The School of Mathematics

THE UNIVERSITY
of EDINBURGH

The Effect of Bilingualism on Executive Function

by

Elena Shaw, s1932415

August 2020

Supervised by
Dr. Michael Allerhand and Dr. Gordon Ross

Acknowledgments

The author wishes to express her gratitude to the advisory team Michael Allerhand, Gordon Ross, and Riki Herliansyah for their help, guidance, and patience. She also thanks Annemiek van Essen for being the author's personal consultant on children psychology.

Own Work Declaration

The author, Elena Shaw, here declares that the contents of this report include only her original work, with inspiration drawn from her advisors and colleagues.

Contents

Executive summary

Background:

There is little existing research on the effect of bilingualism on children with autism (ASD). ASD is a gradated condition commonly associated with difficulties in verbal expression and executive functions. Current findings suggest that children's exposure to bilingualism will not exacerbate their ASD symptoms, contrary to parents' and clinicians' fears.

Research objective:

This report investigates the effect of bilingualism as moderator or mediator on Executive Functions for bilingual children in the study as well as subsetting by gender and diagnosis.

Data:

89 bilingual families in Scotland with either TD or ASD children aged 5 to 13 volunteered to participate in a study consisting of a parent's report on their children as well as a battery of assessments performed with the child by a researcher.

Methods:

Factor analysis is used to identify a plausible model for measuring executive function. Causal Analysis is conducted to investigate the influence of bilingualism as moderator of the effect of ASD on executive functions as well as the influence of bilingualism as mediator between language ability and executive functioning.

Results:

Factor Analysis showed that a 7 metric, single latent factor model showed near perfect fit to the data. However, dimensionally reduced data according to this model showed no evidence of bilingualism having any moderation or mediation affect on executive functions.

1 Introduction

For families with access to multiple language environments, the decision to raise bilingual children can be intricately tied to the child's access to culture, identity, and community. For families with a child diagnosed with autism (ASD), the same considerations are in play but are further confounded by the child's condition (Beauchamp & MacLeod (2017), Hampton et al. (2017)). Both Beauchamp & MacLeod (2017) and Hampton et al. (2017) point out that parents' inability to comfortably express thoughts and sentiments across a language barrier negatively impacted relationship development with their child. Unfortunately, there is not enough research and empirical evidence to inform health care professionals, who are key resources to such parents, the impact of bilingualism on children with autism.

Bilingualism Matters, a research and information center at the University of Edinburgh, currently runs a research project which has recruited just under 100 families residing in Scotland to participate in a study on the effect of bilingualism on both typically developing (TD) children and children with ASD. While this research project studies several cognitive processes associated with ASD, the current report will focus on Executive Functions (EF). Specifically, this report investigates the effect of bilingualism on Executive Functions for TD children and children with ASD.

2 Background

This report attempts to account for several critical aspects of the ASD condition. First, that ASD is not a binary condition. It is generally diagnosed based on a gradated severity of the condition. In practice, clinical diagnosis involves a scaled assessment using the social communication questionnaire (SCQ). Second, that ASD is a condition which interacts with gender. Traditionally, boys made up the majority of diagnosis but recent research has found that there may be differences in how this condition presents in males versus females, allowing the latter the ability to camouflage as asymptomatic (Dworzynski & Happé (2012), Hartley & Sikora (2009), and Carter et al. (2007)). Third, that ASD is often associated with poor mastery of verbal communication and executive functions. Since bilingualism is intricately tied to verbal communication, this report considers bilingualism in the broader context of language ability. However, the main focus of this report is the effect of bilingualism on executive functions.

Executive function (EF) is an umbrella term used to classify a set of cognitive processes, exhibited as skills, which have historically been tied to the prefrontal cortex (Goldstein et al. (2014)). With no shortage of definitions in literature, neither is there a comprehensive, agreed-upon list of skills which fall under this term. As such, this report defaults to the set of nine clinical skills defined by the researchers of this project.

Section 3 will provide the scope of this report as constrained by the data set provided by Dr. Rachael Davis who heads the current research project. This section introduces the assessments under consideration and how they map to the latent constructs of interest, namely bilingualism, EF, and language ability. Section 4 gives a detailed plan on how to proceed with statistical analysis in order to achieve rigour through guided discovery. Lastly, section 5 discusses the results of this process on the provided data set.

3 Data

Both parents and children from bilingually exposed families contribute to data. Parents fill out a survey assessing their children on a Likert scale for provided prompts (e.g. "My child understands the difference between lies and jokes"). Children engage in play-based assessments with a researcher and their responses are recorded in several ways. To accommodate for non-verbal children, eye-tracking technology was used for some assessments. Among the 89 children involved, there were 37 girls and 52 boys, 38 children diagnosed with ASD and 51 TD children. Figure 1a shows the age distribution across gender and diagnosis. Since ASD exists on

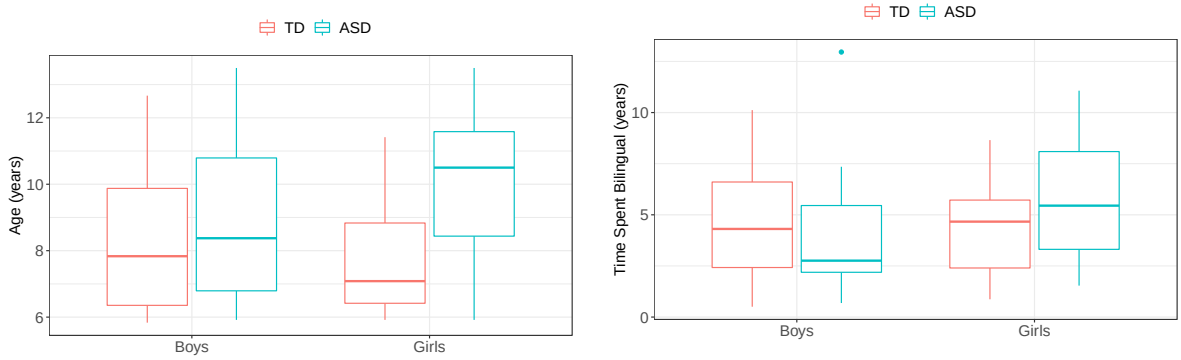
a spectrum, the data set provides two metrics related to the condition. The first is a binary variable indicating whether a child is clinically diagnosed with the condition (ASD) or not (TD). Non-diagnosed children had to complete the SCQ assessment and are only included in the study if they score below 15 [1]. As a result, the second metric, SCQ, is only available for TD children.

3.1 Bilingualism and Language Ability

Since there is no standard scale for measuring bilingualism, families self-reported several variables related to the child's exposure. Given that some children were non-verbal, in this report, exposure is confined to a child's time spent hearing multiple languages. The metric of interest in this report leverages a zero to one scale defined by researchers on the project called "Overall bilingual input". On this scale, a child with equal time exposure to both languages would have a score of one while a child with no exposure to a second language would score a zero. This measure can be interpreted as a proportion of time spent exposed to multiple languages. Based on this interpretation, the author of this report quantifies bilingualism by estimating a child's total exposure to bilingualism as a duration, hereby referred to as "years spent bilingual" (YSB):

$$\text{YSB} = \text{Overall bilingual input} \times (\text{current age} - \text{age at first exposure})$$

Since time and exposure are important aspects of language acquisition (Beauchamp & MacLeod (2017)), this metric provides an intuitive variable to quantify bilingual exposure. The distribution of YSB across gender and diagnosis can be seen in figure 1b.



(a) Distribution of age across gender and diagnosis

(b) Distribution of bilingual exposure (TSB) in years across gender and diagnosis

Figure 1: Distribution of relevant demographics associated with the 89 children involved in the study.

For language ability, only two metrics are collected, both based on a direct assessment done with the child. The assessment is conducted in English using the British Picture Vocabulary Scale (BPVS) in which children match a word to the correct one of four available pictures. There are two scores associated with this task: VCR, the net total number of correct responses (i.e. total correct responses differenced by total incorrect responses), and VRT, the eye-tracked response time until a child looks to and fixates on the correct response. Figure 2 shows the distribution of these 2 scores across gender and diagnosis.

3.2 Executive Function

Executive Function (EF) was measured in three ways, a parent report and two assessments done with the child. The report asks parents to rate on a Likert scale their child's behavioral regulation and metacognition. Behavioral regulation is divided into subcategories on inhibition (BIH), shift (BS; the ability to change between tasks and activities), and emotional control

[1] the typical clinical cutoff for ASD diagnosis.

