E-Risk Study Concept Paper template

Provisional Paper Title: The Role of Early Language Experiences in The Transmission of Family Background Inequality.
Proposing Author: Anna Brown, Prof. Sophie von Stumm
Author's Email: anna.brown2@york.ac.uk; sophie.vonstumm@york.ac.uk
Academic Supervisor: NA (if the proposing author is a student)
E-Risk Sponsor: Helen Fisher (if the proposing author is not an E-Risk co-investigator)
Today's Date: 5/1/2023
Please indicate if you will require an E-Risk independent reproducibility check: ☐

Please describe your proposal in 2-3 pages with sufficient detail for helpful review.

Background & objective of the study:

On average, children from families with fewer resources obtain fewer educational qualifications and are at greater risk of cognitive impairment than children from better-off families. A key pathway for the transmission of this family background inequality could be children's early life language experiences (Kidd et al., 2018). Parents of lower socioeconomic backgrounds tend to speak less often to their children, employ less sophisticated vocabulary, and use a small range of simpler syntactic structures than parents of higher socioeconomic status (SES). Thus, SES correlates with the quality and quantity of the language children are exposed to in their family home (Rowe, 2018). As a result of this SES-related discrepancy in language experiences, children from lower SES backgrounds are thought to develop poorer language skills themselves. Maternal speech affects children's language growth; children who hear longer sentences with a wider vocabulary build their productive vocabularies at a faster rate (Hoff, 2003). Thus, children from low SES backgrounds are at risk for poor cognitive and academic outcomes.

Early life language development plays a vital role in the success of later life outcomes. Sophisticated language is required to develop essential literacy skills (Hulme et al., 2012), further develop cognitive skills (Perszyk & Waxman, 2018), perform well in school and gain access to further education. SES strongly predicts a person's academic performance. It has recently been shown that about half of the influence of SES on later school achievement is mediated by children's early life language ability (von Stumm et al., 2020).

Previous studies in this area often relied on proxy measures to capture children's early life language experiences. In the current study, children's early life language environments will be extracted from recordings of naturalistic samples of their mother's speech. We will take naturalistic speech samples from the E-Risk mothers and extract two markers for the quality of their lexicon (i.e., lexical diversity score and the number of rare words used) and their grammar (i.e., mean length of utterance and clausal diversity). We will focus on grammar and lexicon because variability within these language characteristics describes the children's early life language experiences and is systematically related to families' SES differences (d'Apice et al., 2019; Huttenlocher et al., 2010; Rowe, 2008). We will test how strongly each language marker is associated with SES indicators of education, occupation, and income and how well they predict important developmental outcomes such as non-verbal cognitive ability, reading ability, and academic performance.

To develop a clearer understanding of the mechanism by which early life language experiences relate to family background inequality, we will address three central research questions (RQ):

RQ1: To what extent do markers of family SES, including mothers' and fathers' education, their occupation, and household income, predict children's early life language environments, as indexed by the mother's grammar and lexicon?

RQ2: Do children's early life language environments predict children's cognitive development, including their language abilities (i.e., vocabulary), intelligence, reading ability, and academic performance?

We will test whether children's language environments predict their cognitive development concurrently and over time. We will first test whether children's language environments, extracted from recordings of mothers taken when their children were 5 years old, predict their concurrent vocabulary and intelligence and their later reading ability at age 7 and 10 years. We will also test if early life language environments predict academic outcomes over time, specifically how the language environments relate to teacher ratings of academic performance in English and Maths at the ages 7, 10, and 12 years. We will also test to what extent the prediction of children's cognitive development from their early life language experiences will be attenuated by markers of SES.

RQ3: To what extent are associations between children's family background and their cognitive development, including their language abilities (i.e., vocabulary), intelligence, reading ability, and academic performance, mediated by early life language environments?

Here, we will test the extent to which the grammar and lexicon of early life language environments mediate the association between family background and a child's non-verbal cognitive ability, reading ability, and academic performance.

Significance of the study (for theory, research methods, or clinical practice):

This research will (a) comprehensively quantify children's early life language environments from naturalistic speech samples and (b) empirically demonstrate how early life environments mediate the link between family inequality and children's developmental outcomes. Results from this approach will be able to suggest how much the effect of inequality and developmental outcomes are transmitted through the language environment.

Understanding the extent to which language experiences contribute to the pervasive long-term influences of family background on child development will clarify the suitability of children's language experiences as a target for future interventions. Language experiences are malleable and, therefore, potentially valuable targets for interventions. This research will provide an important evidence base for researchers to include in their design of interventions, as well as define the upper limit of effectiveness that a language intervention can have in reducing inequality.

