The Causal Relationship between Specific Language Impairment (SLI) and the Average Number of Filler Words

HYOEUN PARK, University of Toronto

Speech pathologists are eager to understand the underlying reasons of specific language impairment (SLI) in children. Hence, this paper investigates the causal relationship between the treatment (SLI) and the average number of filler words (outcome) using propensity score matching (PSM) in order to understand how speech planning differs between language impaired children and typically developing (TD) children. Results relay that there is not a significant difference between the average number of filler words between SLI and TD children, and that the treatment is not useful in predicting the effect of it towards the outcome. In conclusion, there is no difference in the process of speech planning between treated and untreated children; however, this could be an issue of a Type II error, due to the limitations of PSM.

Keywords: propensity score matching (PSM), specific language impairment (SLI), typically developing (TD), speech planning, filler words, psycholinguistics, speech pathology, nearest neighbour matching

INTRODUCTION

Psycholinguistics, or the psychology of language, is the study of the human mind processing, producing, and understanding human language (Skalicky, 2020). This interdisciplinary work of psychology and linguistics relies heavily on statistical analysis as many theories and hypotheses are constructed through statistical results from experiments. Moreover, these results often help speech-language pathologists to assess and diagnose speech disorders among people, ranging from children to adults (*About Speech-Language Pathology*, n.d.).

However, one of the quests that remain incomplete by speech pathologists is figuring out the causes of specific language impairment (SLI), which is a language disorder in young children who have no difficulty with hearing or intellectual tasks[1] other than those engaged with language (*What is Specific Language Impairment (SLI)*, 2019). It is crucial to diagnose a child with SLI as soon as possible and treat them accordingly not to delay their language performance (Simpson & Rice, n.d.) and have it persist through adulthood. Past studies suggested that children with SLI have weaker ability in speech planning; however, potential drawbacks of these studies were unable to provide the underlying cognitive cause of SLI (Aziz, Fletcher & Bayliss, 2016).

Hence, psycholinguists are eager to further investigate the causes of SLI with the help from statistics. In other words, causal inferences drawn by statistical analysis can assist psycholinguists to diagnose children with SLI. When one wants to conclude on some causation, randomized experiments would be ideal than observational studies (Sheather, 2009). However, it would be unethical or unrealistic in general to randomly assign children to have SLI or typical development (TD). Thus, there needs to be a remedy which could solve the issues of randomization and yielding causal inference.

Propensity score matching (PSM) is a popular method which often comes into play for making causal inference on observational data (Arbour et al., n.d.). According to Alexander (2020), it allows one to estimate the effect of a treatment on the outcome of interest, by using appropriate observable variables that predict getting the treatment. Hence, this report implements PSM on the Specific Language Impairment in Children Dataset (O'Keeffe, 2017) in order to unveil the causal relationship between SLI children and the average number of filler words.

Filler words, or disfluencies, are one of the tools to assess speech planning. They refer to words like *uh, um, ...*, which fill in the gap between speech. In fact, they are widely used in unscripted speech (Sedivy, 2020, p.729). Although filler words seem often distracting and meaningless, it intrigues many psycholinguists as it possibly unveils the mechanisms of utterance planning (Sedivy, 2020, p.736). Brown-Schmidt and Konopka (2008) suggested that fillers are used to smoothly transition from thought to speech, and that they are

[1]For a long period of time, SLI has been known to be a type of speech disorder caused by poor parenting, hearing loss, or minor brain damage at an early age (Bishop, 2006). Now, experts put more emphasis on genetic as the reasoning, but the cause in general remains unclear.

evidence of speech planning is made on the fly. By using filler words as the outcome variable, psycholinguists would be able to grasp an idea on how speech planning works in SLI children.

It is in one's common sense that speaking without filler words is a challenge, unless one carries around scripted speech all the time. In other words, producing filler words are everywhere in daily conversations. Hence, this study hypothesizes that there would be no significant difference between the average number of filler words of SLI children and TD children. This null hypothesis will be tested and illustrated through the PSM method and various figures and tables.

DATA

Original Data

The observations in the Specific Language Impairment in Children Data[2] is a compilation of three datasets created by a data engineer, David O'Keeffe. These three data sets Conti-Ramsden 4, ENNI, and Gillam were provided by the CHILDES Talk Bank project. Moreover, all three datasets included transcriptions (or corpora) of both SLI (treatment) and TD (control) children doing a wordless picture-description task. According to O'Keeffe, these narrative corpora were converted into one data set since transcriptions in general have good potentials in distinguishing children with SLI.

Specifically, the target population, or the set of units wished to be covered in the main objective of the study, of Conti-Ramsden 4, ENNI, and Gillam were a mixture of TD and SLI British adolescents, Canadian children, and U.S. children, respectively. Then, the frame[3] of all three groups were every participant who did the wordless picture-description task, and the concrete number of participants remain unknown. Unfortunately, not all transcriptions recorded during the task were publicly accessible via the CHILDES website. Thus, the sampled population[4] slightly differs from the frame population. The sampled data set of Conti-Ramsden 4 was 99 TD and 19 SLI British adolescents between roughly 13 to 16 years old. Furthermore, the sample of ENNI and Gillam were 300 TD and 77 SLI Canadian children between 4 to 9 years old, and 497 TD and 171 SLI U.S. children between the age of 5 to 12, respectively. The three datasets added up to a sample of 1163 children (or observations) in total.