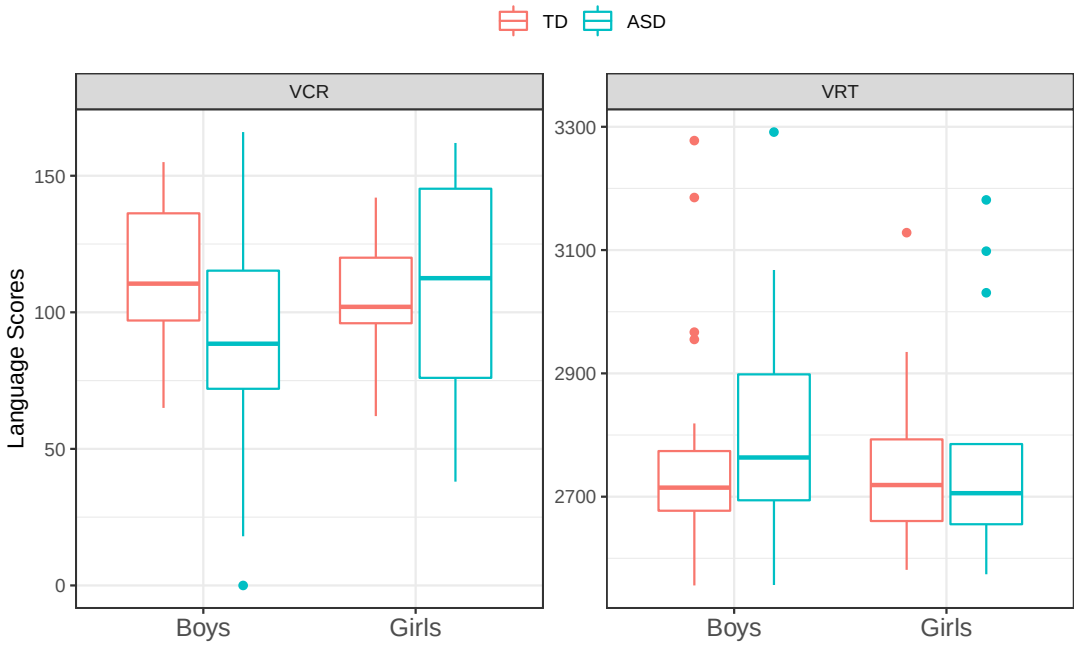


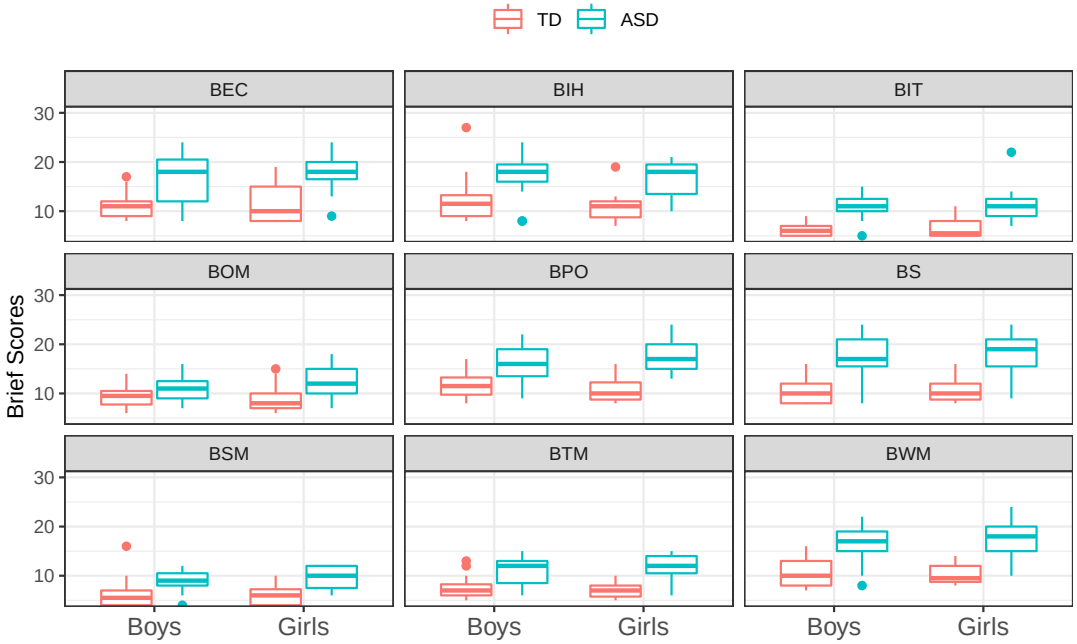
Figure 2: Distribution of language score VCR as counts and VRT in milliseconds across gender and diagnosis.

(BEC). Metacognition included six subcategories: initiation (BIT), working memory (BWM), planning (BPO), organization of materials (BOM), self-monitoring (BSM), and task monitoring (BTM; assessing performance on a completed task). Higher reported scores indicate higher dysfunction.

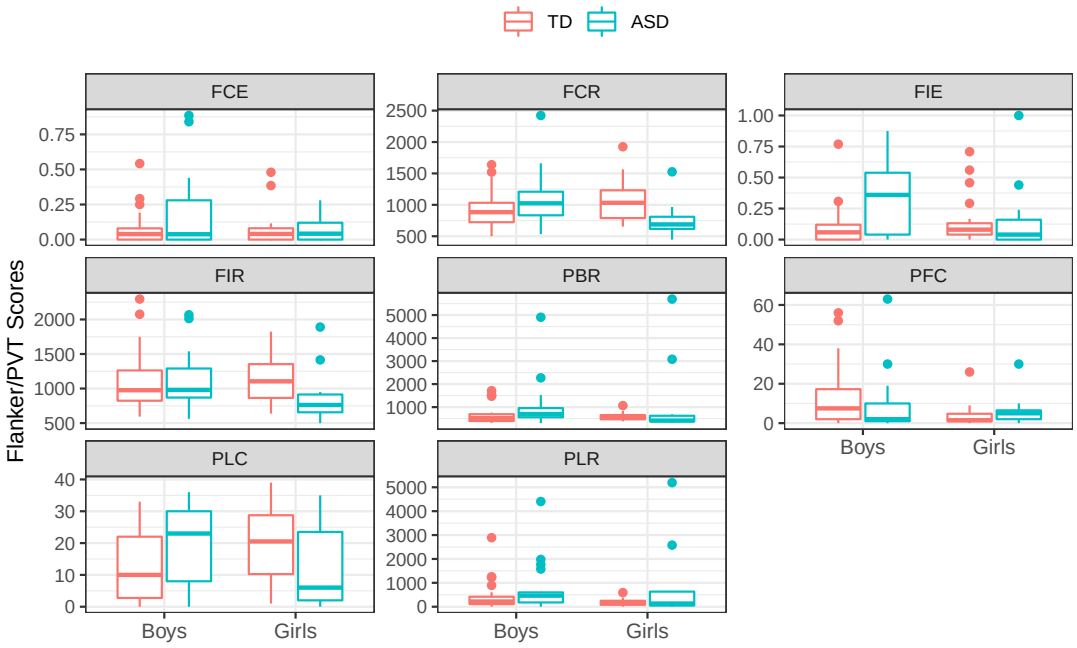
The two assessments conducted with the child are the Flanker Task, a cognitive test commonly used to measure inhibition, and the Psychomotor Vigilence Task (PVT), which is generally used to measure sustained attention. Outcomes of the Flanker Task include mean reaction times (FCR,FIR) and proportion of correct responses (FCE, FIE) for congruent (FCE, FCR) and incongruent (FIE, FIR) trials. While the Flanker Task is a well established assessment of EF, there is less evidence for PVT; PVT is most commonly used in measuring cognitive impairment with sleep deprivation. However, Abad & Guilleminault (2012) reports that EF Impairment is associated with 2 metrics in PVT: lapses, defined to be reaction times that are greater than 500ms, and false responses, responding when there is no stimuli. The data set contains four metrics associated with PVT, an individual baseline metric (PBR; mean reaction time for valid responses), two metrics related to lapses (number of lapses (PLC) and additional reaction time beyond the 500ms threshold (PLR)), and a count metric of false responses (PFC). This report will consider all four PVT metrics with the recognition that these responses may pose validity challenges to the measurement of EF. Altogether, there are 17 metrics associated with EF and their distributions across gender and diagnosis can be seen in figure 3.

4 Methods

This report makes use of causality modeling techniques used in Psychometric research which are based on common statistical modeling methods but differ in its inference. The philosophical details behind this difference are outside of this report's scope aside from the following brief explanation. Because this report is fundamentally concerned with the effect of bilingualism on executive functions, causality, along with its implications on time and counterfactual dependencies, is an inseparable part of the research. However, traditional statistical modeling cannot fully represent such structures. Enter structural equation modeling (SEM), a multivariate technique used to analyze observed variables and their theorized underlying structures and relationships.



(a) Distribution of Brief metrics across gender and diagnosis



(b) Distribution of Flanker and PVT metrics across gender and diagnosis

Figure 3: Box plots showing distribution of relevant factors associated with the 89 children involved in the study.

This report will make use of SEM to conduct Factor Analysis, Moderation Analysis, and Mediation Analysis.

SEM techniques leverage the covariances of observed measurements, herein called "indicators", to quantify overlap among them. In theory, this overlap represents the shared latent variable of interest. However, since data are noisy and tests differ, the variances on each indicator cannot be fully attributed to the latent variable. Thus, the total amount of variance for any given indicator can be divided into two components:

1. Common Variance: attributable to the latent factor.
2. Unique Variance: attributable to the specific indicator caused by noise and error.

SEM models the relationship between indicators and their latent variable as a multivariate linear regression. For k indicators $\{y_1, ..., y_k\}$ and a hypothesized latent variable η,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_k \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_k \end{pmatrix} \cdot \eta + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_k \end{pmatrix} \tag{1}$$

where τ_i is the mean of indicator i, λ_i is the weight coefficient of how strongly the latent variable loads onto indicator i, and ϵ_i is the unique error for indicator i. Rewriting equation 1 in the form of variance, if Σ is the variance-covariance matrix of the observed indicators and Φ is the variance-covariance matrix of the latent variable:

$$\Sigma = \Lambda \Phi \Lambda^T + \Theta_\epsilon \tag{2}$$

$$\text{where } \Lambda = \begin{pmatrix} \lambda_1 & ... & \lambda_k \end{pmatrix} \text{ and } \Theta_\epsilon = \begin{pmatrix} \epsilon_1 & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \epsilon_k \end{pmatrix}$$

Since τ_i represents the mean of indicator i and the variance of a mean is zero, indicator means are not represented in this equation. Furthermore, note the off-diagonals in the unique error variance-covariance matrix consists of zeros – this is in line with the assumption of unique variance. Common variance (denoted h^2) is thus related to the weight coefficient loadings through:

$$h^2 = \sum_{i=1}^{k} \lambda_i^2$$

Expanding these equations to include models with more than one latent variable (η) is trivial: equation 1 would take on more weighted covariate terms while matrices in equation 2 would be expanded.

4.1 Construct Validation: Factor Analysis

The current research topic, specifically related to children with ASD, is not a well studied area. Given the lack of knowledge, it is important to conduct analysis work carefully with minimal assumptions. For this report, substantial effort is put into validating the data and research model using factor analysis prior to conducting causal analysis. This validation work is also empirically useful since the analysis can reduce dimensionality. Given that there are 17 metrics designed to measure EF, dimension reduction is a useful step.

There are two kinds of factor analysis: exploratory and confirmatory. Exploratory Factor Analysis (EFA) is meant to be an exploratory tool to understand the interrelationships among indicators. The EFA outcome in this report is a hypothesized model of how the indicators

interrelate. Confirmatory Factor Analysis (CFA) reverses this knowledge and checks how well the data fit a proposed model. This report uses CFA to quantify the fit of the model found in EFA.

Informed by recommendations from Watkins (2018) and Nerbonne (2016), the author of this report conducts her construct validation work as summarized by the steps below. Steps 1-3 are commonly considered part of the EFA process.

