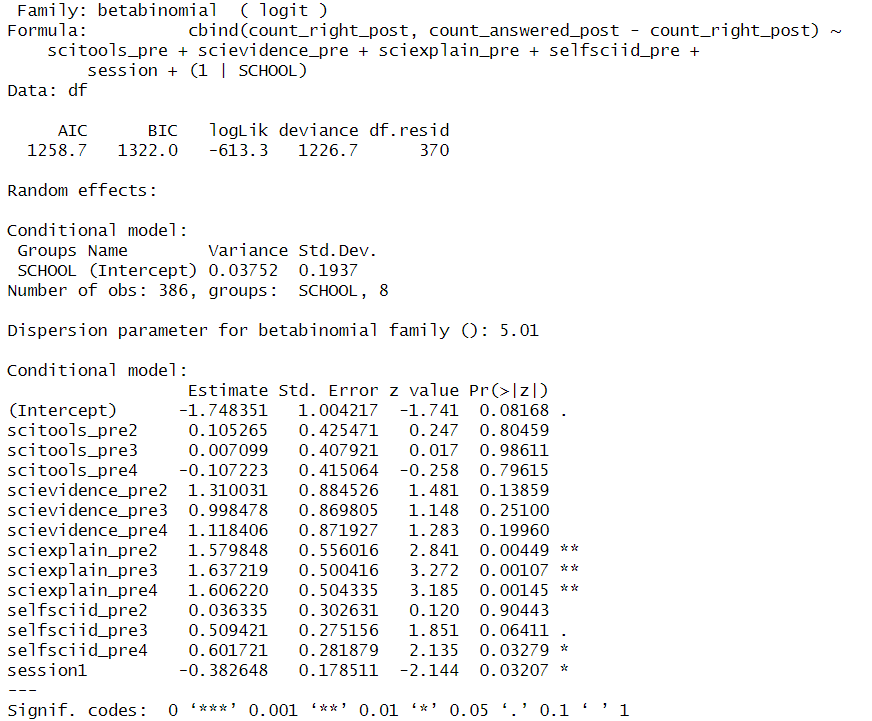
*Table 1.2 The result of a generalized linear mixed-effects model (GLMM) using a binomial logistic regression*.

From *Table 1.2,* regarding the fixed effects, the result indicates sciexplain\_pre is a positive predictor: levels 2 through 4 each significantly increase the probability of correct responses (Estimates ~1.37–1.53, p < 0.001). Likewise, a moderate association appears when science is framed as “gathering evidence” (scievidence\_pre2, Estimate = 1.3474, p = 0.0449), although higher levels show only marginal significance. By contrast, tools-centric views of science (scitools\_pre) fail to reach significance, implying minimal influence. A strong self-science identity (selfsciid\_pre4) exerts a small but significant effect (Estimate = 0.5208, p = 0.0296), while the session (versions of curriculum) variable negatively impacts outcomes (−0.4266, p = 0.0122), potentially reflecting kids’ participated in adaptation course show a lower performance in the post test.

In sum, children who conceptualize science as explaining phenomena show a higher likelihood of correct answers, and those who see science as partially about gathering evidence also benefit. Emphasizing tools alone has no notable impact. These findings underscore the value of fostering an explanatory perspective on science to enhance post-test performance.

DHARMa diagnostic plots (Appendix 1.8) reveal that the initial binomial mixed model with a single random intercept for SCHOOL fails to adequately reflect the observed outcome distribution. Notable deviations include a pronounced departure from the 1:1 line in the Q–Q plot (Appendix 1.8) (Kolmogorov–Smirnov p = 7e-05), a dispersion test indicating significant over- or under-dispersion (p = 0), and an outlier test that also flags irregularities. These results suggest that a standard binomial framework may be insufficient for modeling correct–incorrect response data.

To address these issues, we decide to use glmmTMB, a more flexible platform for generalized linear mixed models, where a beta-binomial approach is employed to handle potential overdispersion and accommodate outcomes bounded between 0 and 1.The resulting AIC (1258.7) and log likelihood (−613.3) indicate a reasonable fit, while the random-effect variance (0.03752, SD = 0.1937) shows moderate variability among schools. By allowing additional complexity, this beta-binomial method more naturally aligns with the data’s structure, thereby offering a more reliable framework.



*Table 1.3 The result of beta-binomial generalized linear mixed-effects model (GLMM).*