Data analysis methods:

Preliminary analyses

Prior to including the extracted language markers in any analysis, we will validate the language markers across and within transcripts. We will calculate intraclass coefficients between the mother's speech about the elder twin and the mother's speech about the younger twin. We shall also calculate correlations between language markers and mothers' wide-range achievement test scores (Snelbaker et al., 2001).

Main Analyses

We will use hierarchical linear regression models and latent growth curve models (LGCM) to assess and compare how early life language environments at age 5, as indicated by language

markers from the mother's speech, predict non-verbal cognitive score, vocabulary score at age 5, and reading ability at ages 7 and 10, as well as educational achievement at ages 7, 10, and 12. Non-verbal cognitive score and vocabulary are measured by the Wechsler Preschool & Primary Scale of Intelligence (WPPSI), and reading ability by the Test of Word Reading Efficiency (TOWRE). Educational outcomes are Teacher's ratings of English and Maths school performance.

We will use the Lavaan package in R (Rosseel, 2012) to compute our models. This allows using full information maximum likelihood (FIML) estimation to deal with missing data, which we expect to be missing at random. No recordings of mothers' speech are available for about 15% of the E-Risk families. To account for non-independence and the correlation of error terms within twin pairs, regression error terms will be clustered at the family level for all models.

From transcripts of mothers' speech recordings, we will extract four language markers that map onto two language characteristics. From each speech sample, we will compute two lexical markers (Lexical Diversity score; the number of rare words) and two grammar markers (Mean Length of Utterance; Clausal diversity*)*. Next, we will test the markers' inter-correlations, predicting that the two grammar markers and the two lexicon markers will correlate above .60, respectively, but less strongly across marker constructs. In this case, we will create composites for grammar and lexicon from their respective markers. In case markers correlate above .60 across constructs (i.e., grammar and lexicon), we will produce one composite from all four markers to reflect children's early life language experiences. If the correlations between language markers are below .60, we will treat them as individual predictors in our models.

To measure family SES, we will include mothers' and fathers' highest educational qualifications, mothers' and fathers' occupations, and overall household income. These SES indicators will be added to the models as individual predictors, as these components of SES may affect the language environment differently and independently.

***RQ1:* To what extent does family SES predict children's early life language environments?**

We will fit a regression model with the language markers extracted from naturalistic speech samples as our outcomes and our SES indicators (mother and father's highest educational qualification, occupation, and combined household income) as predictors. We will first construct a correlation matrix to assess the correlations between all our language outcomes and SES indicators. Then we will construct a multiple regression model with all SES indicators included as individual predictors (model 1). We will interpret beta weights and p values to judge if individual SES predictors make significant, independent contributions

In the next step, we will adjust our models for co-variates, including mothers' age and their IQ as indicated by performance on the Wide Range Achievement Test, as these may confound associations between our SES indicators and language markers (model 2). We will then use the change in R^2 to gauge how much variance is now accounted for and interpret changes in the SES predictor's beta weights between this and the previous model to determine whether the co-variates adjusted our SES indicators' effect on language environments.

RQ2: To what extent do children's early life language environments predict their cognitive development?

We will fit a series of independent hierarchical regression models for our cognitive outcomes (nonverbal cognitive skills and vocabulary at age 5, reading ability at ages 7 and 10). In our baseline model, we will add our language markers as predictors (model 3). We will use the beta weights to assess which language characteristics best predict cognitive outcomes. In the next step, we will add any SES indicators that significantly predicted our language environments in model 1 (model 4). We will then use the change in R^2 to gauge how much variance is now accounted for, and we shall use beta weights to determine if including SES indicators changed the effect our language markers had on cognitive outcomes. Finally, our model will be adjusted for the mother's

marital status, no. of children in the household, and the mother's age, as they may confound any associations (model 5). Once again, we shall use the change in R^2 to gauge how much variance is now accounted for and use beta weights to determine if these co-variates adjust any of our observed effects between language environments, SES indicators and family background.