1. Screen Data: "Do the data meet the conditions for EFA? How can the data be adjusted to improve conditions?"
2. Conduct EFA: "How many components or factors should be retained?"
3. Interpret results: "What do the components or factors represent?"
4. Evaluate fit (CFA): "How well does the resulting model fit the data?"

Screen Data: Watkins (2018) argues that there are several considerations which affect EFA results. In some cases, these conditions may determine whether EFA is the appropriate tool to use:

- Data should be factorable, i.e. correlations should be due to common variances.
- Most factors should have correlations greater than $|\pm .30|$.

Watkins (2018) recommends two objective tests to check that the first condition is met: Bartlett's test of sphericity, which should come back significant against the Null hypothesis that data are randomly generated, and the KMO measure of sampling adequacy, which should yield a ratio of correlation to partial correlation greater than 0.7.

The second condition is more involved and includes considerations regarding linearity and normality. Both conditions can be subjectively examined using scatter plots, but should be supplemented with quantitative metrics. Linearity can be checked using Ramsey's RESET test while normality can be measured by skew and kurtosis. Conditions which affect linearity and normality include missing data and outliers. To handle missing data, this report uses multiple imputation as it is the least controversial approach. For outliers, this author follows the statistics belief that outliers should not be removed from analysis unless there is near certain evidence that they are due to collection error and therefore not a true observation. As will be discussed in section 5, there is only one such case and a sensitivity analysis approach is taken to compare results with and without the outlier replaced by imputed data. To adjust for normality and linearity conditions, this report uses a combination of Pearson's and Spearman's correlations depending on the nature of a given variable. While Pearson's correlation is the more commonly used metric, its assumptions of normality and linearity attenuate the results of variables that violate these assumptions. Spearman's rank-order correlation is recommended as a robust alternative (Watkins (2018)).

Conduct EFA and Interpret results: At its core, the goal of EFA is to reduce the number of dimensions: given a set of indicators, derive the common underlying latent variable. A common statistical tool to accomplish this is Principal Component Analysis (PCA) which uses eigen decomposition to derive orthogonal sets of components. These components represent the data transformed according to non-interacting orientations of maximal variance in decreasing order. A disadvantage of this method concerns interpretability since it is not guaranteed that the optimal orientation maps to meaningful metrics. Furthermore, PCA does not distinguish the subtleties between common versus unique variances. As such, another common EFA technique is Principal Factor Analysis (PFA).

Mathematically, PCA and PFA are similar. The difference originates in the common-unique variance breakdown. PFA assumes that due to unique variances, the diagonal communalities cannot be 1. It estimates the values of this diagonal iteratively. In practice, the matrix results

of PFA are rotated to facilitate factor interpretation. This report uses the R 'psych' package implementation of PFA.

Both Watkins (2018) and Nerbonne (2016) recommend a combination of PCA and PFA to determine the number of components/factors to retain. Specifically, Watkins (2018) recommends using PCA results to determine the number of components to retain and then interpreting those components as factors using the output of PFA. A subjective method is to consider the scree plot. Watkins (2018) recommends supplementing this with parallel analysis and minimum average partials (MAP) criterion. Parallel analysis compares eigenvalues from PCA to the eigenvalues of a simulated data set. If an eigenvalue from PCA is lower than that of simulated data, then the component is accounting for variance less noticeable than random noise. MAP is based on partial covariances and considers the ratios of common to unique variances for a given number of components. If this value is calculated for all subsets of possible components, the minimal value (MAP) will represent the component after which unique variances start to dominate common variances. Since PCA components are ordered, all remaining components would thus be poor representatives of latent variables.

Results from previous steps allow data to be dimensionally reduced. For this report specifically, these reduced variables should represent EF. Reduction results from PCA and EFA will differ. Due to a phenomenon called Factor Indeterminacy, the validity and meaning of reduced results from EFA has been a topic of controversy (Acito & Anderson (1986),Rigdon et al. (2019)). In short, factor indeterminacy prevents a consistent, unique set of reduced values. In the extreme, depending on the choice of rotation, a participant's observed scores could be rotated to take on opposite scores on a given scale – there is no guaranteed meaningful mapping. To avoid this controversy, the author of this report only considers PCA reductions in her moderation and mediation analyses.

Evaluate fit (CFA): Results from the previous EFA steps should reveal the underlying structure of EF in the data, but the quality of this model must be examined. This report uses fit metrics and tests reported in the R 'lavaan' package to evaluate fit. These include statistics related to residuals, likelihood compared to saturated model, as well as information loss. Within each category, several related statistic are reported and highly correlated information is produced. As such, given the data's small sample size and low hypothesized number of latent variables, this author takes recommendations from Hu & Bentler (1999) and Urbano & Kiers (2011) by considering the following three statistics from each category:

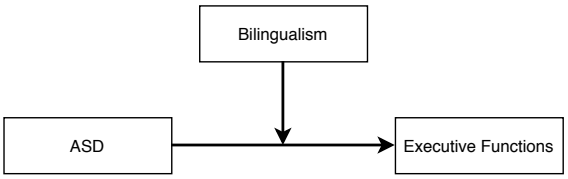
- **Residuals**: standardized root mean squared residual (SRMR); .08 is a recommended threshold for relatively good fit (Hu & Bentler (1999)).
- **Likelihood comparison**: Confirmatory Factor Index (CFI), which is the sample size adjusted chi-square test statistic; CFI of 1 indicates perfect fit and a value greater than .95 is recommended (Hu & Bentler (1999)).
- **Information loss**: Akaike Information Criterion (AIC); since AIC is not standardized, there is no recommendable threshold. As a relative metric, lower values indicate lower information loss and thus a better fit. Urbano & Kiers (2011) recommends AIC over Bayesian Information Criterion (BIC) for small samples with one to three latent variables.

4.2 Causal Analysis: Moderation and Mediation

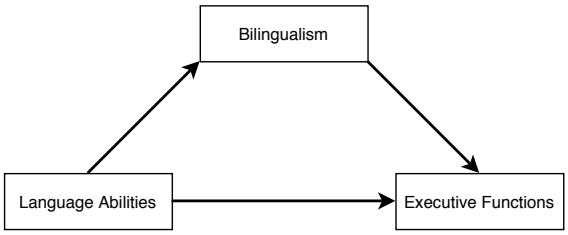
Models for moderation and mediation analyses must be informed by existing evidence of a causal relationship. Beauchamp & MacLeod (2017) cites evidence that "performances on EF tasks appear to be positively correlated with length of time a child has spent as a bilingual". This report investigates whether bilingualism moderates or mediates EF.

As technical terms, moderation and mediation both refer to how a given variable M affects an existing causal relationship between two variables X and Y. Mediation looks at how variable M conditionally changes the nature of the relationship between X and Y, potentially strengthening

or weakening the relationship. It seeks to understand how universal or conditionally dependent the causal relationship is. Figure 4a models how bilingualism could moderate the effect of autism on EF. By contrast, mediation considers how much variable M accounts for the causal relationship between X and Y. In other words, it asks "how much of the existing relationship between X and Y can be explained by variable M?" Figure 4b models how bilingualism would mediate the effect of language ability on EF.



(a) Moderation model.



(b) Mediation model.

Figure 4: Moderation and mediation models to investigate the effect of Bilingualism on Executive Functions.

The statistical technique behind Mediation Analysis is fundamentally a linear regression. For a dependent effect variable Y, independent cause variable X, moderator M_o, coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ and error ϵ,

$$Y = \beta_0 + \beta_1 X + \beta_2 M_o + \beta_3 X \cdot M_o + \epsilon$$

β_3, being the coefficient term for the interaction of X and M_o, would thus be the weight of the moderation effect. In order for a moderation effect to exist, this weight must be non-zero. Moderation need not be linear. If there is reason to believe that linearity is a bad assumption, mediation analysis will commonly add squared mediator terms:

$$Y = \beta_0 + \beta_1 X + \beta_2 M_o + \beta_3 X \cdot M_o + \beta_4 M_o^2 + \beta_5 X \cdot M_o^2 + \epsilon$$

where a non-zero β_5 weight would be evidence for quadratic moderation.

Mediation Analysis takes on a more traditional SEM form as a system of linear regressions, one for each model path in figrue 4b. For mediator M_e, coefficients $\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma_0, \gamma_1, \gamma_2$, and errors $\epsilon_1, \epsilon_2, \epsilon_3$,

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 X + \epsilon_1 \\ M_e &= \beta_0 + \beta_1 X + \epsilon_2 \\ Y &= \gamma_0 + \gamma_1 X + \gamma_2 M_e + \epsilon_3 \end{aligned} \tag{3}$$

Coefficient α_1 measures the total effect of X on Y. The goal of mediation analysis is to break apart this total effect into the "direct effect" (γ_1) and the "indirect effect" ($\beta_1 \cdot \gamma_2$) of X on Y through mediator M_e.

5 Results

5.1 Construct Validation: Factor Analysis

The causal analysis in the following section will consider two latent variables: EF and language ability. While there are 17 metrics for EF, there are only 2 for language ability. A factor with only 2 indicators is undersaturated and cannot yield enough degrees of freedom to calculate fit statistics. Given this limitation, this section pertains primarily to validating the factor model for EF.

As outlined in section 4, the first step is to consider the appropriateness of the data for an EFA. Figure 5 visualizes the Pearson correlations across the 17 EF metrics prior to any corrections – 41% of these correlations are above the $|\pm 0.30|$ threshold. A quick Bartlett test of sphericity comes back significant, rejecting the null hypothesis that the data were randomly generated, and KMO metrics on common versus unique variance reports an overall score of 0.72, which is considered acceptable. A breakdown of the KMO on each metric is visualized in figure 6. The main assessment of concern here is the PVT – 3 of its metrics scored near the unacceptable threshold of .5 and 1 scored an abysmal .13. Even with these results as baseline, there is enough evidence that EFA is appropriate and can be improved with adjustments.