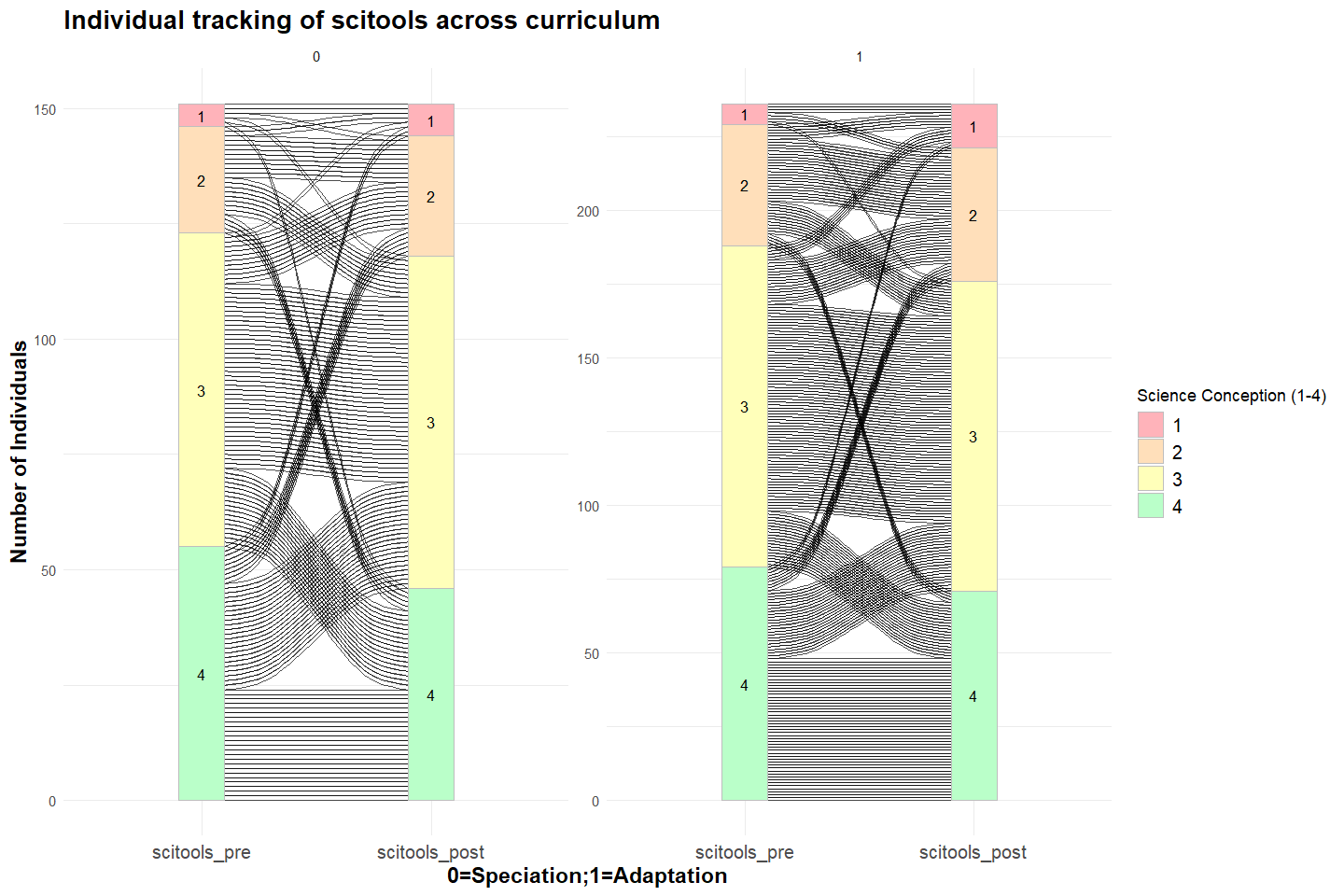
The result (*Table 1.3*) reveals that sciexplain\_pre (a view of science as an explanatory process) emerges as the strongest positive predictor: levels 2, 3, and 4 each exhibit highly significant coefficients (p < 0.01), indicating that students who conceptualize science as explanation tend to achieve higher proportions of correct answers. In contrast, emphasizing tools (scitools\_pre) or evidence (scievidence\_pre) does not consistently elevate performance, implying that these conceptions alone do not drive post-test outcomes in this sample.

Additionally, a strong self-science identity (selfsciid\_pre4) imparts a modest but positive effect (p = 0.0328), suggesting that students who regard themselves as “science people” benefit accordingly. Conversely, the session variable exerts a negative influence (Estimate = −0.3826, p = 0.0321). The beta-binomial dispersion parameter (5.01) and DHARMa diagnostics (dispersion test p = 0.736(Appendix 1.10)) confirm that this model effectively addresses overdispersion, leaving no clear evidence of under- or over-dispersion.

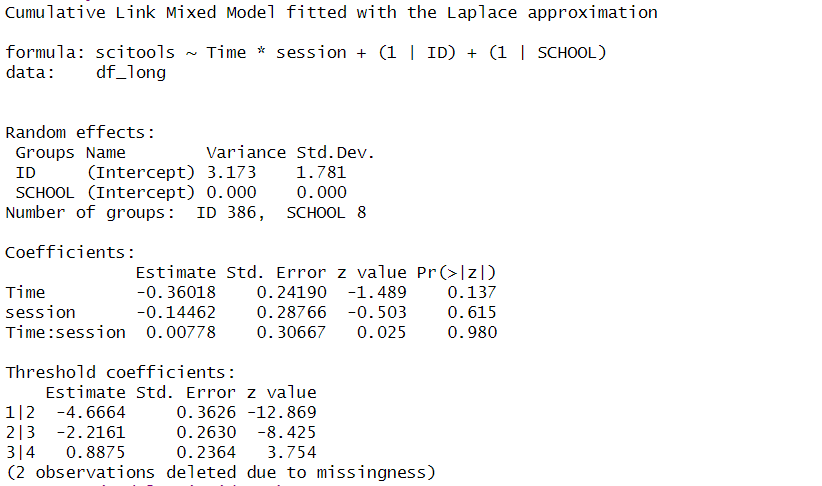
In summary, a beta-binomial approach successfully accommodates the overdispersed correct–incorrect data and demonstrates that viewing science as explanation—together with a strong personal science identity—correlates with improved performance, whereas tool- or evidence-centric perceptions do not yield significant advantages.

**Objective #2:** Whether kids' participation in a longer vs. shorter version of a challenging curriculum changes three general conceptions of science?

To answer this question, we reshaped the dataset into a “long” format: each student’s conception was recorded both at pre (time = 0) and post (time = 1), allowing us to track changes over time for each participant. We then fit a cumulative link mixed model (CLMM), which is appropriate for ordinal outcomes and includes random intercepts for both individual students (ID) and schools (SCHOOL), ensuring that variation at these levels is properly accounted for.

