**Using Metrics From Transcript Data to Predict A Diagnosis of Specific Language Impairment: A Machine Learning Approach**

Mary E. Rodgers

Vanderbilt University

mary.e.rodgers@vanderbilt.edu

Do not delete, move, or resize this block. If the paper is accepted, this block will need to be filled in with reference information.

**ABSTRACT**

Binomial logistic regression can be used to predict diagnostic status when appropriate data are available using R Studio. This study used predetermined linguistic metrics derived from transcripts of children completing a wordless picture task to see if their diagnostic status of Specific Language Impairment or Typical Development could be accurately predicted using binomial logistic regression. Of the available 35 metrics, five were removed for multicollinearity. A Principal Component Analysis categorized 27 of the remaining 30 metrics into three components: Frequency, Length, and Agreement. A machine learning approach using the 10 fold lm method indicated that in a negative correlation, Length and Agreement were significant predictors of Specific Language Impairment. Using this model, the logistic regression indicated that Speech Language Impairment could be accurately predicted at 65%. However, the model could not distinguish between predicting SLI or TD status to a significant degree (kappa = 0.34, p = 0.45).

To promote replication, the data and analytic methods used in this study are freely and publicly available on GitHub [1].

Due to the random selection of the Typically Developing children using the slice\_sample function in R, exact outcomes may differ in a non-significant manner.

**Keywords**

Logistic regression, Machine learning, Principal component analysis, Speech language impairment, Transcript

# INTRODUCTION

Seven to 10 children out of every 100 will be diagnosed with a specific language impairment (SLI) by the time they reach kindergarten [2]. SLI is observed in these children by the symptoms they exhibit, consisting of delays in language development (e.g., speaking, listening, reading and writing skills [2]. Specific symptoms include impairments in vocabulary and grammar, which are negatively correlated with academic achievement [3]. These impairments in vocabulary and grammar are also comorbid with challenging behavior, which are positively correlated with additional challenges in peer relationships and social-emotional development [4].

Several standardized measures can be used to identify delays in expressive (i.e., saying) and receptive (i.e., understanding) language development such as Preschool Language Scales [5] through the production of a standard score and an age equivalence. However, understanding a child's current level of language development can also be assessed through transcripts of their language, viewed in stages of syntactic and morphological development [6] or through descriptive data such as the mean length of utterance or total number of words [7]. These data may be more readily obtained by school staff than standardized assessments, which can only be administered by trained clinicians.

By understanding the significant components found within language metrics derived from transcripts of SLI and typically developing (TD) children, transcript data can be used to predict if a child will receive a diagnosis of SLI.

## Specific Language Impairment

SLI is a developmental delay in a child’s acquisition and use of language communication in the absence of another disability being cited as a reason for this observed delay [3, 4]. SLI has a wide range of proximal and distal effects (Prelock et al., 2008). Proximal effects observed expressively (i.e., what the child says) include impairments in the the development of vocabulary, grammar, phonology, semantics, and pragmatics [3, 8], Proximal effects can also be viewed receptively (i.e., what the child understands) through impairments in decoding and integration [8]. Distal effects of these SLI symptoms are observed in a child’s reading and writing writing abilities [2]. These proximal and distal effects, if left untreated, can lead to long term negative consequences for the child such as challenges in peer relationships, delays in social-emotional development, behavioral challenges, mental health problems, and academic failure [4, 8, 9, 10].

Out of every 100 children entering kindergarten, seven to 10 of them will have already have a diagnosis of SLI [2]. Children with more severe SLI symptoms are likely to how low scores across cognitive, developmental, demographic, and audiometry measures [11]. Similarly, children with the least severe SLI symptoms are likely to score higher in these measures [11]. This would indicate that children diagnosed with SLI from a higher socioeconomic status may have an advantage that children from a lower socioeconomic status do not. This may include when a diagnosis of SLI was made, mediating when treatment and supports for SLI could begin.

While SLI is the most common disability in children, it is frequently the least well detected [8]. Frequently, a ‘wait’ and see approach is used when caregivers have concern over their child’s language development [8], as many children are simply language delayed and not language impaired. Some symptoms of SLI, such as repetitions, interjections, and revisions, are typical of a young child developing their expressive language, but appear in higher rates of children with SLI and impacts their academic performance [3, 8]. Conversely, SLI may be overidentified in diverse populations where English is a second language [9]. Given the proximal and distal effects in SLI, early identification is imperative so treatment may begin.