The data set investigated in this study merges all three aforementioned datasets. Hence, the target population of the Specific Language Impairment in Children Data is British adolescents, Canadian or U.S. children with SLI or TD. Additionally, the frame population is those who did the picture-description task and the sampled population is 896 TD and 267 SLI British adolescents, Canadian or U.S. children. The age distribution in the original data set based on whether the child has the treatment or not is visually displayed below:

[2]See references for the dataset

[3]Frame population: the set of units accessible among the population

[4]Sampled population: The sample that represents the population of the study

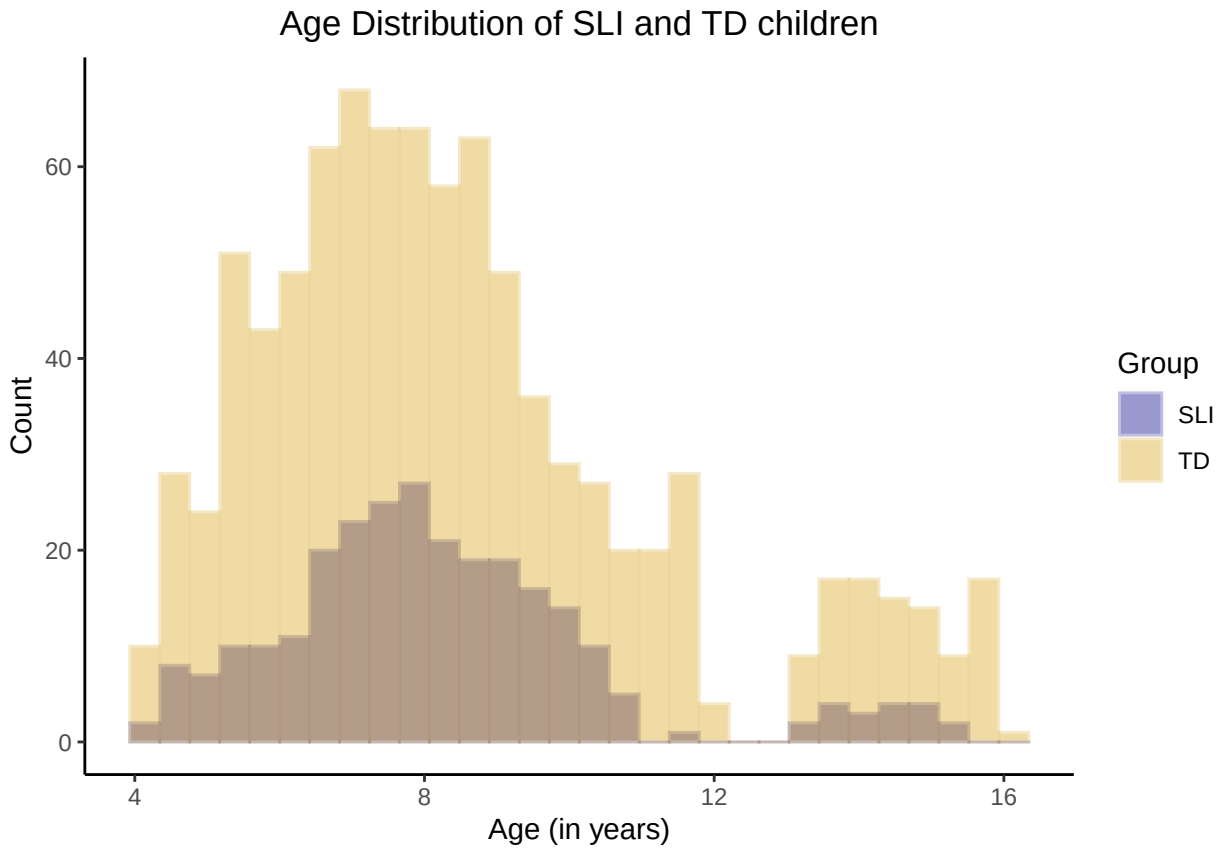

Figure 1: Age (in years) distribution of SLI (specific language impairment) and TD (typically developing) children in the original data set

Based on the numbers given and Figure 1 above, we know that the number of SLI children accounts for approximately 23% of the entire sample and the remaining proportion are TD children. Since it is said that approximately 5-7% of the population in general has SLI (G. Conti-Ramsden and N. Botting, 2006), the proportion of SLI children in the data set is greater than expected. Although the proportion of SLI and TD children in the data set does not represent the "true world", it is still beneficial in a way that there is a substantial amount of SLI observations that can be used in order to investigate the causal relationship with the mean number of filler words.

Each transcription in the sample was analyzed by O'Keeffe through various NLP (Natural Language Processing) and ML (Machine Learning) methods in order to create 64 variables. To be specific, Neural Networks with Feature Extraction, Synthetic Minority Over-sampling Technique (SMOTE), and 10-fold Cross Validation ($k = 2$) were some methods used to create variables with good predictive ability.

Data Wrangling

Not all variables and observations are going to be taken into consideration when analyzing the causal relationship between SLI and the average number of filler words. In other words, there are going to be specific variables selected and observations filtered that meet specific criteria for the purpose of the goal.

To begin with, 1000 samples were randomly selected by the `sample()` function in `R`. This way, it is ensured that the errors are uncorrelated[5]. The number of observations captures approximately 86% of the original dataset, so it is not a serious case of data reduction, and its size is sufficiently large enough the represent the original dataset. Here, note that `set.seed()` is crucial for reproducibility of this randomization process.