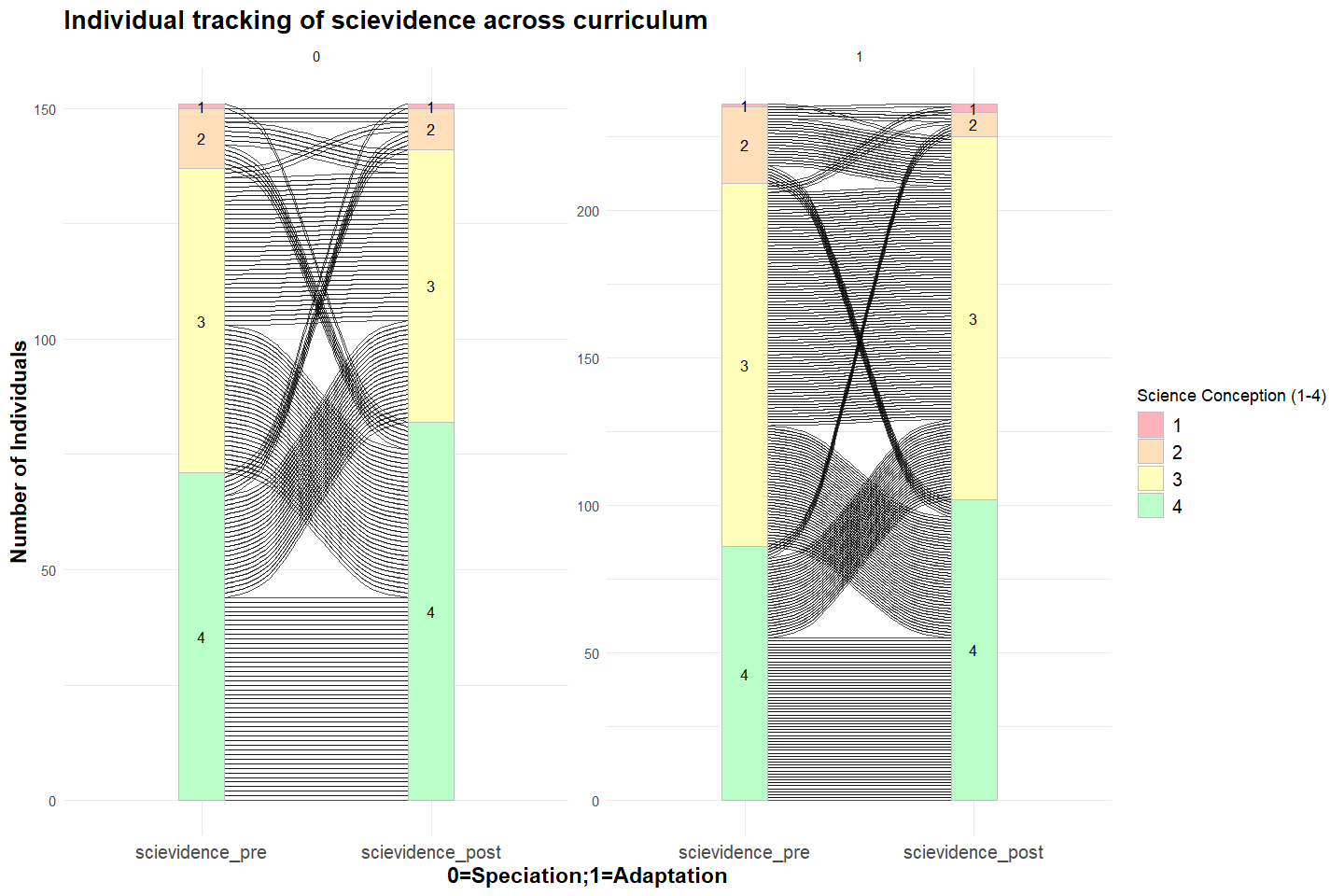


*Figure 2.1 The alluvial diagram tracking individual transitions in science tool conception (scitools) across the curriculum, comparing pre- and post-test distributions for two conditions (speciation = 0, adaptation = 1) and highlighting shifts among four science conception levels (1–4).*

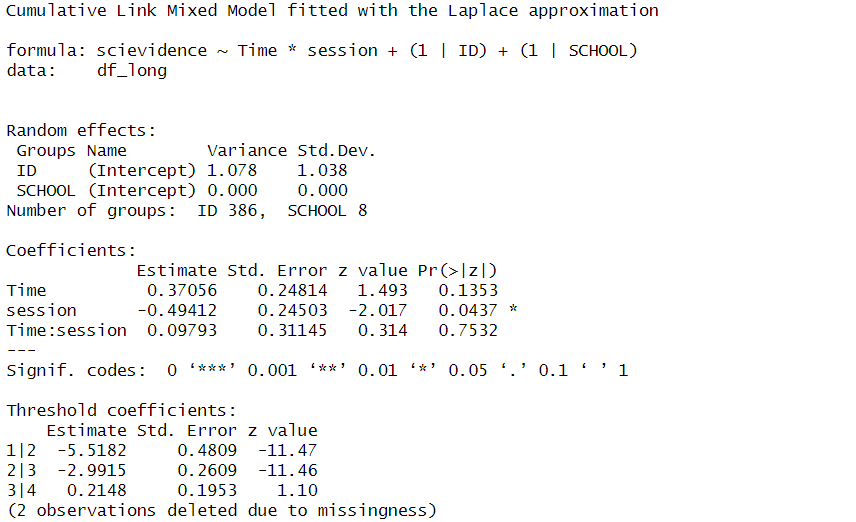


*Table 2.1 Result of a cumulative link mixed model (CLMM) analyzing the effects of time, session, and their interaction on “science is using tool conception” (scitools), incorporating individual and school-level random effects*

The result from *Table 2.1* of fitted CLMM for **scitools** indicates non-significant effects of time (Estimate = −0.36018, p = 0.137) and session(adaptation = 1) (Estimate = −0.14462, p = 0.615), as well as their interaction (Estimate = 0.00778, p = 0.980). This suggests that neither participating in a longer versus shorter curriculum produces a noticeable shift in students’ tool-centric view of science from pre to post. The threshold coefficients (−4.6664, −2.2161, 0.8875) define the boundaries among the ordinal categories, but none of the fixed effects substantively alter how children rank their “tools” conception over time. Overall, this model does not detect a significant change in scitools as children progress from before to after the challenging curriculum.

