Article:
Thorton, E., Matthews, D., Patalay, P. et al. (2026) Unequal educational outcomes for children with similar early childhood vocabulary but different socio-economic circumstances. Journal of Child Psychology and Psychiatry. ISSN: 0021-9630

https://doi.org/10.1111/jcpp.70117

Unequal educational outcomes for children with similar early childhood vocabulary but different socio-economic circumstances

Emma Thornton[1]

Danielle Matthews[2]

Praveetha Patalay[3]

Colin Bannard[4]

Manchester Institute of Education, University of Manchester, United Kingdom.[1]

Department of Psychology, University of Sheffield, United Kingdom.[2]

Social Research Institute and Department of Population Science and Experimental Medicine, University College London, United Kingdom.[3]

Department of Linguistics and English Language, University of Manchester, United Kingdom.[4]

Abbreviated title: Vocabulary, SEC, and educational attainment

Word count: 4,640

Abstract

In a purely meritocratic society educational outcomes would reflect ability, and only ability. Vocabulary size is a common measure of cognitive ability that predicts educational outcomes but is confounded with socioeconomic circumstances (SEC). **Methods.** In preregistered analyses of the nationally representative UK Millennium Cohort Study data (N=15,576), we used a series of multiple linear and logistic regression analyses to investigate the predictive value of age-5 vocabulary for age-16 educational outcomes and assess whether socioeconomic circumstance moderated this relation. **Results.** We show that age-5 vocabulary strongly predicted age-16 educational attainment, even after adjusting for both SEC and caregiver vocabulary (OR = 1.62, 95% CIs = [1.52;1.72]; (β = .22, 95% CIs = [.19;.24]). SEC also predicts educational attainment (OR = 2.05, 95% CIs = [1.92;2.19]), and modifies the association between vocabulary and educational attainment, whereby, a larger vocabulary was most advantageous for those in middle SEC groups (interaction term OR 1.09 [1.03; 1.15). **Conclusions.** Early child vocabulary is a strong predictor of children's educational outcomes - even when controlling for proxy measures of the home environment and genetics. Nonetheless, children who enter school with strong vocabulary skills but disadvantaged socio-economic circumstances still have only about a 50/50 chance of gaining gateway qualifications at age 16.

Keywords: vocabulary, socioeconomic inequalities, birth cohort, longitudinal, education

Abbreviations: Socioeconomic Circumstance (SEC); Millennium Cohort Study (MCS); General Certificate of Secondary Education (GCSE); National Five (N5)

Key points and relevance:

- Early language ability at age 5 predicts educational attainment at age 16, but we do not know if this relation holds equally across socio-economic strata.
- We tested whether children who enter school with strong cognitive abilities, as indexed by vocabulary size, achieve higher grades at age 16, regardless of their socio-economic circumstances. Analyses controlled for factors including caregiver language ability as a proxy for genetic inheritance and linguistic environment.
- In this study of a large, nationally representative sample, higher childhood vocabulary scores uniquely predicted better educational attainment at the end of secondary school.
- Children who were most disadvantaged were less likely to obtain secondary qualifications regardless of their cognitive ability when starting school. Conversely those who were highly socio-economically advantaged tended to obtain qualifications regardless of vocabulary ability at school entry. From a policy perspective, improving early language skills is likely to aid educational outcomes but would need to act in concert with measures to tackle other burdens of socio-economic disadvantage.

Language ability in early childhood is related to later educational attainment. It has often been assumed that this relationship is causal – that children's language skills affect their ability to access the curriculum and exchange their thoughts and ideas which in turn affects their academic performance (e.g., Hulme et al., 2020). However, determining a causal link is complicated by the existence of a third factor that is related to both language and educational attainment – children's socio-economic circumstances (SEC). In this study, we investigate these relationships, using the lens of vocabulary, an early cognitive predictor of educational outcomes. There are socioeconomic inequalities in language ability which are observed from

18 months of age in both the USA and the UK (Pace, Luo, Hirsh-Pasek, & Golinkoff, 2017; McGillion, Pine, Herbert & Matthews, 2017). That there are parallel socioeconomic inequalities in educational attainment has led many to claim that these are driven by early inequalities in cognitive skills, most specifically language, since language is generally the medium of education. However, this understanding rests on two untested assumptions. The first is that the relationship between language ability and educational outcomes isn't simply attributable to their both being driven by a third variable associated with social or other circumstances. The second is that, if there is an unconfounded relationship between language and educational outcomes, it holds across SEC levels equally. In other words, it assumes that we live in a broadly meritocratic society such that children who have similar skills at school entry will, all other things being equal, succeed in the education system, regardless of their SEC.