[5]One of the multiple linear regression (MLR) assumptions that has to be met in order to have a valid model. Since MLR is going to be used in this report, it is essential to have uncorrelated errors.

On top of that, children above the age of 12 were filtered out for two main reasons. As shown in Figure 1, the distribution of age is bimodal, with each peaking around the age of 8 and 14. Since this study is going to carry out a multilevel regression model (MLR), it is best to have normally distributed variables. Hence, only children below the age of 12 will be considered in the study to ensure a unimodal normal distribution in age[6]. Moreover, this study is interested in younger children than older adolescents as one of the focuses of this report is to help speech pathologists diagnose younger children with SLI. Since leaving SLI untreated can affect one's language performance, it is essential to catch it at an early age and treat it accordingly. Filtering age left the sampled data with 901 observations.

Finally, the following variables are selected for data analysis: Treatment group, Number of filler words, Age (in years), TNS of a child, MLU of words, DSS, Number of word errors, and Retracing[7]. First of all, the Treatment group simply states whether a child has SLI or TD[8]. Then, the Number of filler words counts the total number of filler words (or disfluencies) like *um, umm, uh, uhh, etc.*[9] spoken during the wordless picture-description task. Also, the original dataset consists of Age in both months and years, but the latter has been chosen since it is a more commonly used unit. However, since a child's language development can differ by months, the age in years was kept to two decimal places. Moreover, TNS of a child counts the total number of sentences produced during the task. Similarly, MLU of words, introduced by Brown (1973), measure the mean length of utterance of words produced (Ezeizabarrena & Fernandez, 2018). Here, MLU simply measures the average number of words used per utterance[10]. Finally, DSS is an abbreviated form of Developmental Sentence Score, which is a technique used for evaluating the language status of a child (Lee & Canter, 1971). This method is efficient in quantifying the grammatical structure of speech in children. Descriptive reasons for choosing such variables will be further discussed in **Model Specifics**.

Table 1: Pairwise Correlation Coefficients

	Fillers	TNS	Age (in years)	DSS	Word Errors	MLU of Words
Fillers	1.0000	0.1742	0.0893	0.1241	0.0116	0.0996
TNS	0.1742	1.0000	-0.0478	0.2993	-0.0889	0.0528
Age (in years)	0.0893	-0.0478	1.0000	0.2638	-0.1382	0.4923
DSS	0.1241	0.2993	0.2638	1.0000	-0.2315	0.7273
Word Errors	0.0116	-0.0889	-0.1382	-0.2315	1.0000	-0.2378
MLU of Words	0.0996	0.0528	0.4923	0.7273	-0.2378	1.0000

Since all variables except for the Treatment group are quantitative variables, the pairwise correlation coefficients can be calculated as shown in Table 1. Note that the diagonals equal 1 since they are correlation coefficients within themselves, so those will be ignored.

Observe that there are no strong indications of multicollinearity. That is, there is no strong relationship between the predictor variables that are going to be used later in the models. It is essential to check multicollinearity between predictors since they might yield various side effects when creating a multilevel linear regression model[11]. Computing the variance inflation factor[12] by `vif()` in `R` indicates the multicollearity is not strong for all variables. All variables have VIF values less than 2, which are far less than the cut-off, 5. In conclusion, it is said that all variables are not redundant.

The report will now continue with the sampled data (901 observations, 8 variables) from the original dataset.

[6] This means that all observations in the Conti-Ramsden 4 will be omitted

[7] Note that none of the variables had missing entries (NA's).

[8] For convenience, a dummy variable was created for the two groups so that it can be used in the regression models without using `as.factor()`.

[9] i.e., All variants of filler words are taken into account

[10] Specifically, it is computed by dividing the number of words over the number of utterances.

[11] For instance, it could create an unstable model, give a negative estimated coefficient when a positive one is to be expected, or fail to illustrate the predictive ability of each individual predictor in a model (Sheather, 2009)

[12] This technique measures the strength of multicollinearity between variables in a multilevel regression model (*Variance Inflation Factor (VIF)*, 2018).

METHODOLOGY

This report aims to find the causal relationship between Specific Language Impairment children and their average number of filler words. Thus, propensity score matching will be used on the sample from the Specific Language Impairment in Children Dataset. Details will be discussed in the upcoming sub-sections.

Model Specifics

About Propensity Score Matching

Propensity score matching (PSM) is a statistical method introduced by Rosenbaum and Rubin (1983), which remedies the following main issues:

- Randomizing observations to treatment groups[13]
- The inability to draw causal inference from observational data[14] (Streiner, 2012).

Specifically, PSM allows one to probe on the average effect of a treatment on the outcome of interest by using several variables to predict receiving the treatment (Austin, 2011). The process involves computing the probability, or propensity score, of receiving a treatment regardless of actually having the treatment or not. Then, a pair of observations, one with the treatment and one without the treatment, with similar propensity scores will be matched. There are various methods for matching, but this report will be considering nearest neighbour matching. At the end, using a multilevel regression model (MLR), the outcome of interest would be compared between the treatment and control group to probe on the causal relationship between the treatment and the outcome. This means that there will be two multilevel regression models generated throughout the process of PSM: One logistic model for calculating the propensity score of a child having the treatment (SLI) or not (TD), and one for examining the effect of being treated on the average number of filler words.