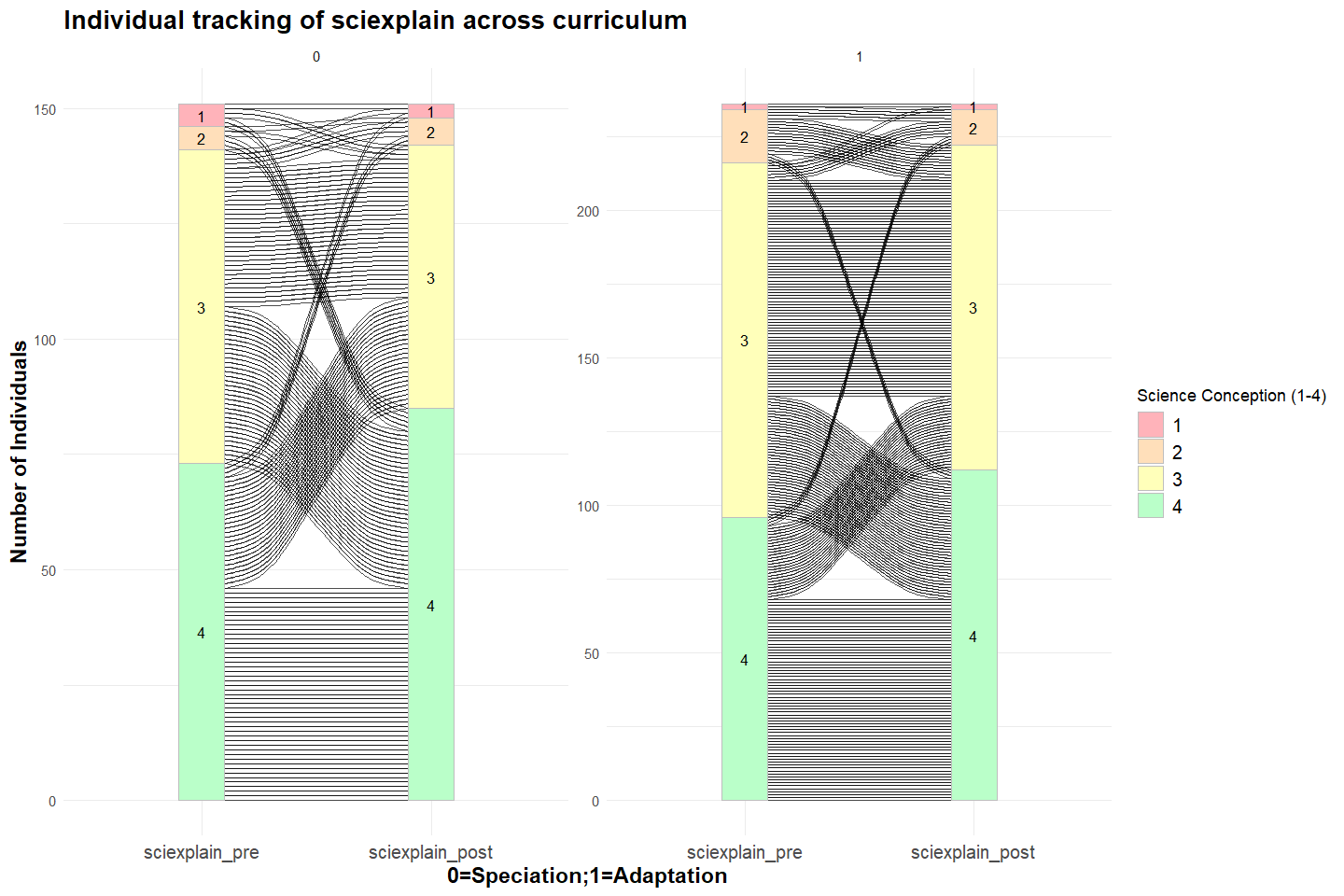


*Figure 2.2 The alluvial diagram tracking individual transitions in scientific evidence conception (scievidence) across the curriculum, comparing pre- and post-test distributions for two conditions (speciation = 0, adaptation = 1) and highlighting shifts among four science conception levels (1–4).*

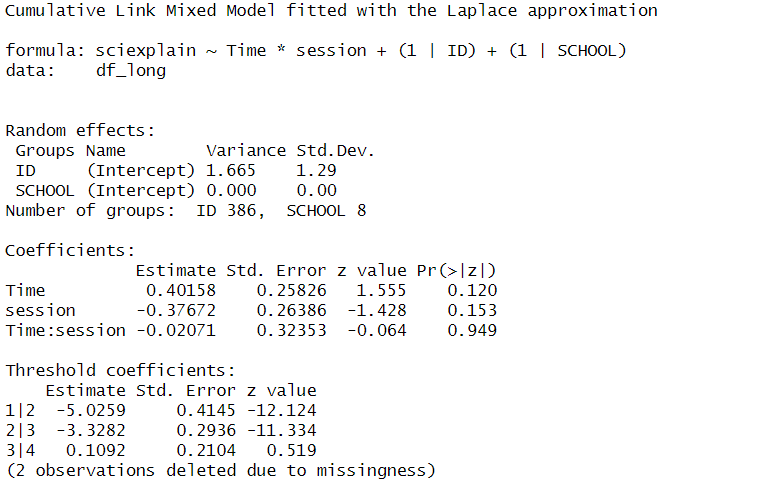


*Table 2.2 Result of a cumulative link mixed model (CLMM) analyzing the effects of time, session, and their interaction on “science is evidence-focused” (scievidence), incorporating individual and school-level random effects*

The result from *Table 2.2* for **scievidence** shows time has a positive but non-significant coefficient (Estimate = 0.37056, p = 0.1353), indicating that, overall, children’s evidence‐focused view of science does not shift substantially from pre to post. By contrast, session (adaptation = 1) produces a negative effect (Estimate = −0.49412, p = 0.0437), suggesting that students in the adaptation session are marginally less inclined to view science as evidence‐based than those in the speciation session, irrespective of time. The interaction between time and session (p = 0.7532) is not significant, which implies that adaptation versus speciation does not alter how children’s conception changes from pre to post.



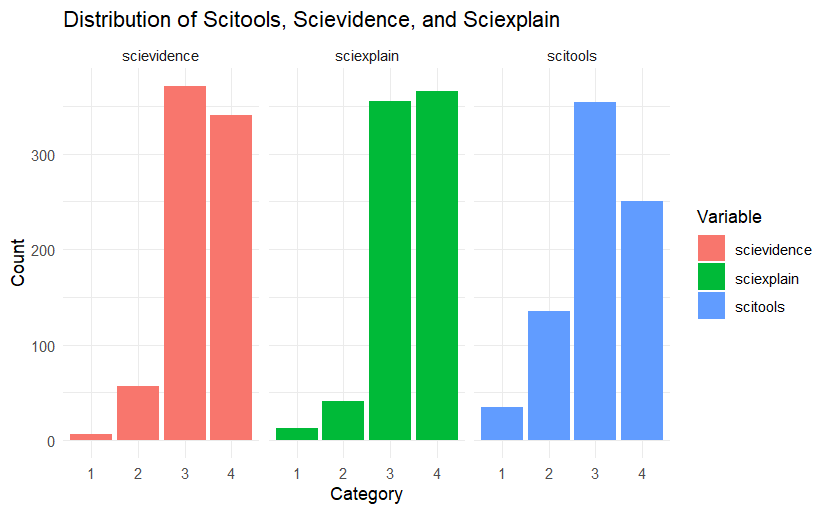
*Figure 2.3 The alluvial diagram tracking individual transitions in scientific explaining conception (sciexplain) across the curriculum, comparing pre- and post-test distributions for two conditions (speciation = 0, adaptation = 1) and highlighting shifts among four science conception levels (1–4).*