Impairments in receptive and expressive language ability are often diagnosed after a battery of standardized assessments, such as the PLS-5 [5]. The benefit of these standardizes assessments are that they produce a standard score and age equivalence, which allows the child’s ability to be compared to their peers. Being a standard deviation below their peers would indicate a marked impairment that should receive intervention. However, trained clinicians must administer these assessments and large sample sizes are required for developmental norms. Both of these requirements may further limit identification of SLI.

Another way to evaluate a child’s language ability is through the analysis of transcripts. This can be done by looking at what is written by the child [12], or by transcribing what is spoken by the child [13, 14]. Advantages of using transcripts include smaller sample sizes for developmental norming [15] and stability in measurement across small transcript lengths [16]. With the advent of free automated transcription using Artifical Intelligence, such as Whisper AI [17], transcripts can be more obtainable than standardized assessments for every child. Identifying SLI through transcripts may be a viable solution to the underidentification problem.

## Purpose Statement

The objective of the current study was to develop a logistic regression model that could predict SLI or TD status based off of pre-determined language metrics derived from transcripts using R Studio [18]. In order to do this, a principal component analysis (PCA) on the transcript data of SLI and TD children will be conducted in order to create predictor variables. Then, a machine learning approach will be used to see if any of the components created from the PCA are significant predictors of SLI. The final model produced will be used to run a logistic regression between the predictor components and the outcome variable of SLI or TD status.

## Research Questions

Our primary research question was: 1) What linguistic features in transcript data can predict SLI status?

Our secondary research question was: 2) Can the model accurately predict SLI or TD status to a significant degree?

These questions are addressed through the analysis of the available data set.

## METHODS

## Data

The current data set was obtained from kaggle.com from user DGOKE1 [19] containing 1163 instances of language data derived from transcripts of children between the ages of four to fifteen completing a wordless picture task [20, 21, 22]. This particular data set was created from three different data sets. The data set description on Kaggle indicated 1163 children, 919 typically developing and 346 with SLI. Further analysis of the data indicated that 267 children with SLI were in the data set. The samples from all 1163 children were used in the PCA, and the 267 samples from children with SLI and 267 randomly selected samples from TD children were included in the final logistic regression. Due to the random selection of the TD children using the slice\_sample function in R, exact outcomes may differ in a non-significant manner.

The variables representing the language metrics derived from the transcripts of the children included in this analysis are included in Table 1.

**Table 1. Variables and Descriptions**

| **Variable** | **Description** |
| --- | --- |
| Y | SLI or TD status |
| child\_TNS | total number of sentences |
| freq\_ttr | count of word types to word token ratio |
| r\_2\_i\_verbs | ratio of raw to inflected verbs |
| mor\_words | count of words in the mor tier |
| num\_pos\_tags | count of different part-of-speech tags |
| n\_dos | count of the word ‘do’ |
| repetition | count of repetitions |
| retracing | count of continued abandoned utterances |
| fillers | count of filler words |
| average\_syl | average count of syllables per word |
| mlu\_words | mean length of words |
| verb\_utt | count of utterances with a verb |
| present\_progressive | count of present progressives |
| preposition\_in | count of the word ‘in’ |
| preposotion\_on | count of the word ‘on’ |
| plural\_s | count of plural words |
| irregular\_past\_tense | count of irregular past tense words |
| possessive\_s | count of possessive words |
| uncontractible\_copula | count of uncontractible copulas |
| articles | count of articles |
| regular\_past\_ed | count of regular past tense words |
| regular\_3rd\_person\_s | count of regular third person words |
| uncontractible\_aux | count of uncontractible auxiliary words |
| contractible\_copula | count of contractible copula words |
| word\_errors | count of word errors |
| n\_v | count of nouns followed by a verb |
| n\_aux | count of nouns followed by an auxiliary verb |
| n\_3s\_v | count of third singular nouns followed by a verb |
| det\_n\_pl | count of determinant nouns followed by a personal pronoun |
| det\_pl\_n | count of determinant pronouns followed by a noun |
| pro\_aux | count of pronouns followed by an auxiliary verb |
| pro\_3s\_v | count of singular nominative pronouns followed by a verb |