According to Streiner (2012), PSM reduces bias as well. It is because PSM attempts to mimic randomization by matching pairs of a controlled and treated unit that have the same propensity score and eventually creates unbiased estimation. That is, PSM reduces selection bias once propensity scores has been calculated by the logistic regression model. For this reason, it is important to choose the appropriate independent variables for calculating propensity scores.

Variable Selection

The variables chosen for MLR's were mentioned in **Data Wrangling**. The Treatment group variable will be treated as dummy variables in the regression models, where 1 and 0 denote SLI and TD, respectively. Then, the Number of filler words is the outcome of interest.

The remaining variables are considered to be predictors for the following reasons. First of all, Age (in years) is one of the crucial predictors required to be included in the model since language development differs based on age (*Language development in children: 0-8 years*, 2020). In fact, it is needless to say that age should be known to figure out whether a child has SLI or not. Moreover, TNS of a child is considered since it is suggested that children with SLI are likely to speak in shorter and less sentences (O'Keeffe, 2017). Similarly, MLU of words is used as well since it has supporting evidence of being a good marker of Specific language impairment in young children (Rice et al., 2010). The higher the MLU value, the higher language proficiency. Furthermore, DSS (Developmental Sentence Scoring) seems to be a crucial predictor in diagnosing a child with SLI, as it quantifies how a child can use complex grammatical structure when producing speech. Similar to MLU, a higher DSS indicates higher language proficiency. Lastly, Number of word errors are also taken into consideration because SLI children tend to have limited knowledge of words (Marshall, 2014). However, note that it is always still possible for someone without SLI to produce word or sound errors[15].

[13] Doing so would arise ethical concerns. For instance, an experimenter cannot randomize people to a smoking group since it have potentials to cause serious aftermaths.

[14] Strong conclusions like causal inference can be made on experimental data instead

[15] E.g., Saying *Let's stop* instead of saying *Let's start*

Equation of the Models

Using the aforementioned variables besides the Number of filler words, the following equation estimates the propensity score of a child receiving SLI:

$$\begin{aligned} y &= \log\left(\frac{p}{1-p}\right) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e \end{aligned}$$

where β_0 = intercept, x_1 = Age (in years), x_2 = Total number of sentences, x_3 = Mean length of utterance of words, x_4 = Number of word errors, x_5 = DSS, and e = Error in measurement. Here, $y = \log\left(\frac{p}{1-p}\right)$ is the log-odds of receiving the treatment. So, β_0 is the log-odds of receiving the treatment when all x_i's $(i = 1, 2, ..., 5)$ are equal to zero. However, β_0 would not technically hold practical interpretation since some predictors like Age, Total number of sentences, Mean length of utterance of words, and DSS cannot equal to zero. In other words, β_0 is meaningless. Also, β_i where $i = 1, 2, ..., 5$ denotes the additive change in receiving the treatment due to the corresponding predictor. So, a positive β_i yields an increase of probability in receiving the treatment, while a negative would make a decrease.

Since the equation above computes the log-odds of a child having the treatment, the equation below must be used in order to find the actual propensity score p of receiving the treatment:

$$p = \frac{\exp(y)}{1 + \exp(y)}$$

Next, the model where the outcome of interest y^f is the average number of filler words will be written as follows:

$$y^f = \beta_0^f + \beta_1^f x_1 + \beta_2^f x_2 + \beta_3^f x_3 + \beta_4^f x_4 + \beta_5^f x_5 + \beta_6^f x_6 + e$$

Here, the x_i's from $i = 1$ to $i = 5$ are the same as the previous model equation, except now x_6 = Treatment group is included to assess the *effect* of being treated on the average number of filler words. This means $x_6 = 0$ when one wants to estimate the average number of filler words of a typically developing (TD) child. Note that the PSM method now allows one to make a causal relationship statement instead of mere correlation. Also, β_i^f where $i = 1, 2, ..., 5$ now measures the additive change in the average number of filler words due to the corresponding predictor. Similar to the logistic model, $\beta_i^f > 0$ increases the average number of filler words but $\beta_i^f < 0$ decreases the average number of filler words. In addition, β_0^f is the average number of filler words of a TD children when all predictors equal zero[16], and e is simply the error of measurement induced by the model.

The benchmark significance level $\alpha = 0.05$ will be used in order to assess the significance of the coefficients.

Nearest Neighbour Matching

Nearest Neighbour Matching is one of the simplest matching methods. As the name itself literally states, a pair of a treatment and control is determined by how *close* they are in terms of their propensity scores. In other words, an treated unit is matched with a controlled unit which has the most similar propensity score[17] (Austin, 2013).

Note that this method can have potential drawbacks of choosing a bad match if the closest neighbour possible is too far away (Glen, 2018). Therefore, it is important to check whether the propensity score has been adequately *balanced* (Austin, 2013). The standardized mean difference[18] will be calculated in

[16]However, the intercept β_0^f does not hold any practical interpretation due to similar reasoning given for the intercept of the logistic model.

[17]If more than one controlled observation has the same propensity score as a treated one, the controlled unit will be selected randomly.

[18]measures whether the independent variables (i.e., covariates) are distributed similarly between the pairs of treated and controlled units (Austin, 2013)

order to diagnose whether the propensity score model has good ability in appropriately matching the pairs. The standardized mean difference can be computed by `stddiff.numeric()` from the `stddiff` package. A standardized mean difference below 0.1 or 0.2 indicates good balance (Du & Hao, 2019).