*Table 2.3 Result of a cumulative link mixed model (CLMM) analyzing the effects of time, session, and their interaction on “view science as explanation” (sciexplain), incorporating individual and school-level random effects*

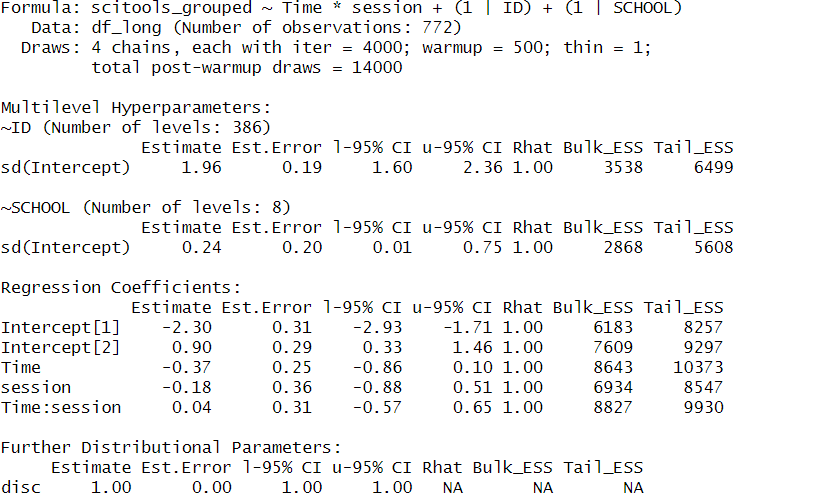
The result from *Table 2.3* for **sciexplain** shows time (Estimate = 0.40158, p = 0.120), session (adaptation = 1, Estimate = −0.37672, p = 0.153), and their interaction (Estimate = −0.02071, p = 0.949) are not statistically significant, indicating that none of these factors robustly affects children’s inclination to view science as explanation. The random-effects structure reveals negligible variance among the eight schools, while the majority of unexplained variability (variance = 1.665, SD = 1.29) lies at the student level. Overall, these results suggest that children’s explanatory conception of science remains stable from before to after the curriculum, with no detected difference between the adaptation and speciation sessions.

Across these three conceptions of science, the CLMM analyses do not reveal a consistent, significant change tied to moving from pre to post, nor to a longer (speciation) versus shorter (adaptation) version of the curriculum. With the exception of a modest session effect on scievidence, children’s views regarding science as tools, evidence, or explanation appear stable in this dataset.



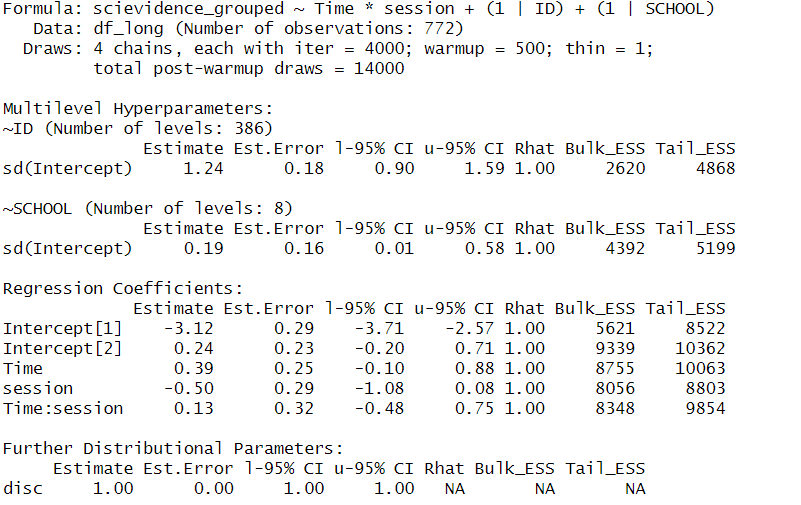
*Figure 2.1 The bar plot of the distribution of student responses across three scientific conceptions: Scitools, Scievidence, and Sciexplain.*

Due to the imbalance(Figure 2.1) in the dataset and the violation of the Proportional Odds Assumption (POA) for regular ordinal regression (CLMM)(Appendix 2.0.1)—which assumes that the effect of predictors remains constant across response categories—we opted to use Bayesian ordinal regression (brms) to refit the model. To further mitigate the effects of data imbalance, we grouped students who scored 1 or 2 on the three conceptions measures into a single category (1\_2), reducing the impact of sparsely populated response categories. Additionally, we compared models across the three conceptions for both grouped and ungrouped data. To ensure robustness and improve interpretability, we ultimately focused on the grouped data in our analysis.