While there are recognised problems with the notion of meritocracy (Sandel, 2020), considerable effort is put into early interventions on the assumption that supporting children's language skills prior to school entry will allow them to access educational and social activities, thereby breaking the cross-generational transmission of disadvantage (e.g., *The Nuffield Early Language Intervention* in the UK; West et al, 2021). It is therefore important to test the basic assumptions that underpin these efforts. In this paper, we first aim to establish whether childhood vocabulary is indeed related to age-16 educational attainment in the UK, once SEC differences have been accounted for in a large, nationally representative sample of children. We then further explore the assumption of meritocracy by examining whether children of similar language ability upon entering the school system achieve similar education outcomes, regardless of their SEC.

We used a large, nationally representative, contemporary British birth cohort (the Millennium Cohort Study, MCS), to explore these research questions in a series of pre-

registered analyses. To address the first question, we tested whether early vocabulary predicted age-16 educational outcomes, adjusting for a range of control variables including SEC and caregiver vocabulary (with the latter used as a proxy for both the genetic component of vocabulary skill, and the language environment that cohort members are exposed to). We addressed the second question by testing whether the relation between early vocabulary and age-16 educational outcomes (in the form of compulsory national qualifications) is moderated by SEC.

Method

The Millennium Cohort Study (MCS) is a longitudinal birth cohort study of 19,518 young people from 19,244 families, born across England, Scotland, Wales and Northern Ireland between 2000-02 (Connelly & Platt, 2014). More information can be found here: https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/. Informed consent was obtained at each sweep. Cohort members with either a response on the vocabulary measure at age 5 *or* an educational attainment outcome at age 17 were included in our analyses. Where there were multiple cohort members from the same family, one of these was selected at random for inclusion, resulting in a total sample of 15,576 cohort members.

Measures

Predictor variable: age 5 vocabulary

At age 5, cohort members completed the Naming Vocabulary BAS test as a measure of expressive language (Elliott, Smith, & McCulloch, 1996). They were shown a series of images, one at a time, and asked to name each item. Progression through this test depends on performance, and poor performance may result in a different, easier set of items being administered. Cohort members were born over a 1.5-year period (September 2000- January

2002) and assessed over a range of months, so age at the time of testing may differ between cohort members. Therefore, we used *t*-scores adjusted for item difficulty and age (as published in the data; Connelly, 2013). These were converted to *z* scores for analyses.

Outcome variables: educational attainment at the end of secondary school.

Using self-reported qualification data, we focused on qualifications usually taken ~age 16 (i.e., GCSEs for England, Wales, and Northern Ireland and National Fives for Scotland). We conceptualised educational attainment in two ways: 1) a binary variable to represent a benchmark commonly used by the UK government to indicate educational achievement:(grade ≥C/4 in English, Maths and Science subjects, see Table 1 for specific subjects); and 2) a continuous variable of average performance across these subjects. The advantages of the binary measure are twofold: methodologically, it allowed us to include those cohort members who did not take the benchmarking subjects (because they took a mix of vocational and academic qualifications or took no qualifications at all). By definition, the average grade (continuous outcome variable) excluded those with no qualifications, since they did not have a grade to include in this analysis, and this is likely to be socioeconomically patterned. This approach also enhances the policy relevance of the findings, as the binary variable aligns with the measure commonly used by the UK government to benchmark educational attainment. Full details on this approach can be found in the Supplementary Methods.