Moreover, the treated and untreated observations within a pair are not independent due to similar propensity scores. Thus, they are prone to have similar outcomes in the multilevel regression with the number of filler words as the outcome of interest (Austin, 2011). The lack of independence must be taken into consideration when estimating the effect of SLI on the outcome of interest. Hence, a paired t-test[19] can be used as well (Austin, 2011) in order to assess the difference between the effect of treatment on the average number of filler words. Using the paired t-test can help one compare the size of effect of treatment yielded by the regression.

This matching method can be applied by `matching()` in the `arm` package of `R`. The default option `replace = FALSE` will be kept although it can be also carried out with replacement of units. This way, all observations will be used only once. Since there are 214 SLI children, the matched dataset is expected to have 428 observations in total.

RESULTS

Propensity Score

Table 2: Propensity Score

Coefficients:	Estimate	Standard Error	P-value
Intercept	1.487019	0.588897	0.0116
Age (in years)	0.541022	0.068703	3.41e-15
TNS	0.004328	0.003866	0.263
MLU of words	-1.118567	0.118614	$<$ 2e-16
Word errors	0.905256	0.120856	6.87e-16
DSS	0.034653	0.063383	0.5846

Based on the regression output displayed in Table 2, the following expression can be written for estimating the propensity score:

$$\begin{aligned}\hat{y} &= \log\left(\frac{\hat{p}}{1-\hat{p}}\right) \\ &= 1.4870 + 0.5410x_1 + 0.0043x_2 - 1.1186x_3 + 0.9053x_4 + 0.0347x_5\end{aligned}$$

Using the benchmark significance level $\alpha = 0.05$, the significant predictors are \$x_1 = \$ Age, x_3 = MLU of Words, and x_4 = Word errors. On the other hand, x_2 = Total number of sentences and x_5 = DSS are not significant in predicting the propensity score. Note that it is crucial to use appropriate or significant variables when calculating the propensity score.

Also, observe that all coefficient estimates are positive except for MLU of words. Hence, children who are older, produce more number of sentences, create more word errors and have higher DSS are more likely to receive SLI diagnosis. In contrast, an increase in MLU of words attributes to a decrease in the percentage of a child getting diagnosed with SLI. Specifically, for every 1 unit increase in the MLU of words, the log-odds of a child being diagnosed with the treatment decreases by 1.1186, given that all the remaining predictors stay fixed. For instance, a child with an MLU of 4 would have 1.1186 log-odds higher to be diagnosed with SLI than some other child with an MLU scoring 5. This logically aligns with the real world, since a higher MLU value indicates higher language proficiency (as mentioned already).

On the contrary, there seems to be some discrepancies found with the real world among two of the positive coefficient estimates. Specifically, the estimate ($\beta_5 = 0.034653$) of the DSS predictor (x_5) and the

[19] It is used when one wants to know the mean difference between two groups (Paired t-test, 2016)

estimate ($\beta_2 = 0.004328$) of the TNS predictor (x_2) interpret that for every unit increase in DSS and TNS, the log-odds of getting SLI diagnosis increases by approximately 0.035 and 0.004, respectively, given all other independent variables are fixed. However, DSS gives higher scores to more complex sentence structures and later developing forms (Eisenberg, Guo & Mucchetti, 2018). Hence, it is more expected to have a negative β_5 instead. Similarly, β_2 is expected to be positive since SLI children tend to speak in shorter and less number of sentences. Based on Table 1, one could say the coefficient estimate has an opposite sign due to multicollinearity[20] between MLU of Words and DSS. Thus, this is indeed a drawback of this propensity score model.

Matching Propensity Score

Table 3: Matched Pairs

Matched Pair	Propensity Score
(SLI, TD)	(0.007608927, 0.007676580)
(SLI, TD)	(0.023416538, 0.023445083)
(SLI, TD)	(0.024419437, 0.024641698)
(SLI, TD)	(0.032538402, 0.032694478)
(SLI, TD)	(0.037377058, 0.037449990)

Table 3 above shows the first few pairs resulted from the nearest neighbour matching method. As expected, there are 428 out of 901 observations matched in total. In other words, 214 pairs were matched. Note that this data reduction is a drawback, since 473 (roughly 52%) of the observations were not able to get matched. Further discussions of this will be made in the **Discussions** section.