## Data Source

<https://www.kaggle.com/datasets/dgokeeffe/specific-language-impairment> [9]

## Preparing The Data

The complete csv file was opened in R studio [18] and analyzed using several packages [23-32]. Using packages tidyverse [23] and dplyr [25], a tibble was created that only contained the predictor variables of interest (n = 35) for all 1163 children in the sample. A correlation matrix was then created for all 35 available variables. Multicollinearity could then be assessed by looking for correlations between all predictor variables and identifying variables with a correlation of 0.90 or higher. If correlations among the predictor variables were 0.90 or higher, those variables would be removed from the analysis. A new tibble could then be created that only contained the variables that did not have multicollinearity between them. Another correlation matrix was created to verify that multicollinearity was no longer present. The data in the second tibble could then be scaled so that the metrics the variables represented could be compared to one another, and descriptive data could be assessed. Bartlett’s Test for homogeneity of variance could then be used to see if the variables were related enough to be combined together into components. If so, the Kaiser-Meyeer-Olkin (KMO) Assessment of sampling adequacy could then be done to see if the sample size of each predictor variable was large enough to conduct a PCA. If the sample size of any predictor variable was not adequate, it would be removed from the analysis. Once all parameters were met, the PCA could then be conducted.

## The PCA

A Baseline PCA could then be created to check for loadings above one to determine the number of components in the PCA, and could be visually checked using a sree plot. After the number of components could be determined, the distribution of the residuals could then be checked for normality. If the data were normally distributed, the identified components could then be labeled and joined with the original csv file for analysis.

## Machine Learning

A subset of the data were selected, only including the 267 SLI children and the 267 randomly selected TD children and the predictor variables of interest using the dplyr package [25]. This resulted in a sample of 534 children and 30 predictor variables. A machine learning approach could then be used to create a model in which only significant predictors of SLI status would be contained. This was done by creating a cross-validated model with feature selection using the 10 fold lm method using caret [26]. The machine learning approach allowed for significant components to be identified, which could then be included in the final model used in the logistic regression.

## Logistic Regression

Before running the logistic regression, the model was assessed for suppression effects by looking at the correlations between the predictor component variables and the outcome variable (SLI or TD), to see if any negative correlations between the predictor and outcome variables existed. If so, the predictor variable causing the suppression effects would be removed from the model.

This would produce the final model, used for the logistic regression. The estimates produced in this model would indicate the direction of effect for the predictor components.

# RESULTS

Variables with correlations above 0.90 were identified through the creation of a correlation matrix and removed to keep multicollinearity out of the model (i.e., mor\_words, child\_TNS, mlu\_morphemes, uncontractible\_aux, and det\_n\_pl).

Descriptive data for all current variables were produced using the psych package [31].

Bartlett's test revealed that the R-matrix was not an identity matrix (p < .05), indicating that the remaining variables were similar enough to be grouped together in components.

The KMO assessment indicated that the sample size of the remaining variables were large enough to conduct a PCA, as all variables were above 0.50.

The baseline PCA and sree plot indicated that three components could be made. Visual assessment of the histogram verified that the residuals were normally distributed, indicating that the PCA could be conducted.

The resulting three components of the PCA were named: 1), frequency sentence components (Frequency1), 2) sentence length indicator components (Length2), and 3) pronoun and verb agreement components (Agreement3) based off of the similarities observed between the predictor variables and what they were hypothesized to stand for.

The first component, Frequency1, contained articles (0.86), child\_TNW (0.84), retracing (0.83), freq\_ttr (-0.80), repetition (0.75), n\_v (0.73), present\_progressive (0.71), uncontractible\_copula (0.71), irregular\_past\_tense (0.56), pro\_3s\_v (0.56), n\_aux (0.56), propositions\_in (0.54), propositions\_on (0.51), and n\_dos (0.49).

The second component, Length2, contained verb\_utt (0.76), mlu\_words (0.75), num\_pos\_tags (0.67), plural\_s (0.63), contractible\_aux (0.49), contractible\_copula (0.42), average\_syl (0.41), and possessive\_s (0.33).