INSERT TABLE 1 HERE

Modification variables: Socioeconomic circumstances

Using the *lavaan* package in R (Rosseel, 2012), confirmatory factor analysis was used to compute a factor score for parental SEC (see Supplementary Methods). In short, the latent variable was made up of highest household education, income, wealth, occupational status, and relative neighbourhood deprivation. Although the normed χ^2 statistic indicated a poor model fit (normed χ^2 (χ^2/5)) = 20.83), this test is sensitive to large sample sizes, rendering it an impractical fit statistic here(McNeish, 2018; Yuan & Chan, 2016). The remaining fit indices indicated the model was a good fit to the data (RMSEA = 0.036; SRMR = 0.023; TLI = 0.938; CFI= 0.969). Standardised factor loadings indicate that all variables loaded onto the latent construct (see Figure S1). Such a factor score has been shown to be better than any one indicator when predicting child vocabulary (Thornton et al., 2024). Individual indicators are as follows.

Parental education: Highest household NVQ level by cohort member age 3 (both academic and vocational qualifications derived into NVQ levels 1-5, with level 1 equating to GCSE grades D-G or NVQ level 1 vocational equivalents, and level 5 equating to higher degree qualifications: Rosenberg, 2012).

Income: UK OECD weighted income quintiles at age 3 (an indication of household income 1=lowest, 5=highest, accounting for family size). If data was missing, OECD weighted income quintiles at age 9 months were used instead.

Occupational status: highest household occupational status (NS-SEC 4 categories: higher managerial; intermediate; routine; unemployed) at 3 years. If data was missing, occupational status at age 9 months was used instead.

Wealth: a measure of total net wealth, taken from the age 11 sweep of the MCS, following the approach outlined by Moulton et al (2021), and Thornton et al (2024).

Relative neighbourhood deprivation: Indices of Multiple Deprivation (IMD) deciles, ranging from the most deprived decile to the least deprived decile taken from the age 3 sweep of the data. If data were missing, IMD deciles at age 9 months were used instead

Potential confounding variables.

Demographic confounders. Cohort member's sex at birth (male, female); ethnicity (White, mixed, Indian, Pakistani and Bangladeshi, Black or Black British, other ethnic group (including Chinese); whether English was spoken as an additional language in the home (English only, English and another language, only another language); and the country that the cohort member lived in (England, Wales, Scotland, Northern Ireland) were included as potential confounders.

Caregiver vocabulary.

Caregiver vocabulary was measured in the age-14 sweep of the MCS, using the Word Activity Test (Closs, 1986): caregivers had to identify the correct synonym from a choice of five, for 20 target words. In the MCS, there is a main respondent (usually the mother) and a partner respondent (usually the father), and both the main and partner respondents had the opportunity to complete the word activity test, with a different set of 20 words each. We used the mean score across respondents as a proxy for the home language environment and the heritable component of vocabulary skill.

Data analysis

All analyses were pre-registered on the Open Science Framework website (OSF number blinded for peer review), and all R code can be found on GitHub (blinded for peer review). Analyses consisted of a series of multiple logistic and multiple linear regression models.

Missing data strategy. Multiple imputation using chained equations with the *mice* package in R was used (van Buuren & Groothuis-Oudshoorn, 2011). Imputations accounted for the

interaction between SEC variables and age 5 vocabulary, due to the focus of our analyses (van Buuren & Groothuis-Oudshoorn, 2011). In line with guidance that the number of imputed datasets should exceed the overall proportion of missing data (16.5%), the dataset was imputed 25 times (White et al., 2011). Missing data for individual variables, including auxiliary imputation variables, can be found in Figure S2 (Supplementary Methods). We conducted sensitivity analyses using complete cases for each outcome variable, due to the high proportion of missing data (36.65% for the binary outcome variable; 50.73% for the continuous outcome variable). These analyses did not change the overall pattern of results (see Supplementary File, Sections 3 and 4).

Combined sampling and attrition weights were applied to the data to account for the stratified clustered design of MCS cohort data and the oversampling of subgroups, as well as for missing data due to attrition from the MCS before the age of 5, when the vocabulary measure of interest was measured.

Analysis plan.