[20] Even though VIF < 5

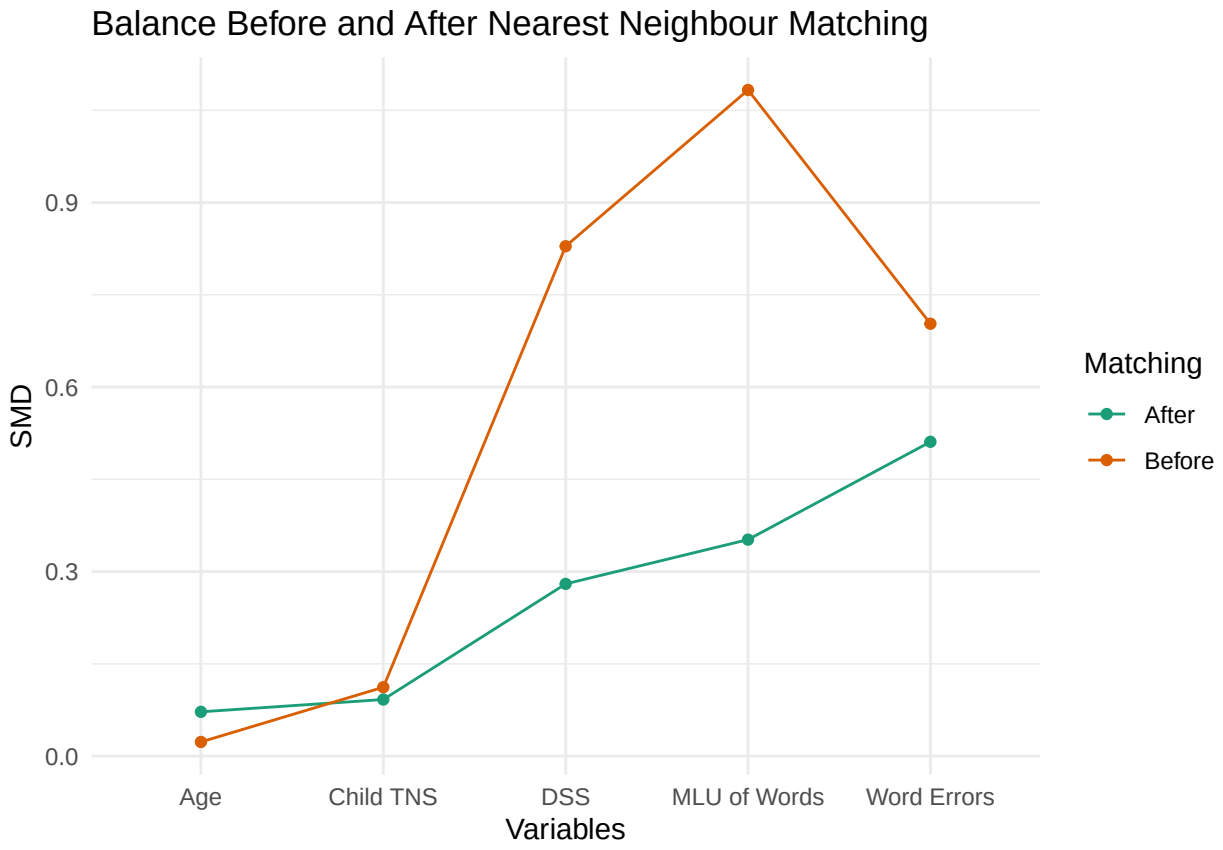

Figure 2: Standardized mean difference of variables before and after nearest neighbour matching

Figure 2 above depicts the standardized mean differences of the dataset before and after matching observations. In general, imbalances has lessened for most variables (Total number of sentences, DSS, MLU of words, and Word errors) except for Age. Nevertheless, the standardized mean differences of DSS (x_5), MLU of words (x_3), and Word errors (x_4) are still above 0.1, which is not a good indication of balance.

Effect of Treatment on the Outcome

Table 4: Causal Relationship

Coefficients:	Estimate	Standard Error	P-value
Intercept	-4.36408	2.56327	0.0894000
Age (in years)	0.35917	0.30123	0.2338000
TNS	0.06606	0.01656	0.0000781
MLU of words	0.90466	0.51340	0.0788000
Word Errors	0.22535	0.31929	0.4807000
DSS	-0.10388	0.28331	0.7140000
Treatment	1.51616	0.90194	0.0935000

According to Table 4, the following equation estimates the effect of receiving SLI on the average number of filler words:

$$y^f = -4.3641 + 0.3592x_1 + 0.0661x_2 + 0.9047x_3 + 0.2254x_4 - 0.1039x_5 + 1.5162x_6$$

Here, the coefficient estimate of β_6^f is approximately 1.5162. Therefore, it is said that receiving the

treatment, SLI, causes a child to increase the average number of filler words by roughly 1.5 (or between 1 to 2 for practical interpretation). However, also note that the p-value of this coefficient is greater than $\alpha = 0.05$. Hence, it does not have strong evidence against the null hypothesis ($H_0 : \beta_6^f = 0$) saying there is no significance difference in the average number of filler words between the treated and untreated groups. Moreover, the treatment is not significantly useful in predicting the average number of filler words.

Then, computing the paired t-test by the `t.test()` function in `R`, the mean difference of the number of filler words between the treated and untreated group is roughly 1.0935 with a 95% confidence interval of (-0.6904, 2.8773). This means that children with SLI produce approximately 1 filler word more on average than those who are typically developing. However, the p-value of this paired t-test is $0.2283 < \alpha = 0.05$, indicating there is weak evidence against the null hypothesis (H_0), like the regression method above. In other words, 1.0935 does not hold strong significance.

Table 5: Comparing the Size of Effect

Method	Difference (95% Confidence Interval)	P-value
Regression	1.5162 (-0.2567, 3.2890)	0.0935
Paired T-test	1.0935 (-0.6904, 2.8773)	0.2283

Table 5 above is an overview of the effect of receiving a treatment on the average number of filler words based on two different methods. In general, both methods suggest children with SLI does not cause themselves to produce significantly different average number of filler words than those who are typically developing.

DISCUSSIONS

Conclusion

The main findings of this paper suggests that the effect of treatment on a child causes the average number of filler words to increase by approximately 1 to 1.5, given that all other conditions are fixed. However, the p-values of the regression method and paired t-test both indicates the effect of receiving a treatment is not strong compared to the control. That is, there is weak evidence against the null hypothesis that says receiving a treatment causes no significance difference in the average number of filler words (i.e., null treatment effect).