*Table 2.4 The Bayesian ordinal regression model for Scitools Grouped was fitted to assess the impact of Time (pre/post-test) and Session (longer vs. shorter curriculum) while accounting for individual and school-level variations.*

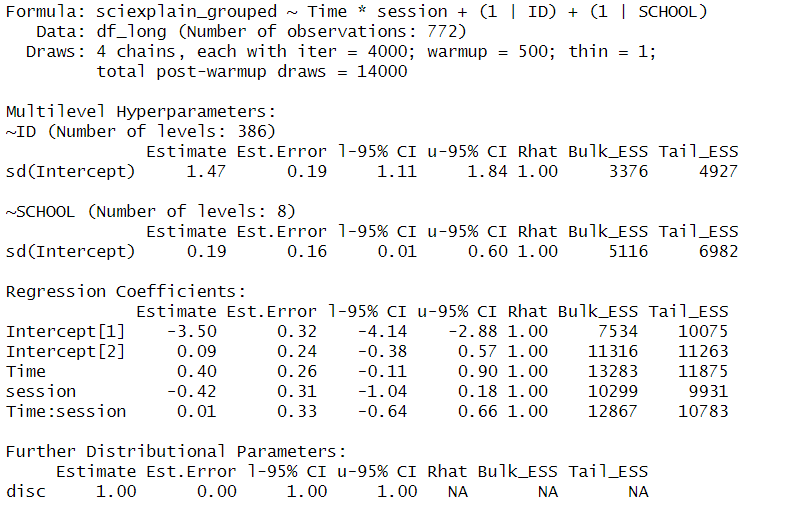
The result from Table 2.4 indicates that Time coefficient (-0.37, 95% CI: -0.86 – 0.10) indicates a possible decline in **Scitools** over time, but the confidence interval includes zero, meaning this effect is not statistically significant, which also supported by the plot (Appendix) , the line is almost parallel to the x-axis. Similarly, the Session effect (-0.18, 95% CI: -0.88 – 0.51) suggests that students in the longer (Speciation) curriculum did not significantly differ from those in the shorter (Adaptation) version. The interaction (0.04, 95% CI: -0.57 – 0.65) shows no meaningful difference in how students’ tool-based conceptions evolved across curriculum versions. With good model convergence (Rhat = 1.00, sufficient ESS) and a constant variance assumption (disc = 1.00), the findings indicate that curriculum length does not substantially shift students’ perception of science as a tool over time.



*Table 2.5 The Bayesian ordinal regression model for Scievidence Grouped was fitted to assess the impact of Time (pre/post-test) and Session (longer vs. shorter curriculum) while accounting for individual and school-level variations.*

The result from Table 2.5 indicates that Time coefficient (0.39, 95% CI: -0.10 – 0.88) suggests a potential positive shift in **Scievidence** over time, but the wide confidence interval crossing zero implies uncertainty in this effect. The Session coefficient (-0.50, 95% CI: -1.08 – 0.08) suggests that students in the longer curriculum may have slightly lower evidence-based conceptions, though this effect is also not conclusive. The interaction term (0.13, 95% CI: -0.48 – 0.75) shows no clear difference in how students' evidence-based views evolved depending on curriculum length. Model diagnostics show good convergence (Rhat = 1.00, adequate ESS), and the proportional odds assumption is met (disc = 1.00), reinforcing that curriculum version does not significantly alter students’ tendency to view science as evidence over time.

sciexplain



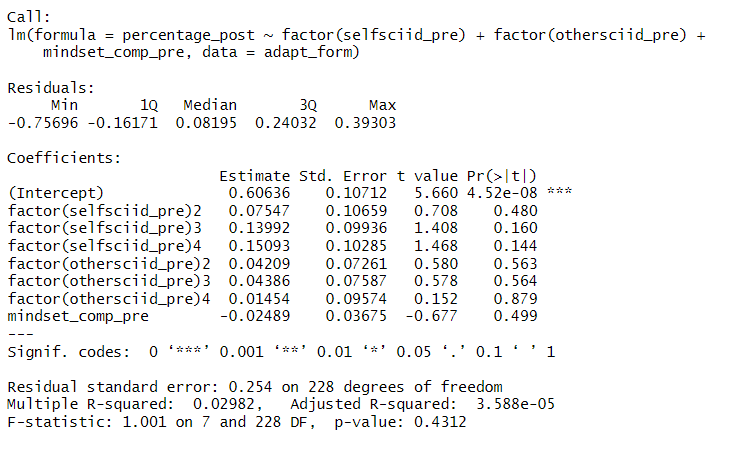
*Table 2.6 The Bayesian ordinal regression model for Sciexplain Grouped was fitted to assess the impact of Time (pre/post-test) and Session (longer vs. shorter curriculum) while accounting for individual and school-level variations.*

The result from Table 2.6 indicates that the Time effect (0.39, 95% CI: -0.10 – 0.88) suggests a possible increase in **Sciexplain** from pre- to post-test, but the confidence interval includes zero, meaning the effect is not statistically significant. Similarly, the Session effect (-0.50, 95% CI: -1.08 – 0.08) suggests no clear advantage of the longer curriculum. The interaction (0.13, 95% CI: -0.48 – 0.75) confirms that both versions of the curriculum led to similar learning trends. With good model convergence (Rhat = 1.00, sufficient ESS) and a constant variance assumption (disc = 1.00), the findings suggest that curriculum length does not significantly shape students’ explanatory reasoning in science.

In summary, the Bayesian ordinal regression results indicate that curriculum length does not significantly impact students' three conceptions of science, as none of the Time, Session, or Interaction effects showed strong statistical significance across the three models (*Tables 2.4–2.6*). The **non-NPO model** was chosen based on LOO and Bayes factor comparisons (Appendix 2.1.2, 2.2.2, 2.3.2), which favored the proportional odds assumption and reduced model complexity. Additionally, model diagnostics (Appendix) confirm good convergence (Rhat = 1.00, adequate ESS) and a good fit, supporting the robustness of these findings.

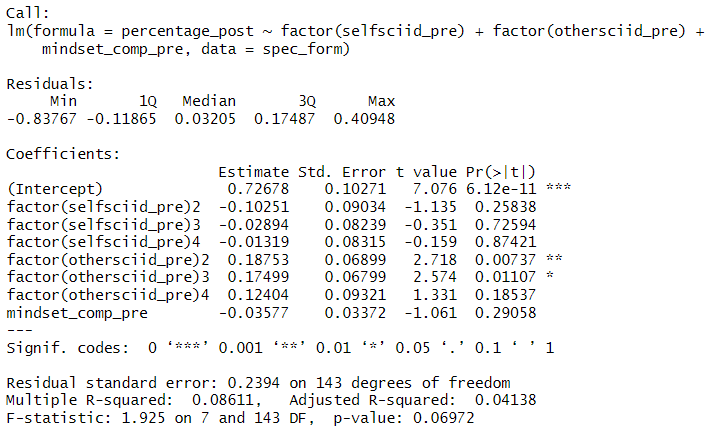
**Objective #3:** Whether pre-test science conceptions are stronger predictors of a final test grade than pre-test science identity or beliefs about the fixedness of intelligence (fixed mindset)