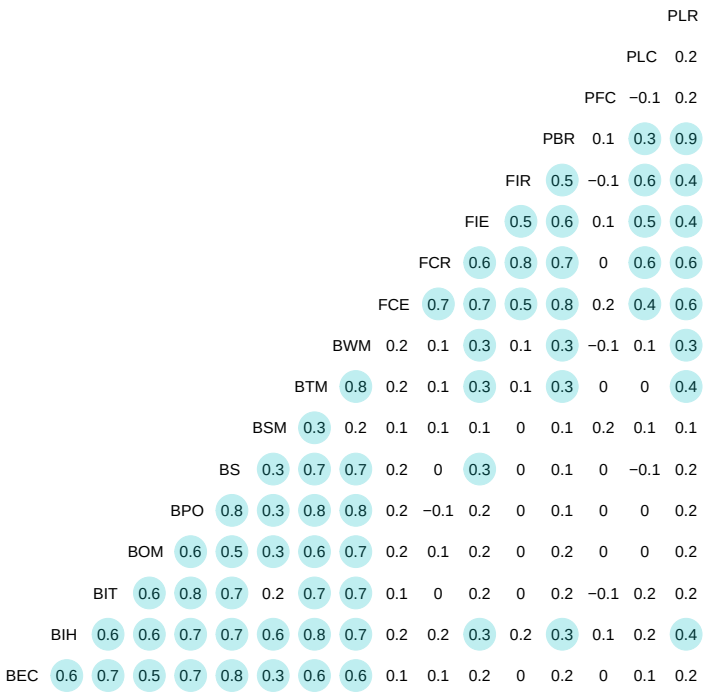


Figure 5: (Unadjusted) correlation matrix for raw data on 17 EF metrics. Correlations which meet the $|\pm 0.3|$ threshold are denoted with a shaded circle.

Data screening shows that several correlations are good candidates for adjustment. A subjective approach is to consider the scatter plots (Appendex A). Visual inspection shows a clear outlier in BSM. A quick investigation confirms that this point is likely to be a data collection error [2]. To supplement the scatter plots, Ramsey's RESET test identified 16 correlation pairs which are non-linear; calculations of skew and kurtosis identified 4 non-linear variables. In total, 63 Pearson's correlations should be replaced with Spearman's correlations, which are more robust to non-linearity and non-normality. Although parents completed Brief report using a Likert scale, which would make these variables good candidates for polychoric correlation, the final conversion of those responses into scores provide a large enough number of categories that the Brief variables can be treated as continuous.

Any missing data are grouped by individuals, indicating a child who did not participate in a given activity. Across the 3 activities, 7 children (8%) did not participate in the Brief survey, 13 (15%) in the Flanker task, and 15 (17%) in PVT. In the end, 5 imputed data sets were pooled to calculate a matrix of correlations with Spearman's correlations replacing the 63 instances mentioned. Another 5 data sets were imputed after treating the outlier identified above as missing. These too were pooled into a similar correlation matrix. These pooled correlation matrices (with (case 1) and without (case 2) the outlier replaced) can be seen in figure 7. Running the same tests as before, the Bartlett test was significance for both pooled correlation

[2] the outlier takes on a value of 48, which is 3 times higher than the penultimate value in that category, and almost 2 times higher than the penultimate value across the whole Brief test.

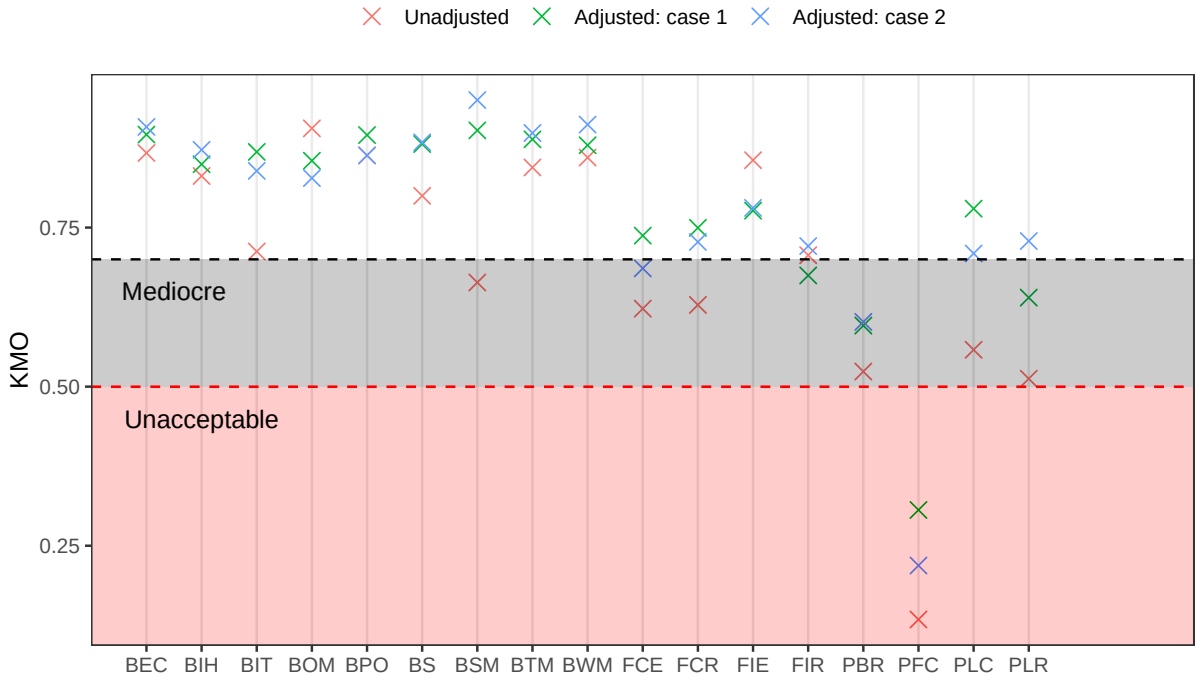
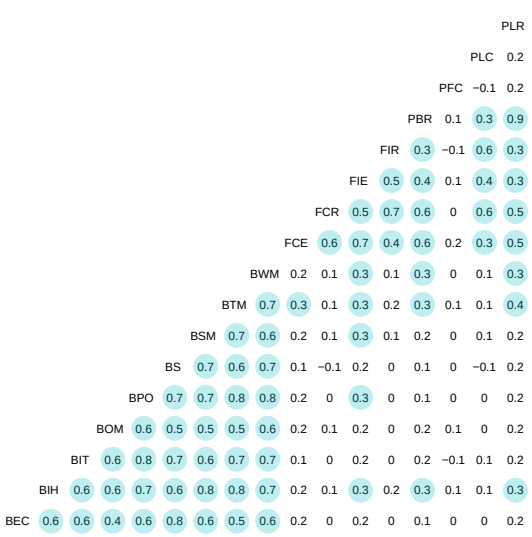


Figure 6: KMO metrics for unadjusted and adjusted (case 1 and case 2) correlation matrices. Unacceptable and mediocre regions are shaded in red and grey respectively. The desired region is left unshaded.

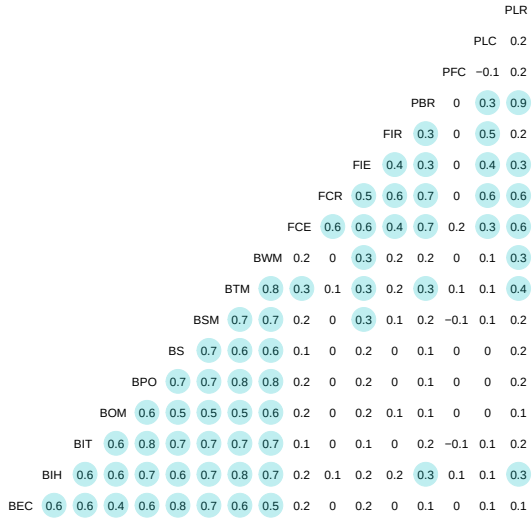
matrices and the KMO metrics can be seen in figure 6. The overall KMO scores were .81 and .82 for cases 1 and 2 respectively, a significant increase from the baseline value of .72. Examining the individual EF metrics (figure 6), while most changes were not large, there are several patterns of note. Multiple imputation alone was able to increase the KMO for BSM, regardless of the outlier. Of the 3 PVT metrics which were near the cutoff, the adjustments were able to improve performance on 2 (PLC, PLR), raising them closer to the desired .7 threshold. However, these same adjustments were not enough to improve the acceptability of PBR to this desirable level nor that of PFC, which remains well below .5.

Thus far, there are two reasons to drop PFC as an indicator for EF. Even after adjustment, all correlations (figure 7) for PFC remain below the $|\pm 0.30|$ threshold. Furthermore, the KMO metric for PFC indicates that much of the correlation in this metric is poorly explainable by any shared, underlying correlations apparent in other metrics. The PCA results reported in figure 8b confirm that dropping PFC could result in an additional 2% to 3% overall gain in proportion of variance explained. By dropping PFC, 64% of total variance would be explained by 2 components – the first component contributes 48%, with the second contributing the remaining 22%. Given that the first component explains less than 50% of the variance, these results encourage more than 1 underlying factor to explain the 16 remaining EF metrics. A subjective inspection of the scree plot (figure 8a) suggests that 3 components may be reasonable. The proportion of variance gained by this third component is 8% for both adjustment cases, bringing the cumulative variance proportion from the previous 64% to 72%. However, results from a Parallel Analysis show that the third component accounts for less variance than simulated random data, suggesting that 2 components may be enough. This conclusion is less clear when considering the MAP (figure 8c); the minima values occurs at both the second and third component. In this case, interpretation of the third component using PFA may help with the decision on whether to retain a third factor.

Given the small sample size of the data set, this report takes advice from Watkins (2018) and uses iterated Principal Axis at 1000 iterations for its PFA. An oblique rotation shows the first factor consistently loads heavily on all 9 Brief metrics – this is true whether 2 factors are assumed



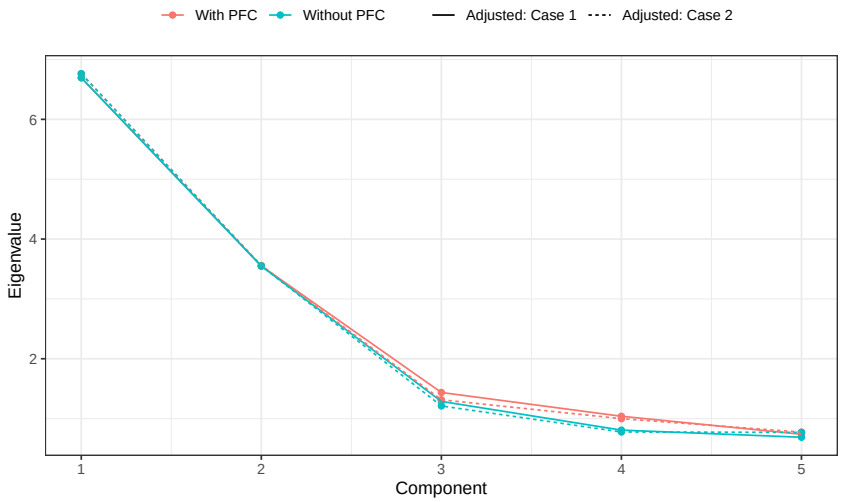
(a) Case 1: Pooled correlation matrix with Spearman adjustments.