To reiterate, filler words or disfluencies like *uh, um, …* are conventionally used to transition thought to articulatory speech. This means speech planning is done on the fly instead of planning everything before start speaking. Fillers can possibly help psycholinguists to identify whether a person is planning speech in smaller chunks (e.g., prepositions, adjectives, nouns) than larger ones (clauses, phrases) if there are more number of filler words produced in the midst of speech. In the case of this study, the number of filler words produced was not significantly different between the treated and untreated groups. Thus, there is no significant difference in the process of speech planning between SLI and TD children. In other words, SLI and TD children plan similar amount of chunks before uttering them. However, it remains unclear in general why there is not a difference of speech planning between SLI and TD children.

Weaknesses and Next Steps

This paper includes some limitations coming from the dataset and the propensity score matching technique, which are further discussed in the upcoming subsections.

Original Dataset

The aim of this paper is to investigate how speech planning differs between language impaired children and typically developing children, so that speech pathologists can diagnose and treat the children with SLI accordingly. This does *not* limit to those who are speaking English. Unfortunately, this dataset only includes children who speak English. Even though English shares some universal linguistic features with other languages in the world, English cannot be a representative of all other languages. In other words, SLI

children that speak English cannot represent SLI children that speak other languages. Hence, this dataset can be improved by including datasets of SLI and TD children speaking other native languages, so that it leaves space for cross-linguistic studies as well. However, note that finding transcription data of SLI and TD children are rare, so this indeed is going to be a challenge.

Propensity Score Matching

Although Propensity Score Matching is a widely used method in statistics, Austin (2008) states that this method tended to be poorly applied in the medical field from 1996 to 2003. Likewise, PSM arises many potential drawbacks in this paper.

One of the most common cons of PSM is reducing the number of observations in the dataset (Streiner, 2012). Similarly, this paper addresses a serious issue of data reduction. The sampled dataset included 901 observations but ended with 428 matched pairs after nearest neighbour matching, meaning the dataset size reduced by roughly 52%. This sample size might not only be representative of the population, but also increase the level of imbalance (King, 2019). King (2019) further argues that imbalance also arises biases in the results. Indeed, this report also includes some imbalances as shown in the computation of standardized differences. Furthermore, data reduction can also plausibly make statisticians fail to find differences between groups that may actually exist in the population. This is called Type II error (Banerjee et al., 2009). This paper could have potentials of a Type II error as well since the main results suggested there was not a significance difference in speech planning between SLI and TD children. For such problems, increasing the sample size or finding a dataset with more samples can be a suggestion. Since PSM has great possibilities of omitting many subjects, it is best to find a dataset with the largest size as possible .

On top of that, the number of pairs matched based on the matching method used can change the results of the study. This paper uses one of the simplest matching methods, nearest neighbour matching, which can often match pairs of treated and untreated subjects with significantly different propensity scores. Such issues could have potentially made Type II error for this paper as well. Hence, other matching methods such as caliper matching, radius matching, stratification matching, and kernel matching can be taken into consideration as well to compare the results. The details of each matching method will not be elaborated here. However, in a nutshell, each method has different ways in choosing closeness of propensity scores and matched pairs, meaning they all come up with different subsets of observations (Streiner, 2012). Therefore, all matching techniques can slightly differ or even change the results. So, it is ideal to compare the findings from each matching method and yield conclusions from them.

Lastly, it is also crucial to ensure appropriate covariates have been chosen. According to Austin (2011), covariates that effect *both* the outcome (i.e., number of filler words) *and* treatment (SLI) or covariates that effect only the outcome are more ideal than covariates that effect only the treatment, especially because they reduce the value of mean squared error (MSE)[21]. This report could have also chose inappropriate variables since some variables showed insignificance in the logistic regression model. However, it is hard to conclude such statement in general since precisely selecting predictors that affect both the treatment and outcome are difficult. Therefore, a suggestion could be including all potential predictor variables that are in the dataset (Austin, 2011) and apply forward AIC or BIC[22] to find those that are significant for calculating the propensity score.

[21] Lower MSE indicates a more stable regression model.

[22] A variable selection method which chooses the variables in a multilevel regression model based on small p-values, low AIC/BIC values, etc.(Sheather, 2009)

REFERENCES

Alboukadel (n.d.). How To Create Histogram By Group In R. *Data Novia*. Retrieved from https://www.datanovia.com/en/blog/how-to-create-histogram-by-group-in-r/

Alexander, R. (2020, November 5). Matching and Difference-in-Differences. *Telling Stories with Data*. Retrieved from https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html#matching

Arbour et al. (2014). *Propensity Score Matching for Causal Inference with Relational Data.* Amherst, MA: University of Massachusetts. DOI http://ceur-ws.org/Vol-1274/uai2014ci_paper5.pdf

Austin, P. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*, 46(3), 399–424. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/

Aziz, S.A., Fletcher, J. & Bayliss D.M (2017). Self-regulatory speech during planning and problem-solving in children with SLI and their typically developing peers. *International Journal of Language & Communication Disorders*, 52(3), 311-322. Retrieved from https://pubmed.ncbi.nlm.nih.gov/27511872/

Banerjee et al. (2009). Hypothesis testing, type I and type II errors. *Indian Journal Of Psychiatry*, 18(2), 127–131. https://doi.org/10.4103/0972-6748.62274

Bishop, D.V.M (2006). What Causes Specific Language Impairment in Children? *Current Directions in Psychological Science*, 15(5), 217-221. Retrieved from https://www.jstor.org/stable/20183118?seq=1

Boston University School of Public Health. (2016). *Paired t-test*. Retrieved from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/sas/sas4-onesamplettest/SAS4-OneSampleTtest7.html#:~:text=A%20paired%20t%2Dtest%20is,variables%20are%20separated%20by%20time.&text=Since%20we%20are%20ultimately%20concerned,the%20one%20sample%20t%2Dtest.