The third component, Agreement3, contained regular\_3rd\_person\_s (0.86), n\_3s\_v (0.83), pro\_aux (0.56), and regular\_past\_ed (-0.49). The variables fillers, r\_2\_i\_verbs, det\_pl\_n, and word\_errors were not strong enough to be included in any component. The predictor variables within the components are available in Table 2.

**Table 2. Components**

| **Frequency** | **Length** | **Agreement** |
| --- | --- | --- |
| articles | verb\_utt | regular\_3rd\_person\_s |
| child\_TNW | mlu\_words | n\_3s\_v |
| retracing | num\_pos\_tags | pro\_aux |
| freq\_ttr | plural\_s | regular\_past\_ed |
| repetition | contractible\_aux | - |
| n\_v | contractible\_copula | - |
| present\_progressive | average\_syl | - |
| uncontractible\_copula | posessives\_s | - |
| irregular\_past\_tense | - | - |
| pro\_3s\_v | - | - |
| n\_aux | - | - |
| propositions\_in | - | - |
| preposotion\_on | - | - |
| n\_dos | - | - |

A machine learning approach was used to create a cross-validated model with feature selection using the 10 fold lm method using [16], which indicated that the predictor components of Length2 and Agreement3 were significant predictors of SLI status.

This model was then analyzed for suppression effects. No negative correlations between the predictor and outcome variables were present. The final model used Length2 and Agreement3 as predictors of SLI status in the logistic regression. The negative estimates indicated a decrease in the components Length2 and Agreement3 between the SLI to TD groups.

The logistic regression indicated that Kappa was 0.34, Precision was 0.65, Recall was 0.68, and F1 was 0.66. The output is displayed in Table 3.

**Table 3. Logistic Regression Metrics**

| **Kappa** | **Precision** | **Recall** | **F1** |
| --- | --- | --- | --- |
| 0.3371 | 0.6479 | 0.6758 | 0.6616 |

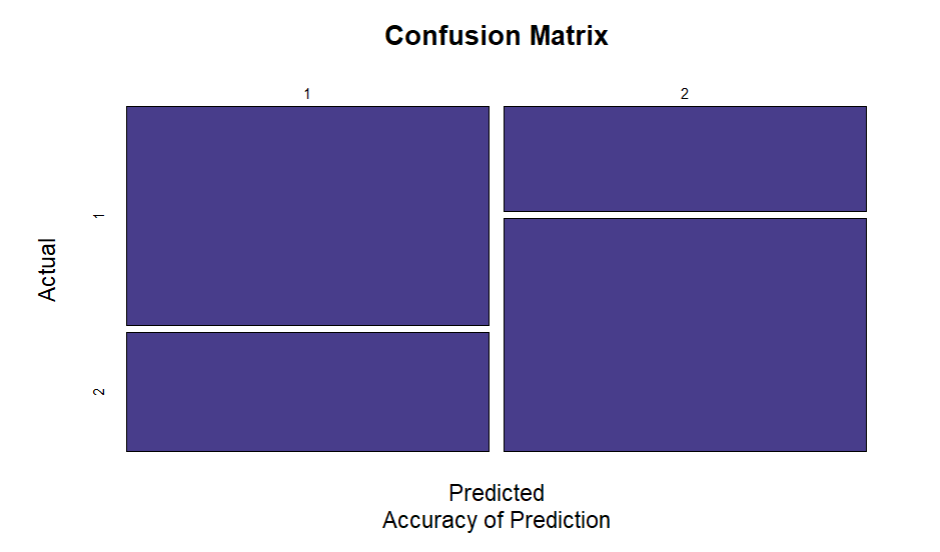
A confusion matrix was created to see the observed and predicted results of the logistic regression. The confusion matrix indicated that of the 267 children with SLI, the model correctly predicted 173, Of the 267 TD children, the model correctly predicted 184.The confusion matrix is displayed in Table 4.

**Table 4. Confusion Matrix**

| **Reference** | **1** | **2** |
| --- | --- | --- |
| **Prediction** |  |  |
| **1** | 173 | 94 |
| **2** | 83 | 184 |

The confusion matrix was then visualized through the use of a mosaic plot. The mosaic plot is displayed in Figure 1.