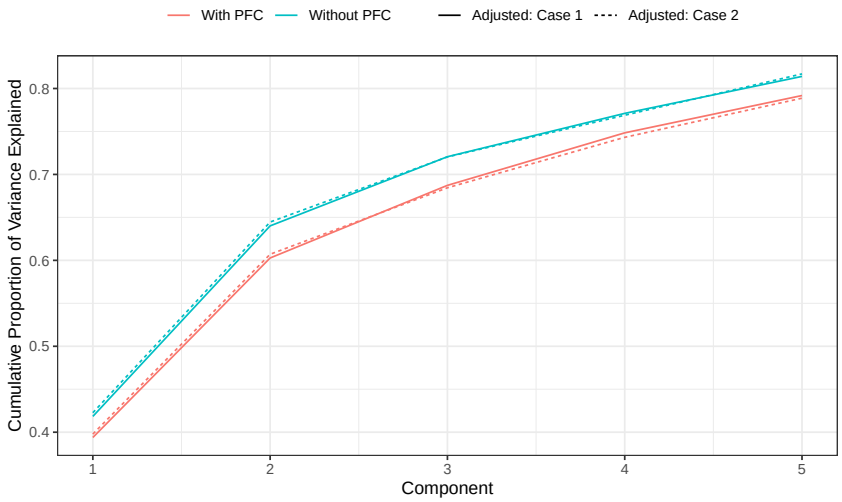
(b) Case 2: Pooled correlation matrix with Spearman adjustments and outlier replaced with imputation.

Figure 7: Correlation matrices for 17 EF metrics. Correlations which meet the .30 threshold are denoted with shaded circle.

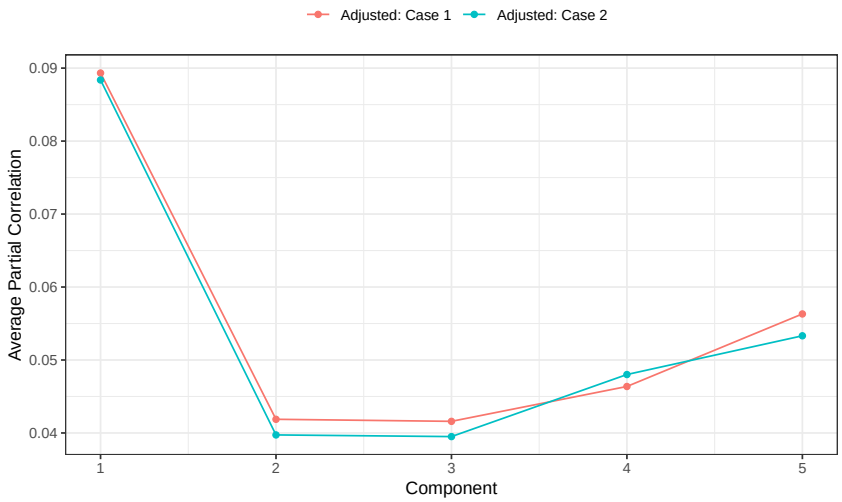
or 3, and remains true for both adjustment cases. Similarly, if 3 factors are assumed, the third factor consistently loads on both PVT reaction time metrics (PBR and PLR). Interpretation of the second factor is less clear (see tables 1 and 2). If only 2 factors are assumed, the second factor consistently loads heavily on all Flanker metrics as well as the 3 remaining PVT metrics. This factor could be interpreted to represent all directly assessed metrics. However, if the third factor is allowed, not only do PBR and PLR move to the third factor as mentioned, but the factor loading on FCE is split nearly in half, indicating that both the second and third factors would load on this metric comparably. Given that fit statistics can only be produced for oversaturated models, the results for the third factor would not allow for CFA validation on adjustment case 1. However, it may be enough to compare the fit on the second factor assuming 2 factors as well as 3.



(a) Scree plot for both adjustment cases, with and without PFC



(b) Cumulative proportion of explained variance for both adjustment cases, with and without PFC



(c) MAP values for adjustment cases, with and without PFC

Figure 8: PCA results for 16 versus 17 EF metrics indicate that PFC should be dropped.

	2 Factors		3 Factors		
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
BEC	0.725		0.728		
BIH	0.795		0.794		
BIT	0.831		0.826		
BOM	0.647		0.642		
BPO	0.902		0.907		
BS	0.845		0.848		
BSM	0.794		0.806		
BTM	0.801		0.787		
BWM	0.852		0.845		
FCE		0.706		0.490	0.306
FCR		0.926		0.812	
FIE		0.614		0.625	
FIR		0.693		0.821	
PBR		0.698			0.931
PLC		0.616		0.777	
PLR		0.564			0.928

Table 1: Factor loadings for adjustment case 1. Only loadings > .300 are shown.

	2 Factors		3 Factors		
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
BEC	0.740		0.746		
BIH	0.803		0.798		
BIT	0.853		0.847		
BOM	0.652		0.649		
BPO	0.912		0.912		
BS	0.817		0.823		
BSM	0.829		0.837		
BTM	0.826		0.816		
BWM	0.850		0.844		
FCE		0.745		0.449	0.393
FCR		0.932		0.707	0.328
FIE		0.564		0.633	
FIR		0.585		0.789	
PBR		0.766			0.941
PLC		0.572		0.704	
PLR		0.658			0.905

Table 2: Factor loadings for adjustment case 2. Only loadings > .300 are shown

This report conducts its CFA directly on z-scored data, again imputing 5 data sets for missing data. As before, the analysis is done twice, with and without the outlier removed. The fit statistics on individual factors (table 3) indicate that removing PBR and PLR from the second factor does not universally improve fit. For both adjustment cases, the removal of PBR and PLR minimizes information loss as indicated by the decrease in AIC as well as improves likelihood per the CFI. However, this improvement is at the expense of increased residuals (SRMR). Furthermore, it should be noted that the fit statistics for all scenarios under consideration for the second factor indicate poor fit – none of the statistics pass acceptability thresholds. By contrast, fit statistics on the first factor indicate much better fit, with the outlier removed (adjustment case 2) performing best with a CFI of .95, a SRMR below .05, and having a 5% lower AIC than adjustment case 1. These results could be further improved to achieve near perfect fit by removing the 2 metrics with the lowest loadings: BEC and BOM. Several more complex models were evaluated, the results for which can be seen in Appendex B, but none were able to achieve acceptability thresholds. Only the model for the first factor indicated by 7 Brief metrics (excluding BEC and BOM) with the outlier removed is able to surpass all acceptability thresholds.

	Number of indicators	Adjustment	CFI	SRMR	AIC
Factor 1	7	Case 1	0.90	**0.07**	1298.40
		Case 2	**0.99**	**0.03**	1231.79
	9	Case 1	0.88	**0.07**	1683.94
		Case 2	**0.95**	**0.04**	1620.84
Factor 2 (3 factors case)	4	Case 1	0.87	0.09	836.28
		Case 2	0.88	0.09	847.98
Factor 2 (2 factors case)	5	Case 1	0.88	**0.08**	1043.01
		Case 2	0.88	**0.08**	1049.41
	7	Case 1	0.00	0.14	1465.33
		Case 2	0.61	0.13	1453.07

Table 3: CFA fit statistics for single factor model. Statistics which pass acceptability threshold values are indicated by bolded font.

As analysis in this report will consider subset populations across gender and diagnosis, it is important to note that the fit statistics (table 4) show that this 7-metric model generally remains a good fit for data split by gender or diagnosis.

Number of indicators	Population	Adjustment	CFI	SRMR	AIC
7	All	Case 1	0.90	**0.07**	1298.40
		Case 2	**0.99**	**0.03**	1231.79
	Girls	Case 2	**0.98**	**0.04**	471.93
	Boys	Case 2	**0.95**	**0.04**	754.47
	ASD	Case 2	**1.00**	**0.05**	645.12
	TD	Case 2	0.91	**0.08**	879.25
9	All	Case 1	0.88	**0.07**	1683.94
		Case 2	**0.95**	**0.04**	1620.84
	Girls	Case 2	0.88	**0.05**	626.53
	Boys	Case 2	0.94	**0.05**	991.10
	ASD	Case 2	0.94	**0.08**	841.97
	TD	Case 2	0.81	0.10	1135.65

Table 4: CFA fit statistics for Factor 1, loading on 7 versus 9 Brief metrics across gender and diagnosis. Values which pass acceptability threshold are bolded.

To conclude, a combination of EFA and CFA has shown that the best construct on all 17 EF metrics is one which considers a single factor loading on 7 of the 9 Brief indicators with the outlier removed. While it is true that the Brief indicators cover a wider range of EF subcategories than the Flanker and PVT assessments, the nature of these indicators is such that this model would bias towards parent reportings, with no information gleamed from assessments done directly with the child. Unfortunately, complex models seeking to unite indicators from all three assessments show poor fit (Appendix B). In the absence of domain expertise, this author continues into casual analysis using the best fit model with the recognition that the results may be biased.