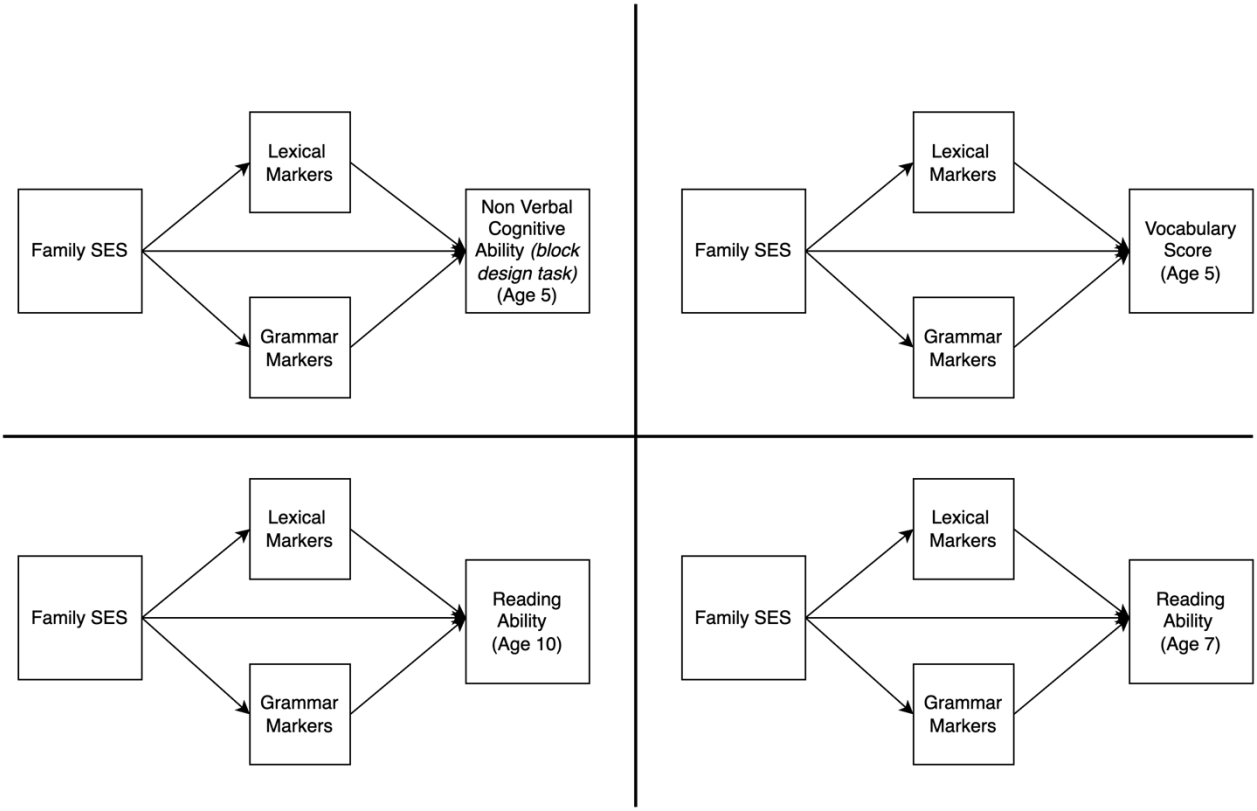

Figure 1. Hypothesised models for the relationship between Family SES, Language markers, and cognitive outcomes.

To test the relationship between a child's early language environment and educational outcomes, we will fit latent growth curve models to school performance, as rated by teachers, at ages 7, 10 and 12 years. Latent growth curve models differentiate individual differences in a construct, here school performance, that are stable across assessments or over time, and systematic differences in the rate of change or growth in school performance. Latent growth curve models require at least three assessment time points to be fitted. They will enable testing whether early life language experiences predict individual differences in school performance that are stable from age 7 through to 12 years and whether they predict gains or losses in school performance over time.

Two latent variables will be created. A latent intercept variable will represent individual differences in academic performance that are stable from age 7 through to 12 years by constraining all factor loadings from the observed age 7, age 10, and age 12 variables to be equal to 1. The second latent factor will be the slope of change, which captures systematic individual differences in linear change (i.e., growth or decline) in academic performance over time. The latent slope will be estimated by constraining the factor loadings from the age 7, age 10, and age 12 variables to reflect the distance between assessments in years (i.e., 0, 3, 5, respectively; model 10). We will also explore extracting a third latent growth factor, a quadratic term, whose loadings are specified as the square of the slope's loadings (i.e., 0, 9, and 25; model 11). Model comparisons will be calculated to determine the best-fitting model, and we will examine the significance of the variances of the latent factors. We will retain the model solution that fits best and includes latent factors with significant variance.