To answer the first part of the objective, the base model will include only the pre-test science identities and the pre-test fixed mindset composite scores as predictors. The result of the base model is shown below:



**Adaptation base model summary**

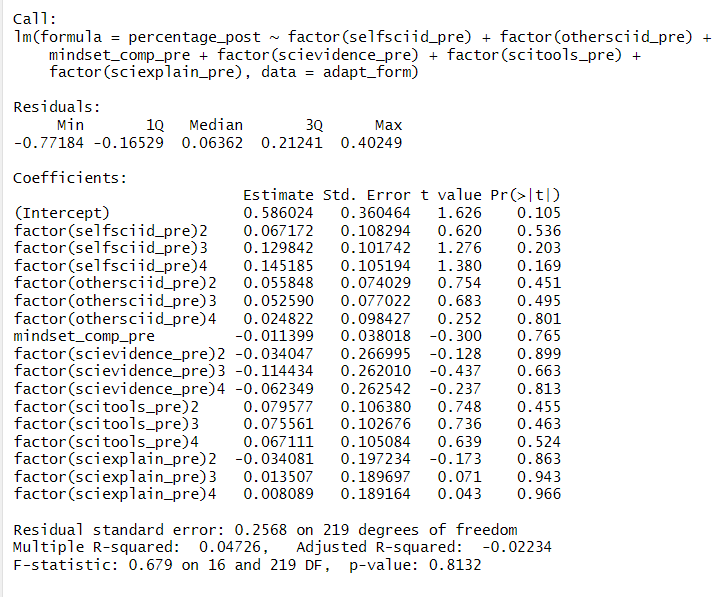
The summary for the adaptation base model shows that none of the predictors are significant, only the intercept. This means that the value when all categorical predictors are at their reference levels (level 1 for selfciid\_pre, othersciid\_pre) and when mindset\_comp\_pre = 0, the predicted percentage\_post score is 0.606 with a significant p-value of 4.52e-08. The estimates for the predictors have large standard errors and therefore, indicate uncertainty of drawing conclusions from the predictors.



**Speciation base model summary**

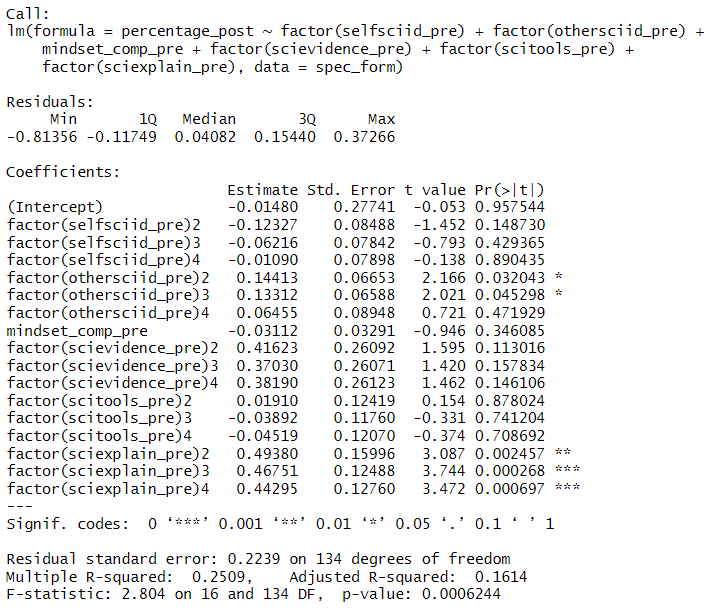
The summary for the speciation base model shows that two predictors are significant (others science identity rating of 2 and 3), and the intercept. For the intercept interpretation, this means that the value when all categorical predictors are at their reference levels (level 1 for selfciid\_pre, othersciid\_pre) and when mindset\_comp\_pre = 0, the predicted percentage\_post score is 0.726 with a significant p-value of 6.12e-11. This is slightly higher than the intercept of the adaptation curriculum of 0.606. The predictors of groups rating others’ science identity score 2 and 3 also show significance with low standard error compared to the estimates. The interpretation for othersciid\_pre2 means that keeping all other predictors constant, if a child rates a 2 of how others view his/her science identity, the percentage post score is predicted to increase by 0.188 (similarly for othersciid\_pre 3 with a 0.175 predicted increase in percentage post score).

To answer the second part of the objective, the full model will include the pre-test science identities, pre-test fixed mindset composite scores, and pre-test science conceptions as predictors. The result of the full model for adaption is shown below:



**Adaptation full model summary**

The summary for the adaption full model shows that none of the predictors or the intercept are significant (large standard errors for the estimates).

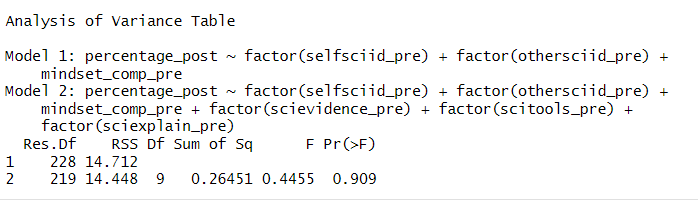


**Speciation full model summary**

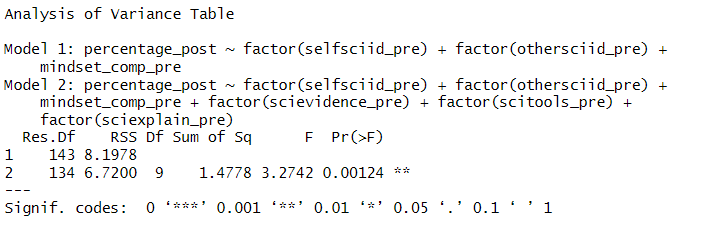
The summary for the speciation full model shows that the intercept is not significant (large standard errors for the estimates). However, there are some significant predictors such as othersciid\_pre2 and othersciid\_pre3 (similar to the speciation base model), along with sciexplain\_pre2, sciexplain\_pre3, and sciexplain\_pre4. The interpretation of the predictors is similar to the adaptation model.