Botting, N. & Conti-Ramsden, G. (2006). Specific Language Impairment. In K., Brown (Ed.). *Encyclopedia of Language & Linguistics* (2nd ed., pp. 629-632). https://www.sciencedirect.com/science/article/pii/B0080448542008440

Canter, S. M. & Lee, L. L (1971). Developmental Sentence Scoring: A Clinical Procedure for Estimating Syntactic Development in Children's Spontaneous Speech. *Journal of Speech and Hearing Disorders*, 36(3), 315-340. https://doi.org/10.1044/jshd.3603.315

Du, Z. & Hao, Y. (2019). *Package 'stddiff'*. Retrieved from https://cran.r-project.org/web/packages/stddiff/stddiff.pdf

Eisenberg, S. L., Guo, L. & Mucchetti, E. (2018). Eliciting the Language Sample for Developmental Sentence Scoring: A Comparison of Play With Toys and Elicited Picture Description. *American Journal of Speech-Language Pathology*, 27(2), 633–646. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6105120/

Ezeizabarrena, M. & Fernandez, I. G (2018). Length of Utterance, in Morphemes or in Words?: MLU3-w, a Reliable Measure of Language Development in Early Basque.

Frontiers in Psychology, 8(2265). https://doi.org/10.3389/fpsyg.2017.02265

Fox et al. (2020). *Package 'car'*. Retrieved from https://cran.r-project.org/web/packages/car/car.pdf

Gelman, A., et al. (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. https://cran.r-project.org/package=arm

Glen, S. (2018, June 13). *Nearest Neighbor Matching: Definition*. Retrieved from https://www.statisticshowto.com/nearest-neighbor-matching/

Hayes, A. (n.d.) *broom v0.7.3*. RDocumentation. Retrieved from https://www.rdocumentation.org/packages/broom/versions/0.7.3

Investopedia. (2018). Variance Inflation Factor (VIF). *Investopedia*. Retrieved from https://www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=Variance%20inflation%20factor%20(VIF)%20is,only%20that%20single%20independent%20variable

Keeffe, D. (2017). Diagnose Specific Language Impairment in Children. (6). [Data file]. Retrieved from https://www.kaggle.com/dgokeeffe/specific-language-impairment

King, G. & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4). https://doi.org/10.1017/pan.2019.11

Marshall, C. R (2014). Word production errors in children with developmental language impairments. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 369(1634). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866419/

Miller, S. V (2020, September 11). Another Academic R Markdown Article/Manuscript Template [Blog post]. Retrieved from http://svmiller.com/blog/2020/09/another-rmarkdown-article-template/

Miller, S. V (2020, November 5). *Package 'stevetemplates'*. Retrieved from https://cran.rstudio.com/web/packages/stevetemplates/stevetemplates.pdf

Rice et al. (2010). Mean Length of Utterance Levels in 6-month Intervals for Children 3 to 9 Years with and without Language Impairments. *Journal of Speech, Language, and Hearing,* 53(2), 333-349. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2849178/

Rice, M. L. & Simpson, J. (n.d.). Top 10 Things you should know about children with Specific Language Impairment. *The Merrill Advanced Studies Center*. Retrieved from https://merrill.ku.edu/top-10-things-you-should-know

Sedivy, J. (2019). *Language in Mind: An Introduction to Psycholinguistics* (2nd ed.). New York, NY: Oxford University Press

Sheather, S.J. (2009). *A Modern Approach to Regression with R* (3rd ed.). New York, NY: Springer Science & Business Media

Skalicky, S. (2020). Psycholinguistics. In S., Pritzker & M., Runco (Ed.). *Encyclopedia of Creativity* (3rd ed., pp. 399-403). Retrieved from https://www.sciencedirect.com/science/article/pii/B9780128093245219135

Specific Language Impairment. (2019, 21 October). *National Institute on Deafness and Other Communication Disorders (NIDCD).* Retrieved from https://www.nidcd.nih.gov/health/specific-language-impairment

Speech-Language Pathologists. (n.d.). *American Speech-Language-Hearing Association (ASHA).* Retrieved from https://www.asha.org/students/speech-language-pathologists/

Streiner, D. L & Norman, G. R (2012). The Pros and Cons of Propensity Scores. CHEST Journal, 142(6), 1380-1382. https://doi.org/10.1378/chest.12-1920

Wickham, H., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Xie, Y. (2020). *knitr: A General-Purpose Package for Dynamic Report Generation in R.* R package version 1.30. Retrieved from https://CRAN.R-project.org/package=knitr

APPENDIX

A. Github Repository

Here is the link to this report's Github repository: https://github.com/hynprk/SLI-and-Filler-Words