We will then follow the same procedure as the analysis for cognitive outcomes. First, the language markers will be added as a predictor to the latent growth curve model (model 12). We will use the beta weights to asses if early life language experiences predict the latent slope factors. In the next step, our SES indicators will also be added to the model as predictors. Only SES indicators were significant predictors in model 1 from the initial analysis will be added (model 13). The beta weights will be used to assess if SES predicts initial academic performance (intercept) and the rate of change in academic performance (slope), and whether it changes the effect of our language markers. Finally, our model will be adjusted for the mother's marital status, no. of children in the household, and the mother's age, as they may confound any associations (model 12). The change in beta weights will be used to assess if adding the co-variates changed whether language markers and SES indicators predict initial academic performance (intercept) and the rate of change in academic performance (slope).

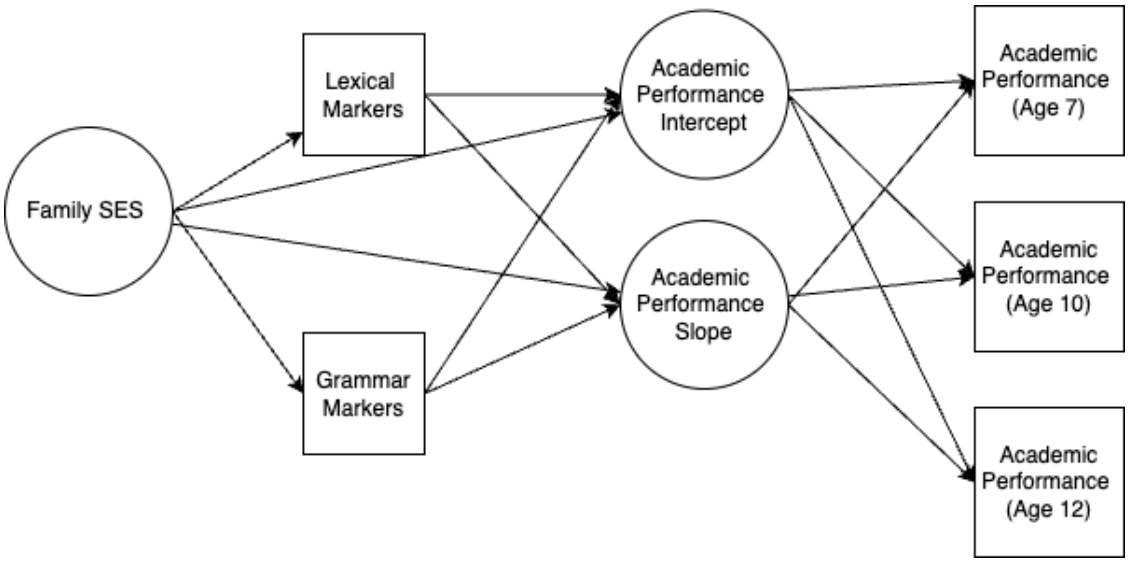

Figure 2. Hypothesised model for the relationship between Family SES, Language markers, and academic performance over time.

(RQ3)* To w*hat extent are associations between family SES and children's cognitive, educational, and literacy outcomes mediated by early life language environments?

Next, we will assess to what extent the effect of family SES on cognitive outcomes and academic performance is mediated via early life language environments.

To carry out a mediation analysis, three conditions must be met. First, for a marker of early life language environment must be significantly predicted by a family SES indicator; this can be identified in the analysis for RQ1. Secondly, the language marker must significantly predict the cognitive or academic outcome; this can be identified in the analysis for RQ2. Thirdly, the family SES indicator must also significantly predict the cognitive or academic outcome; to identify this, a further preparation step must be carried out.

We will run the same models as RQ2 (multiple regression for cognitive outcomes and LGCM for academic outcomes). Except in this analysis, language markers won't be included as predictors. All models will have SES indicators added as individual predictors as well as the relevant predictors. Beta weights will be used to assess whether an SES indicator significantly predicts a cognitive or academic outcome and, therefore, can be included in the regression analysis.

To test whether our SES indicators predict cognitive outcomes via early life language environments, we will construct a mediation model. The mediation model will include any language marker and SES indicators that satisfies the three conditions previously stated. For each outcome, the direct effect of the language marker on the cognitive outcome and SES on the cognitive outcome will be computed, as well as the indirect effect that SES predicts a cognitive outcome via

a language marker. The significance of the indirect effect will be tested through bootstrapping procedures. If the indirect effect is significant, this suggests a significant mediation.