The F-statistics is smaller on the full model vs. the base model of the adaptation curriculum (0.679 vs. 1.001), indicating that the null hypothesis should take precedence under both models.

An analysis of variance (ANOVA) table was calculated and the results (p-value of 0.909 on model 2) showed that adding additional predictors in the full model does not improve model fit and it is likely not a strong predictor of the outcome.



**Adaptation ANOVA table**



**Speciation ANOVA table**

The ANOVA table for the speciation curriculum has a higher F value for the full model (3.27) than the adaptation ANOVA. A significant p-value of 0.00124 on the full model indicates that new predictors significantly improve the full model compared to the base model and adds meaning to the interpretation.

# **Discussion & Limitation:**

Below is a list of potential biases from the dataset to take into account when drawing inferences from the models.

* **Selection Bias** – Schools had discretion in choosing which classrooms participated, introducing potential bias. The sample may not fully represent all third graders(N = 387). This limits the external validity of findings, as voluntary participation could lead to systematic differences between participating and non-participating schools.
* **Teacher Differences** – Teachers had autonomy in selecting whether to teach the shorter (Adaptation) or longer (Speciation) version of the curriculum. Since participation was voluntary, differences in teaching style, subject-matter expertise, and instructional delivery could introduce confounding effects on student learning outcomes.
* **Testing Differences** – Not all classrooms completed pre/post CIER test questions, which could introduce inconsistencies in measuring students' understanding. Additionally, test completion rates vary, as some students may have skipped or failed to complete all six CIER questions, potentially biasing results. CIER Scores were calculated as the number of correct responses divided by the number of questions answered. However, since the denominator varies due to missing information, direct comparisons across students may introduce noise into the analysis.
* **Final Sample Size** – The analysis only includes students who completed both pre- and post-tests, slightly reducing the dataset. This poses a specific limitation for Objective 2, which investigates whether participation in a longer vs. shorter curriculum influences the three general conceptions of science. The dataset is both imbalanced and relatively small, making it unsuitable for fitting a Non-Proportional Odds (NPO) model, which typically requires a larger sample size to estimate category-specific effects reliably.

These biases raise a potential concern about the generalizability of the results to the target population due to a limited sample size. The target population is elementary school children in the U.S. However, since the dataset contains only schools in the Boston area, it is important to note that we might only be able to draw conclusions about the sample and might not be able to generalize it to the population. The desired outcome is to be able to draw conclusions from the models to understand whether any of the three different conceptions of science bear at all on how kids perform on the post CIER and whether the performance differs between curricula. The results could potentially shape what gets targeted in science programs at the schools in the sample.

# 

# **Appendix:**

**Data preprocessing**

| **0.1** |  |
| --- | --- |
| **0.2** |  |
| **0.3** |  |

**Objective #1**

| **1.1** |  |
| --- | --- |
| **1.2** |  |
| **1.3** |  |
| **1.4** |  |
| **1.5** |  |
| **1.6** |  |
| **1.7**  **(Bion)** |  |
| **1.8** |  |
| **1.9** |  |
| **1.10**  **(Beta-binomial)** |  |
| **1.11** |  |
| **1.12** |  |

**Objective #2**

| **2.0.1 Likelihood Ratio Test (Comparing PO and NPO Models)** |  |
| --- | --- |
| **Scitools Distribution** |  |
| **2.1.1 NPO model** |  |
| **2.1.2 model comparison (NPO vs NPO)** |  |
| **2.1.3 Plots of visualize how Session and Time affect the probability of students being in different science conception categories (scitools\_grouped)** |  |
| **2.1.4 Plot of Posterior Predictive Checks (PPCs) for Model(scitools\_grouped)** |  |
| **2.1.5 Posterior Predictive Checks (PPCs) for the Model(scitools\_grouped)** |  |
| **Scievidence Distribution** |  |
| **2.2.1 NPO model** |  |
| **2.2.2 model comparison (NPO vs NPO)** |  |
| **2.2.3 Plots of visualize how Session and Time affect the probability of students being in different science conception categories (scitools\_grouped)** |  |
| **2.2.4 Plot of Posterior Predictive Checks (PPCs) for Model(scitools\_grouped)** |  |
| **2.2.5 Posterior Predictive Checks (PPCs) for the Model(scitools\_grouped)** |  |
| **Sciexplain Distribution** |  |
| **2.3.1 NPO model** |  |
| **2.3.2 model comparison (NPO vs NPO)** |  |
| **2.3.3 Plots of visualize how Session and Time affect the probability of students being in different science conception categories (scitools\_grouped)** |  |
| **2.3.4 Plot of Posterior Predictive Checks (PPCs) for Model(scitools\_grouped)** |  |
| **2.3.5 Posterior Predictive Checks (PPCs) for the Model(scitools\_grouped)** |  |

**Objective #3**

| **Adaptation Curriculum** |  |
| --- | --- |
|  |
| **Speciation Curriculum** |  |
|  |
| **Adaptation Curriculum** |  |
|  |
| **Speciation Curriculum** |  |
|  |
| **Adaptation Curriculum** |  |
|  |
| **Speciation Curriculum** |  |
|  |
| **Adaptation Curriculum** |  |
|  |
| **Speciation Curriculum** |  |
|  |
| **Adaptation Curriculum** |  |
|  |
| **Speciation Curriculum** |  |
|  |
| **Adaptation Curriculum** |  |
| 1 1.33 1.67 2 2.33 2.67 3  17 37 49 88 26 11 8 |
| **Speciation Curriculum** |  |
| 1 1.33 1.67 2 2.33 2.67 3 3.33  15 30 22 34 24 15 8 3 |

**Extra plots for Objective #1**

**Plots for Distribution and shiftment(ID-tracking)**

| **Curriculum** |  |
| --- | --- |
| **scitools\_pre** |  |
| **scievidence\_pre** |  |
| **sciexplain\_pre** |  |
| **selfsciid\_pre** |  |
| **othersciid\_pre** |  |
| **mindset\_comp\_pre (Converted into factor)** |  |
| **scitools\_post** |  |
| **scievidence\_post** |  |
| **sciexplain\_post** |  |
| **selfsciid\_post** |  |
| **othersciid\_post** |  |
| **mindset\_comp\_post (Converted into factor)** |  |