5.2 Causal analysis: Moderation and Mediation

Moderation Analysis

The moderation effect of bilingualism on EF based on the binary 'Diagnosis' variable amounts to fitting linear regression to two categories. Since moderation need not be linear, this report considers the following two moderation models:

$$\begin{aligned}\text{Model 1: } \text{EF} &= \alpha_1(\text{Diagnosis}) + \alpha_2(\text{YSB}) + \alpha_3(\text{YSB}) \times (\text{Diagnosis}) \\ \text{Model 2: } \text{EF} &= \beta_1(\text{Diagnosis}) + \beta_2(\text{YSB}) + \beta_3(\text{YSB}) \times (\text{Diagnosis}) + \beta_4(\text{YSB})^2 \\ &\quad + \beta_5(\text{YSB})^2 \times (\text{Diagnosis})\end{aligned}$$

Since diagnosis is binary, α_1 and β_1 represent the baseline EF difference between TD children and children with ASD. Pooled fit results on 9 imputed data sets show that while both YSB and Diagnosis separately predict EF performance, there does not appear to be a moderation affect, either linear or quadratic. Specifically, none of the estimates for α_3 (-0.083 ± 0.002), β_3 (-0.128 ± 0.066), nor β_5 (0.119 ± 0.161) reject the t-test null hypothesis at 5% significance at 96% power, indicating these coefficients are unlikely to be non-zero. Results from two Wald tests confirm this – there is no evidence to support either moderation model as a better fit for predicting EF than a baseline model which uses bilingualism and diagnosis as independent predictors (see Appendix C). These results cannot confirm moderation, either linear or quadratic, by bilingualism on EF based on a binary diagnosis. This holds true even when the same testing procedure is redone on data from boys and girls separately (see Appendix C).

The other variable related to ASD is SCQ, which is only measured for TD children. Since ASD is defined on a spectrum, this report supplements the above analysis by conducting a moderation analysis on TD children by replacing diagnosis with the child's SCQ score:

$$\begin{aligned}\text{Model 3: } \text{EF} &= \gamma_0 + \gamma_1(\text{SCQ}) + \gamma_2(\text{YSB}) + \gamma_3(\text{YSB}) \times (\text{SCQ}) \\ \text{Model 4: } \text{EF} &= \delta_0 + \delta_1(\text{SCQ}) + \delta_2(\text{YSB}) + \delta_3(\text{YSB}) \times (\text{SCQ}) + \delta_4(\text{YSB})^2 \\ &\quad + \delta_5(\text{YSB})^2 \times (\text{SCQ})\end{aligned}$$

where the intercepts γ_0 and δ_0 are added back in since the independent variable is no longer binary. Given that the sample size is smaller, it should be noted that the associated tests will be less powerful than in the above moderation analysis on the full population. To quantify how much smaller, a baseline model with no moderation term yields 15% power for a 5% significance level. Ultimately, this lack of power is not a main concern given that none of the coefficients in either model were able to reject the t-test null hypothesis at 5% significance. Naturally, as a result, any and all nested Wald tests were unable to reject the null hypothesis (see Appendix C). These conclusions also remained the same when considering data from boys and girls separately. As such, this report concludes its moderation analysis stating that there is no evidence to suggest a moderation affect from bilingualism on EF.

Mediation Analysis

As a reminder, the mediation effect under analysis considers if and how bilingualism mediates language ability on EF. As originally proposed by Baron & Kenny (1986), a mediation model must satisfy 3 specific conditions, with significance achieved at each step, before a mediation relationships can be established. The equations representing these 3 steps are:

$$\begin{aligned}
&\text{Step 1: } \text{EF} = \alpha_0 + \alpha_1(\text{Language}) \\
&\text{Step 2: } \text{YSB} = \beta_0 + \beta_1(\text{Language}) \\
&\text{Step 3: } \text{EF} = \gamma_0 + \gamma_1(\text{Language}) + \gamma_2(\text{YSB})
\end{aligned}$$

Traditionally, the first step is to confirm that the independent variable – here language ability – is a significant predictor of the dependent variable, EF. This is proven when $\alpha_1 \neq 0$. However, in recent practice, this requirement has been relaxed given that absence of correlation does not necessarily mean an absence of causality. In fact, for the current data set, results from this step do not support $\alpha_1 \neq 0$ at the 5% significance level for a t-test with only 26% power (Appendix C), thus this report invokes current practice and moves on to step two.

Results for the linear regression in step two does indeed yield a t-test which is able to reject the null hypothesis at the 5% level with 78% power, allowing a reasonable conclusion that $\beta_1 \neq 0$. This results remains true even when fitted to data for girls and children with ASD separately, though the power is lower ($\leq 50\%$) given the smaller sample sizes. However, when step two's linear regression is fitted to data from boys and TD children, the estimate for β_1 is no longer able to reject the t-test null hypothesis at the 5% significance level, indicating that language ability is not a good predictor of bilingualism in these two subsets (Appendix C). For the purpose of mediation analysis, this distinction may not matter as linear regression results from step three were unable to show that $\gamma_2 \neq 0$. This result remains true across the full population as well as when subset by gender or diagnosis (Appendix C). Ideally, a valid mediation analysis requires that $\gamma_2 \neq 0$ and $\gamma_1 < \alpha_1$. In the current case, neither of these conditions are achieved. This is further confirmed by results from the Wald test – no superset of data is able to achieve better fit with its covariates than the null model with zero covariates (Appendix C).

This 3-step method proposed by Baron & Kenny (1986) is increasingly falling out of favor. Thus for completeness, this report also considers the mediation model fitted by the 'lavaan' package. However, results also point to the same conclusion – there is no evidence of mediation between language ability and EF by bilingualism. Specifically, pooled results from an SEM mediation model fitted to 9 imputed data sets estimates the total effect to be $-0.186 \pm .110$. Assuming for the moment there is indeed a non-zero direct effect (which is suspect given that its 95% confidence interval $(-0.401, 0.030)$ contains 0), the indirect effect is estimated to be $.041 \pm .047$, where the standard error makes the effect essentially non-existent. These estimates only worsen (becoming smaller in magnitude) when subsetting by either gender or diagnosis. Thus, this report concludes its mediation analysis stating that there is no evidence to suggest that bilingualism mediates the relationship between language ability and EF.

6 Conclusions

This report's analysis began with a robust validation of the data set against the hypothesized underlying construct between 17 indicators and the latent EF variable. 1 indicator (PFC) was removed given its low correlation with the remaining 16 metrics. A principal component analysis (PCA) indicated that these 16 metrics metrics may in fact represent 2 or 3 latent factors. Rotated loading matrices were used to interpret these latent factors under Principal Factor Analysis (PFA) and results from both 2 and 3 factor models were reasonable. A Confirmatory Factor Analysis (CFA) showed that the fit for most of the models under consideration were not good approximations, with the exception of a single factor loading on 7 Brief indicators. Though this model would yield highly biased results, the author of this report continued on to causal

analysis using this model with the hope that a more accordant model would yield more reliable results. Unfortunately, this report concluded its causal analysis with no significant findings of bilingualism either moderating or mediating executive functions in children across gender and diagnosis.

The author wishes to conclude this report by making explicit the conservative stance she took in her analysis work, especially in the interpretation of CFA results and her decision to consider only the best fit model when conducting casual analysis. She recognizes that this may not be a successful strategy in the application of psychometrics, particularly in exploratory work such as this. As such, there remains much to be explored if CFA fit tolerances were lowered and more latent variables were considered in the causal analysis. Furthermore, given that linear regressions are the basis of causal analysis, there remain many covariates not considered in this report which may prove to be good predictors or moderators of bilingualism and executive functions.

References

Abad, V. C. & Guilleminault, C. (2012), *Chapter 33 - Polysomnographic Evaluation of Sleep Disorders*, W.B. Saunders, London, pp. 727–762.
URL: *http://www.sciencedirect.com/science/article/pii/B9781455703081000339*

Acito, F. & Anderson, R. D. (1986), 'A simulation study of factor score indeterminacy', *Journal of Marketing Research* **23**(2), 111–118.
URL: *http://www.jstor.org/stable/3151658*

Baron, R. M. & Kenny, D. A. (1986), 'The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.', *Journal of personality and social psychology* **51**(6), 1173.

Beauchamp, M. L. & MacLeod, A. A. (2017), 'Bilingualism in children with autism spectrum disorder: Making evidence based recommendations.', *Canadian Psychology/psychologie canadienne* **58**(3), 250.

Carter, A. S., Black, D. O., Tewani, S., Connolly, C. E., Kadlec, M. B. & Tager-Flusberg, H. (2007), 'Sex differences in toddlers with autism spectrum disorders', *Journal of autism and developmental disorders* **37**(1), 86–97.

Dworzynski, Katharina, R. A. B. P. & Happé, F. (2012), 'How different are girls and boys above and below the diagnostic threshold for autism spectrum disorders?', *Journal of the American Academy of Child Adolescent Psychiatry* **51**(8), 788 – 797.
URL: *http://www.sciencedirect.com/science/article/pii/S0890856712004121*

Goldstein, S., Naglieri, J. A., Princiotta, D. & Otero, T. M. (2014), Introduction: A history of executive functioning as a theoretical and clinical construct, *in* 'Handbook of executive functioning', Springer, pp. 3–12.

Hampton, S., Rabagliati, H., Sorace, A. & Fletcher-Watson, S. (2017), 'Autism and bilingualism: A qualitative interview study of parents' perspectives and experiences', *Journal of Speech, Language, and Hearing Research* **60**(2), 435–446.

Hartley, S. L. & Sikora, D. M. (2009), 'Sex differences in autism spectrum disorder: an examination of developmental functioning, autistic symptoms, and coexisting behavior problems in toddlers', *Journal of autism and developmental disorders* **39**(12), 1715.

Hu, L. & Bentler, P. M. (1999), 'Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives', *Structural Equation Modeling: A Multidisciplinary Journal* **6**(1), 1–55.
URL: *https://doi.org/10.1080/10705519909540118*

Nerbonne, J. (2016), 'Exploratory factor analysis'. URL last visited on 29/7/2020.
URL: *https://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/Factor-Analysis-Kootstra-04.PDF*

Rigdon, E. E., Becker, J.-M. & Sarstedt, M. (2019), 'Factor indeterminacy as metrological uncertainty: implications for advancing psychological measurement', *Multivariate behavioral research* **54**(3), 429–443.