To test whether our SES factors predict academic outcomes via early life language environment, we will construct a mediation model to calculate the language markers' mediation effect on the latent growth factors. The mediation model will include any language marker and SES indicator that satisfies the three conditions previously stated. For each outcome, the direct effect of SES on the outcome and the direct effect of the language marker on the outcome will be calculated, as well as the indirect of Family SES on the academic outcome via the language marker. The significance of the indirect effect will be tested through bootstrapping procedures. If the indirect effect is significant, this suggests a significant mediation.

Variables needed and at which ages:

Age 5:

Variable	Description
FAMILYID	Unique family identifier
ATWINID	Twin A ID
BTWINID	Twin B ID
SAMPSEX	Sex of Twins: In sample
ZYGOSITY	Zygosity
MAINLANG	Main Language Spoken to Twins
SESWQ35	Family Socio-economic status
HIEDM5	Mothers' Highest educational qualification
HIEDPM5	Partner's highest educational qualification
MSCGM5	Mother's social class group
PSCGM5	Partner's social class group
ED56M5	Total Household Income
MAGEM5	Mother's age at P5 assessment
PARM5	Current Partnership status mother
UNDER18L5	Number of people living in household under 18 years of age (including twins) - Phase 5
PSTATEL5	Partner status at end of LHC
ERDTM5	Mother's total Reading Score from WRAT-3
STEDRTM5	Standard Reading Score (Mother)
STEDRTGM5	Standard Reading Score (Mother) - Grouped
VOCTOTE5	Twin's vocabulary total score - Elder (WPPSI)
VERBALE5	Twin's age adjusted vocabulary score – Elder (WPPSI)
BDTOTE5	Twin's block design total score - Elder (WPPSI)
PERFE5	Twin's age adjusted block design score – Elder (WPPSI)
FMSS digitized recordings	Mother's FMSS Speech Recordings for each twin at age 5

Age 7:

Variable	Description
TRFPEE7	English school performance at age 7
TRFPME7	Maths school performance at age 7
ETRWM7	Raw Reading Scores - Real Words - Elder Twin (SWE)
STRWEM7	Standard Reading Scores - Real Words - Elder Twin (SWE)

STRWGEM7	Grouped Standard Reading Scores - Real Words - Elder Twin (SWE)
ETNWM7	Raw Reading Scores - Nonsense Words - Elder Twin (PDE)
·STNWEM7	Standard Reading Scores - Nonsense Words - Elder Twin (PDE)
STNWGEM7	Grouped Standard Reading Scores - Nonsense Words - Elder Twin (PDE)

Age 10:

Variable	Description
TRFPEE10	English school performance at age 10
TRFPME10	Maths school performance at age 10
ETRWM10	Raw Reading Scores - Real Words - Elder Twin (SWE)
STRWEM10	Standard Reading Scores - Real Words - Elder Twin (SWE)
STRWGEM10	Grouped Standard Reading Scores - Real Words - Elder Twin (SWE)

Age 12:

Variable	Description
TRFPEE12	English school performance at age 12
TRFPME12	Maths school performance at age 12

References cited:

d'Apice, K., Latham, R. M., & von Stumm, S. (2019). A naturalistic home observational approach to children's language, cognition, and behavior. *Developmental Psychology*, *55*, 1414–1427. https://doi.org/10.1037/dev0000733

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, *74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The Causal Role of Phoneme Awareness and Letter-Sound Knowledge in Learning to Read. *Psychological Science*, *23*(6), 572–577. https://doi.org/10.1177/0956797611435921

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Perszyk, D. R., & Waxman, S. R. (2018). Linking Language and Cognition in Infancy. *Annual Review of Psychology*, *69*(1), 231–250. https://doi.org/10.1146/annurev-psych-122216-011701

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill*. *Journal of Child Language*, *35*(1), 185–205. https://doi.org/10.1017/S0305000907008343

Rowe, M. L. (2018). Understanding Socioeconomic Differences in Parents' Speech to Children. *Child Development Perspectives*, *12*(2), 122–127. https://doi.org/10.1111/cdep.12271

Snelbaker, A. J., Wilkinson, G. S., Robertson, G. J., & Glutting, J. J. (2001). Wide Range Achievement Test 3 (wrat3). In W. I. Dorfman & M. Hersen (Eds.), *Understanding Psychological Assessment* (pp. 259–274). Springer US. https://doi.org/10.1007/978-1-4615-1185-4_13

von Stumm, S., Rimfeld, K., Dale, P. S., & Plomin, R. (2020). Preschool Verbal and Nonverbal Ability Mediate the Association Between Socioeconomic Status and School Performance. *Child Development*, *91*(3), 705–714. https://doi.org/10.1111/cdev.13364