**Figure 1. Confusion Matrix Visualized via A Mosaic Plot**



Alternative text: A confusion matrix displays the predicted versus observed probabilities of accurately predicting a participant's diagnostic status of speech language impairment. The top left quadrant is large, showing the true positive proportion of predicting speech language impairment. The bottom right quadrant is also large, showing the true negative proportion of predicting speech language impairment. The figure represents how over half of the participants had their diagnostic status of speech language impairment predicted.

# DISCUSSION

In response to the primary research question, "What linguistic features predict SLI?", the 10 fold lm method using caret [26], indicated that the components of Length2 and Agreement3 were the significant predictors, in that decreases in utterance length and agreement were indicative of SLI. The final model produced revealed that out of the 35 original predictor variables, only 12 were predictive of SLI. They included a) the number of utterances consisting of verbs, b) the mean length of utterance, the number of different part-of-speech tags, the number of times a plural was used, how many auxiliary verbs contractible, how many copulas were contactable, the average number of syllables, the number of possessives with an s, the number of regular third person singular verbs, the number of nouns followed by third person singular verbs, the number of pronouns followed by an auxiliary verb, and the number of regular past tense verbs. This indicates that the length and agreement of a child’s utterances is more predictive of SLI, as compared to simply the frequency of the child’s utterances. This ‘quality over quantity’ theory is in line with current research on child language development [6, 7].

However, Length2 and Agreement3 could not significantly differentiate between true positives (accurately predicting SLI status) and false positives (predicting SLI status when actually TD). With this information, we can say that the answer to our secondary research question, "Can the model accurately predict SLI or TD status? to a significant degree" is no. This tells us that while Length and Agreement are in fact significant predictors of a SLI diagnosis, they cannot substantially differentiate between SLI and TD status.

The confusion matrix produced from this model indicated that out of the current sample, in 267 children with SLI the model correctly predicted 173 (65%) and incorrectly predicted 94 (35%). Of the 267 TD children, the model correctly predicted 184 (69%) and incorrectly predicted 83 (31%). The confusion matrix showed that Mcnemar's Test p-value was greater than 0.05 (p = 0.45), which indicated that there was not a significant difference between true positives and false positives.

This indicates that while the Length2 and Agreement3 variables produced from the PCA were significant predictor of SLI status, the model could not distinguish between predicting SLI or TD status to a significant degree. A mosaic plot of the confusion matrix was created, which visualized this result.

While the current approach could not be used to reliability identify SLI, it does demonstrate that using transcript data to identify SLI may be a valid approach that warrants further research [12-15]. Especially with new technologies such as free AI powered transcription [16], what we can learn from transcript data should be explored.

Future research should consider using more components to predict SLI status, or create components from different measures taken from the transcripts such as the number of unique subjects or the number of unique subject-verb combinations [6, 7, 33]

This research was limited in the small SLI sample available (n = 267), and the predetermined variables that went into the PCA from the available data set. Additionally, no pre-post data were available for analysis.