Urbano, Lorenzo-Seva, T. M. E. & Kiers, H. A. L. (2011), 'The hull method for selecting the number of common factors', *Multivariate Behavioral Research* **46**(2), 340–364. PMID: 26741331.
URL: *https://doi.org/10.1080/00273171.2011.564527*

Watkins, M. W. (2018), 'Exploratory factor analysis: A guide to best practice', *Journal of Black Psychology* **44**(3), 219–246.
URL: *https://doi.org/10.1177/0095798418771807*

Appendices

All code used for this report may be found at: https://github.com/eshaw2/BilingualChildren

A Appendix 1

Scatter plot matrix for raw data across 17 EF metrics from section 5.1

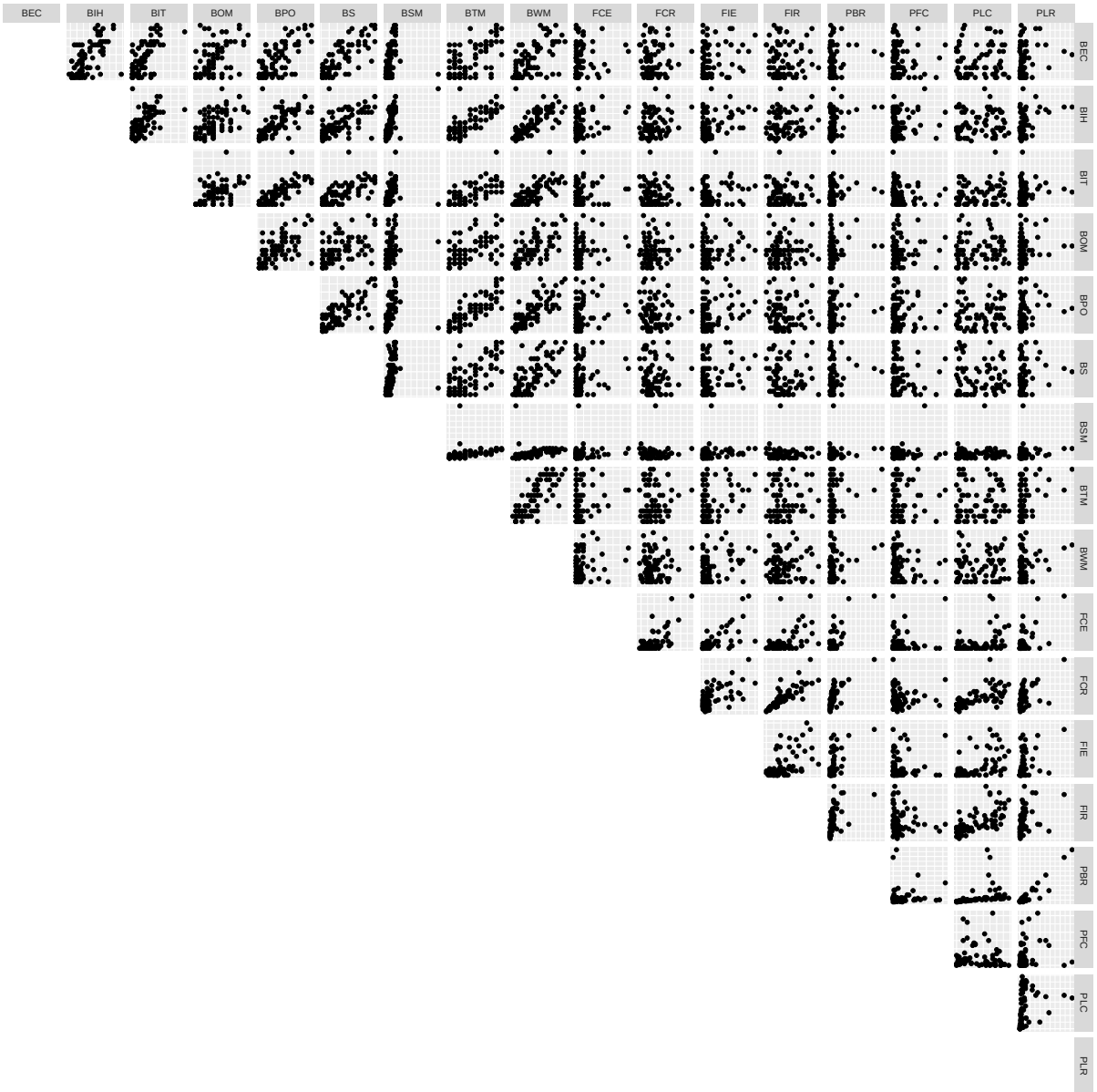


Figure 9: Scatter plot matrix of unadjusted data.

B Appendix 2

CFA fit statistics for complex multi-factor models.

B.1 Correlated factor models

B.1.1 2 correlated factors

Model 1:

$$\text{corr}(f1, f2) = c_1$$
$$f1 = a_1(\text{BPO}) + a_2(\text{BS}) + a_3(\text{BWM}) + a_4(\text{BIT}) + a_5(\text{BSM}) + a_6(\text{BIH}) + a_7(\text{BTM})$$
$$\text{Model 1a: } f2 = b_1(\text{FCE}) + b_2(\text{FCR}) + b_3(\text{FIR}) + b_4(\text{FIE}) + b_5(\text{PLC})$$
$$\text{Model 1b: } f2 = b_1(\text{FCE}) + b_2(\text{FCR}) + b_3(\text{FIR}) + b_4(\text{FIE}) + b_5(\text{PLC}) + b_6(\text{PBR}) + b_7(\text{PLR})$$

Fit statistics:

	Population	Outlier	CFI	SRMR	AIC
Model 1a	All	Not removed	0.89	0.10	2341.58
		Removed	**0.95**	0.09	2281.23
	Girls	Removed	**1.00**	0.14	1318.05
	Boys	Removed	0.82	0.10	1797.34
	ASD	Removed	**1.00**	0.10	1114.00
	TS	Removed	0.89	0.10	1381.73
Model 1b	All	Not removed	0.65	0.12	2761.82
		Removed	0.80	0.10	2683.59
	Girls	Removed	**1.00**	0.14	1318.05
	Boys	Removed	0.82	0.10	1797.34
	ASD	Removed	0.93	0.12	1443.13
	TS	Removed	0.69	0.12	2003.24

Table 5: Statistics which surpass acceptability thresholds are bolded.

B.1.2 3 correlated factors

Model 2:

$$\begin{aligned}
\text{corr}(f1, f2) &= d_1 \\
\text{corr}(f1, f3) &= d_2 \\
\text{corr}(f2, f3) &= d_3 \\
f1 &= a_1(\text{BPO}) + a_2(\text{BS}) + a_3(\text{BWM}) + a_4(\text{BIT}) + a_5(\text{BSM}) + a_6(\text{BIH}) + a_7(\text{BTM}) \\
f2 &= b_1(\text{FCR}) + b_2(\text{FCE}) + b_3(\text{FIR}) + b_4(\text{FIE}) + b_5(\text{PLC}) \\
\text{Model 2a: } f3 &= c_1(\text{PBR}) + c_2(\text{PLR}) + c_3(\text{FCE}) \\
\text{Model 2b: } f3 &= c_1(\text{PBR}) + c_2(\text{PLR}) + c_3(\text{FCE}) + c_4(\text{FCR})
\end{aligned}$$

Fit statistics:

	Population	Outlier	CFI	SRMR	AIC
Model 2a	All	Not removed	0.89	0.10	2601.35
		Removed	**0.95**	0.10	2535.16
	Girls	Removed	1.00	0.11	1152.27
	Boys	Removed	0.90	0.11	1509.65
	ASD	Removed	**1.00**	0.11	1206.84
	TS	Removed	0.81	0.12	1692.23
Model 2b	All	Not removed	0.91	0.10	2584.02
		Removed	**0.96**	0.10	2521.35
	Girls	Removed	**1.00**	0.11	1339.89
	Boys	Removed	0.91	0.11	1503.00
	ASD	Removed	**1.00**	0.10	1200.80
	TS	Removed	0.84	0.12	1685.35

Table 6: Statistics which surpass acceptability thresholds are bolded.

B.2 Second order factor models

B.2.1 2nd order with 2 factors

Model 3:

$$f1 = a_1(\text{BPO}) + a_2(\text{BS}) + a_3(\text{BWM}) + a_4(\text{BIT}) + a_5(\text{BSM}) + a_6(\text{BIH}) + a_7(\text{BTM})$$

Model 3a: $f2 = b_1(\text{FCE}) + b_2(\text{FCR}) + b_3(\text{FIR}) + b_4(\text{FIE}) + b_5(\text{PLC})$

Model 3b: $f2 = b_1(\text{FCE}) + b_2(\text{FCR}) + b_3(\text{FIR}) + b_4(\text{FIE}) + b_5(\text{PLC}) + b_6(\text{PBR}) + b_7(\text{PLR})$

$$f4 = f1 + f2$$

Fit statistics:

	Population	Outlier	CFI	SRMR	AIC
Model 3a	All	Not removed	0.88	0.20	2351.24
		Removed	0.94	0.21	2290.71
	Girls	Removed	**0.95**	0.38	963.39
	Boys	Removed	0.90	0.11	1356.19
	ASD	Removed	**1.00**	0.16	1115.17
	TS	Removed	0.85	0.27	1062.88
Model 3b	All	Not removed	0.67	0.17	2766.15
		Removed	0.79	0.17	2689.41
	Girls	Removed	**1.00**	0.29	1196.53
	Boys	Removed	0.80	0.13	1561.34
	ASD	Removed	0.90	0.18	1307.67
	TS	Removed	0.68	0.25	1152.55

Table 7: Statistics which surpass acceptability thresholds are bolded.

B.2.2 2nd order with 3 factors

Model 4:

$$f1 = a_1(\text{BPO}) + a_2(\text{BS}) + a_3(\text{BWM}) + a_4(\text{BIT}) + a_5(\text{BSM}) + a_6(\text{BIH}) + a_7(\text{BTM})$$
$$f2 = b_1(\text{FCR}) + b_2(\text{FCE}) + b_3(\text{FIR}) + b_4(\text{FIE}) + b_5(\text{PLC})$$

Model 4a: $f3 = c_1(\text{PBR}) + c_2(\text{PLR}) + c_3(\text{FCE})$

Model 4b: $f3 = c_1(\text{PBR}) + c_2(\text{PLR}) + c_3(\text{FCE}) + c_4(\text{FCR})$

$$f4 = f1 + f2 + f3$$

Fit statistics:

	Population	Outlier	CFI	SRMR	AIC
Model 4a	All	Not removed	0.88	0.21	2617.31
		Removed	0.93	0.21	2558.45
	Girls	Removed	1.00	0.34	1146.98
	Boys	Removed	0.87	0.15	1527.79
	ASD	Removed	**1.00**	0.21	1259.82
	TS	Removed	0.77	0.28	1117.00
Model 4b	All	Not removed	0.90	0.25	2595.37
		Removed	**0.95**	0.25	2534.10
	Girls	Removed	**1.00**	0.32	1144.27
	Boys	Removed	0.90	0.17	1506.27
	ASD	Removed	**1.00**	0.22	1202.60
	TS	Removed	0.77	0.29	1725.06

Table 8: Statistics which surpass acceptability thresholds are bolded.