## REFERENCES

1. Rodgers, M. E. FinalProject\_PCA\_ML\_LogisticRegression\_TranscriptData. <https://github.com/Mary-E-Rodgers/FinalProject_PCA_ML_LogisticRegression_TranscriptData>
2. National Institute of Health, Deafness and Other Communication Disorders. NIDCD Fact Sheet: Voice Speech and Language, Specific Language Impairment. U.S. Department of Health and Human Services.
3. Curtis, P. R., Roberts, M. Y., Estabrook, R., & Kaiser, A. P. The longitudinal effects of early language intervention on Children's Problem Behaviors. \*Child Development\*. 2017; 90(2): 576-592.
4. Ervin, M. SLI: what we know and why it matters. \*American Speech and Hearing Association\*. 2001; 6(12).
5. Zimmerman, I. L., Steiner, V. G., & Pond, R. E. Preschool Language Scale, Fifth Edition (PLS-5) [Database record]. PsycTESTS. 2011.
6. Hadley, P. Exploring sentence diversity at the boundary of typical and impaired language abilities. \*Journal of Speech, Language, and Hearing Research\*. 2020; 63: 3236-3251.
7. Hadley, P., McKenna, M., & Rispoli, M. Sentence diversity in early language development: Recommendations for target selection and progress monitoring. American Journal of Speech-Language. 2018; 27: 553-565.
8. Prelock, P. A., Hutchins, T., Glascoe, F. P. Speech-Language Impairment: How to Identify the Most Common and Least Diagnosed Disability of Childhood. Medscape Journal of Medicine. 2008; 10(6); 136.
9. Bourdreau, D., Hedberg, N. L. A Comparison of Early Literacy Skills in Children With Specific Language Impairment and Their Typically Developing Peers. American Journal of Speech-Language Pathology. 1999; 8; 249-260.
10. Shonkoff, J. P., Philips, D. A. From Neurons to Neighborhoods: The Science of Early Childhood Development. National Research Council and Institute of Medicne Committee on Integgrating the Science of Early Childhood Development: National Academic Press. 2000.
11. Beitchman, J. H., Hood, J., Rochon, J., Peterson, M., Mantini, T., Majumdar, S. Empirical Classification of Speech/Language Impairments in Children I. Identification of Speech/Language Categories. Journal of the American Academy of Child & Adolescent Psychiatry. 1989; 28: 1: 112-117.
12. Connelly, V., Dockrell, J. E., Walter, K., Critten, S. Predicting the Quality of Composition and Written Language Bursts from Oral Language, Spelling, and Handwriting Skills in Children With and Without Specific Language Impairment. Written Communication. 2012; 29(3).
13. Miller, C. A., Deevy, P. Structural Priming in Children with and Without Specific Language Impairment. Clinical Linguistics & Phonetics. 2006: 20(5) 387-399.
14. Hassink, J., M., Leonard, L. B. Within-treatment Factors as Predictors of Outcomes Following Conversational Recasting. American Journal of Speech-Language Pathology 2010: 19.
15. Tucci, A., Plante, E., Heilmann, J. J., Miller, J. F. Dynamic Norming for Systematic Analysis of Language Transcripts. Journal of Speech, Language, and Hearing. 2022: 65(1).
16. Heilmann, J., Nockerts, A., Miller, J. F. Language Sampling: Does the Length of the Transcript Matter? Language, Speech, & Hearing Services in School. 2010: 41.
17. Radford, A., Kim, J. W., Xu, Tau, Brockman, G., McLeavey, C., Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. 2022.
18. RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, 2020.
19. Dgoke1. (2018). Diagnose Specific Language Impairment in Children, Version 6. Retrieved April 21, 2023 from https://www.kaggle.com/datasets/dgokeeffe/specific-language-impairment.
20. D. Wetherell, N. Botting, and G. Conti-Ramsden. Narrative skills in adolescents with a history of SLI in relation to non-verbal IQ scores. \*Child Language Teaching and Therapy\*. 2007; 23(1): 95--113.
21. Schneider, D. Hayward, and R. V. Dub. Storytelling from pictures using the Edmonton Narrative Norms Instrument, 2006.
22. Gillam, R. Pearson, N. Test of Narrative Language. Austin, TX: Pro-Ed Inc., 2004.
23. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). “Welcome to the tidyverse.” \_Journal of Open Source Software\_, \*4\*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.
24. Wickham, H.. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. <https://ggplot2.tidyverse.org>.
25. Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D. (2023). \_dplyr: A Grammar of Data Manipulation\_. R package version 1.1.0, <https://CRAN.R-project.org/package=dplyr>.
26. Kuhn, M. (2022). \_caret: Classification and Regression Training\_. R package version 6.0-93, <https://CRAN.R-project.org/package=caret>.
27. Fox, J., Weisberg, S. (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
28. Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. Journal of Statistical Software, 17(1), 1--27.
29. Peterson, B. G., Carl, P. (2020). \_PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis\_. R package version 2.0.4, <https://CRAN.R-project.org/package=PerformanceAnalytics>.
30. JFriedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL https://www.jstatsoft.org/v33/i01/.
31. Revelle, W. (2023). \_psych: Procedures for Psychological, Psychometric, and Personality Research\_. Northwestern University, Evanston, Illinois. R package version 2.3.3, <https://CRAN.R-project.org/package=psych>.
32. Wickham, H. (2022). \_stringr: Simple, Consistent Wrappers for Common String Operations\_. R package version 1.5.0, <[https://CRAN.R-project.org/package=stringr](https://cran.r-project.org/package=stringr)>.
33. Kaiser, A., Roberts, M, & Hadley, P. Maximizing outcomes for preschoolers with developmental language disorder: Testing the effects of a sequentially targeted naturalistic intervention. In Preparation