C Appendix 3

Causal analysis linear regression statistics for various population subsets.

C.1 Moderation analysis results

C.1.1 ASD by Diagnosis

$$\begin{aligned}
\text{Null model: } & \text{EF} = a_1 \\
\text{Baseline model: } & \text{EF} = a_1(\text{Diagnosis}) + a_2(\text{YSB}) \\
\text{Model 1: } & \text{EF} = a_1(\text{Diagnosis}) + a_2(\text{YSB}) + a_3(\text{YSB}) \times (\text{Diagnosis}) \\
\text{Model 2: } & \text{EF} = a_1(\text{Diagnosis}) + a_2(\text{YSB}) + a_3(\text{YSB}) \times (\text{Diagnosis}) + a_4(\text{YSB})^2 \\
& \quad + a_5(\text{YSB})^2 \times (\text{Diagnosis})
\end{aligned}$$

Baseline regression results:

$$\text{Baseline model: } \text{EF} = a_1(\text{Diagnosis}) + a_2(\text{YSB})$$

Results:

Population	Sample size	a_1 estimate	t-statistic	p-value	a_2 estimate	t-statistic	p-value	Power
All	89	0.086	4.041	**0.000**	-1.195	-6.431	**0.000**	1.000
Girls	37	0.114	3.773	**0.001**	-1.601	-6.030	**0.000**	1.000
Boys	52	0.075	2.562	**0.017**	-0.975	-4.088	**0.001**	0.990

Wald test: Baseline vs. Null

$$\begin{aligned}
H_0 &: a_1 = a_2 = 0 \\
H_A &: a_1 \neq a_2 \neq 0
\end{aligned}$$

Results:

Full Model	Nested Model	Population	F-statistic	p-value
Baseline	Null	All	23.372	**0.000**
		Girls	19.365	**0.000**
		Boys	9.552	**0.001**

Table 9: Wald test of Baseline non-moderation model against Null model. Significant p-values under .05 are bolded for visual ease.

Interpretation: Both YSB and Diagnosis are good predictors of EF. Baseline Model is a good baseline against which to compare moderation models. This is instead of comparing against the Null model, which will always come back significant for either moderation models.

Moderation (linear regression) results:

	Population	Sample size	a_3 estimate	t-statistic	p-value	a_5 estimate	t-statistic	p-value	Power
Model 1	All	89	-0.083	-1.508	0.142				0.957
	Girls	37	-0.101	-1.433	0.162				0.758
	Boys	52	-0.042	-0.506	0.619				0.713
Model 2	All	89	-0.128	-1.942	0.058	0.119	0.741	0.463	0.965
	Girls	37	-0.139	-1.263	0.219	0.102	0.343	0.735	0.695
	Boys	52	-0.082	-0.864	0.394	0.076	0.394	0.696	0.724

Table 10: Pooled moderation results. Significant p-values under .05 for moderation effect are bolded for visual ease (there are none).

Wald test: Moderation Models vs. Baseline

$$H_0 : a_3 = a_4 = a_5 = 0$$
$$H_A : a_3 \neq a_4 \neq a_5 \neq 0$$

Results:

Full Model	Nested Model	Population	F-statistic	p-value
Model 1	Baseline	All	2.273	0.205
		Girls	2.053	0.225
		Boys	0.256	0.639
Model 2	Baseline	All	1.349	0.268
		Girls	0.693	0.564
		Boys	0.392	0.759
Model 2	Model 1	All	0.825	0.223
		Girls	0.123	0.885
		Boys	0.483	0.620

Table 11: Wald test of moderation models against Baseline model. Significant p-values under .05 are bolded for visual ease (there are none).

C.1.2 ASD by SCQ

$$\begin{aligned}
\text{Null model: } & \text{EF} = b_0 \\
\text{Baseline: } & \text{EF} = b_0 + b_1(\text{SCQ}) + b_2(\text{YSB}) \\
\text{Model 3: } & \text{EF} = b_0 + b_1(\text{SCQ}) + b_2(\text{YSB}) + b_3(\text{YSB}) \times (\text{SCQ}) \\
\text{Model 4: } & \text{EF} = b_0 + b_1(\text{SCQ}) + b_2(\text{YSB}) + b_3(\text{YSB}) \times (\text{SCQ}) + b_4(\text{YSB})^2 \\
& \qquad + b_5(\text{YSB})^2 \times (\text{SCQ})
\end{aligned}$$

Baseline regression:

$$\text{Baseline model: } \text{EF} = b_0 + b_1(\text{SCQ}) + b_2(\text{YSB})$$

Results:

Population	Sample size	b_1 estimate	t-statistic	p-value	b_2 estimate4	t-statistic	p-value	Power
All	51	0.005	0.115	0.909	-0.012	-0.264	0.793	0.149
Girls	21	0.036	0.598	0.558	0.041	0.632	0.537	0.030
Boys	30	-0.011	-0.215	0.832	-0.055	-0.876	0.390	0.101

Table 12: Pooled moderation results. Significant p-values under .05 for moderation effect are bolded for visual ease (there are none).

Wald test: Baseline vs. Null

$$\begin{aligned}
H_0 &: b_1 = b_2 = 0 \\
H_A &: b_1 \neq b_2 \neq 0
\end{aligned}$$

Results:

Full Model	Nested Model	Population	F-statistic	p-value
Baseline	Null	All	0.044	0.957
		Girls	0.354	0.707
		Boys	0.387	0.683

Table 13: Wald test of Baseline non-moderation model against Null model for TD children. Significant p-values under .05 are bolded for visual ease (there are none).

Moderation Analysis on TD children results:

	Population	Sample size	b_3 estimate	t-statistic	p-value	b_5 estimate	t-statistic	p-value	Power
Model 3	All	51	0.013	0.423	0.676				0.139
	Girls	21	0.028	0.669	0.514				0.079
	Boys	30	0.004	0.091	0.929				0.097
Model 4	All	51	-0.042	-0.270	0.790	0.006	0.367	0.717	0.163
	Girls	21	-0.670	-1.922	0.084	0.089	2.022	0.070	0.087
	Boys	30	0.010	0.047	0.964	-0.001	-0.031	0.976	0.108

Table 14: Pooled moderation results on TD children. Significant p-values under .05 for moderation effect are bolded for visual ease (there are none).

Wald test: Moderation Models vs. Null

$$H_0 : b_1 = b_2 = b_3 = b_4 = b_5 = 0$$
$$H_A : b_1 \neq b_2 \neq b_3 \neq b_4 \neq b_5 \neq 0$$

Results:

Full Model	Nested Model	Population	F-statistic	p-value
Model 3	Null	All	0.093	0.963
		Girls	0.379	0.769
		Boys	0.239	0.869
Model 4	Null	All	0.142	0.981
		Girls	1.212	0.358
		Boys	0.138	0.982

Table 15: Wald test of moderation models against Null model for TD children. Significant p-values under .05 are bolded for visual ease (there are none).

C.2 Mediation analysis results

$$\begin{aligned}
\text{Null Model: } & \text{EF} = c_0 \\
\text{Step 1: } & \text{EF} = c_0 + c_1(\text{Language}) \\
\text{Step 2: } & \text{YSB} = c_0 + c_1(\text{Language}) \\
\text{Step 3: } & \text{EF} = c_0 + c_1(\text{Language}) + c_2(\text{YSB})
\end{aligned}$$

3-step mediation analysis results:

	Population	Sample size	c_1 estimate	t-statistic	p-value	c_2 estimate	t-statistic	p-value	Power
Step 1	All	89	-0.177	-1.405	0.170				0.263
	Girls	37	-0.090	-0.468	0.644				0.068
	Boys	52	-0.236	-1.320	0.212				0.315
	ASD	38	-0.087	-0.525	0.610				0.109
	TD	51	-0.161	-1.521	0.135				0.236
Step 2	All	89	-0.875	-2.533	**0.021**				0.776
	Girls	37	-0.985	-2.110	**0.046**				0.451
	Boys	52	-0.773	-1.833	0.081				0.410
	ASD	38	-1.022	-2.468	**0.019**				0.503
	TD	51	-0.751	-1.380	0.203				0.381
Step 3	All	89	-0.189	-1.446	0.157	-0.014	-0.328	0.744	0.274
	Girls	37	-0.158	-0.770	0.448	-0.069	-0.938	0.355	0.157
	Boys	52	-0.217	-1.198	0.251	0.025	0.462	0.646	0.269
	ASD	38	-0.079	-0.472	0.644	0.008	0.138	0.891	0.096
	TD	51	-0.169	-1.513	0.138	-0.010	-0.251	0.803	0.251

Table 16: Pooled mediation results from 3-step linear regressions from Baron & Kenny (1986). Significant p-values under .05 are bolded for visual ease.

Wald test: Step i model vs. Null model:

$$\begin{aligned}
H_0 &: c_1 = c_2 = 0 \\
H_A &: c_1 \neq c_2 \neq 0
\end{aligned}$$

Full Model	Nested Model	Population	F-statistic	p-value
Step 1	Null	All	1.975	0.233
		Girls	0.219	0.664
		Boys	1.743	0.257
		ASD	0.276	0.627
		TD	2.314	0.203
Step 2	Null	All	6.417	0.064
		Girls	4.454	0.102
		Boys	3.361	0.141
		ASD	6.091	0.069
		TD	1.903	0.240
Step 3	Null	All	1.164	0.321
		Girls	0.539	0.588
		Boys	1.181	0.326
		ASD	0.171	0.844
		TD	1.149	0.327

Table 17: Wald test of models from 3-step linear regressions against Null model. Significant p-values under .05 are bolded for visual ease (there are none).