Age 5 vocabulary and age 16 educational attainment

Logistic and linear regression models were used to estimate the relation between age-5 vocabulary and educational attainment. These were built in the same way, employing the first outcome (achieving ≥grade 4/C in the core subjects) in the logistic regressions, and the second outcome (mean grade across core subjects) in the linear regressions. First, the unadjusted relationship was estimated to assess whether there was an association. This model was extended to include sociodemographic confounding variables (sex, ethnicity, EAL, country, parent education, income, occupational status, wealth, and neighbourhood deprivation). Caregiver vocabulary was then added in a second model, and age-5 vocabulary was added to a model containing all potential confounding variables to see if an association remained after adjusting for these variables.

Nested model comparisons for imputed data (Meng and Rubin, 1992) were used to compare each model to the previous one, to see if the additional parameters predicted unique variance in educational attainment (determined by an improvement in model fit at each stage), revealing whether any relation holds above and beyond SEC and caregiver vocabulary factors. As a planned additional analysis, we investigated attainment in English, Maths and Science separately, rather than as a factor score (see Supplementary File, Section 7).

To establish whether findings were contingent on any analytic decisions, we ran two planned sensitivity analyses:1) including Welsh as a core subject for cohort members in Wales, where this is a compulsory GCSE (see Supplementary File, Section 5); and 2) analysing each country separately (England, Wales, Northern Ireland, and Scotland), due to the different education systems and examinations taken. This established whether any one country was driving any particular finding (Supplementary File, Section 6).

The moderating role of SEC

To assess if vocabulary is an equally important predictor across SEC groups, the above regression models were extended to include a parent SEC factor score*age-5 vocabulary interaction term. A model with the interaction term was compared to a model without using nested model comparisons for imputed data to test the significance of the interaction (Meng & Rubin, 1992*)*. To unpack any moderating effect of the factor score and establish if any one SEC indicator in particular drives any interaction, we included each indicator of SEC as a moderator in separate models, in a series of additional pre-registered analyses. Nested model comparisons for imputed data were again used to test the significance of the interaction term (Meng & Rubin, 1992).

Sensitivity analyses including interaction terms between potential confounders (remaining SEC variables) and between the predictor (vocabulary), and potential confounders and the moderator (each SEC variable in turn) are reported, to ensure the confounding effect

of SEC on the interaction term was accounted for (Keller, 2014) (see Supplementary File, Section 10).

Results

Descriptive statistics.

Descriptive statistics were pooled across 25 imputed datasets and can be found in Table 2. Proportions were similar between the whole cohort and our analytical sample (see supplementary file). However, means (±SD) for the average GCSE and average N5 grades were slightly higher in the whole cohort compared to our analytical sample.

INSERT TABLE 2 HERE

Age 5 vocabulary and age 16 educational attainment

Binary indication of whether cohort members achieved benchmark educational qualifications

In an unadjusted logistic regression model (not including any potential confounding variables), vocabulary was a significant predictor of attainment, such that with every SD unit increase in age-5 vocabulary, the odds of passing the benchmark of ≥ grade 4 on core subjects increased by 86% (see Table 3). Full model results for models containing sociodemographic and caregiver vocabulary factors can be found in Table S1. We subsequently assessed whether age 5 vocabulary explained variance over and above these variables.

Sociodemographic confounding variables improved the model fit compared to the null model ($Dm(36,3473.53)=39.87$, $p<.001$). Model comparisons further revealed caregiver vocabulary to predict variance in achieving the benchmark above sociodemographic variables ($Dm(1,81.25)=127.75$, $p<.001$). Further, adding age-5 vocabulary improved model fit ($Dm(1,86.05)=237.67$, $p<.001$) such that higher vocabulary was associated with increased

odds of passing the benchmark. After controlling for sociodemographic and caregiver vocabulary factors, with every SD unit increase in age-5 vocabulary, the odds of passing the benchmark of ≥ grade 4 on the core subjects increased by 62% (see Table 3).

Average grade across the core subjects (continuous outcome)

In an unadjusted linear regression model (not including any potential confounding variables), a positive relation was observed, such that higher age-5 vocabulary was associated with higher achievement (see Table 3). Full model results for models containing sociodemographic and caregiver vocabulary factors can be found in Table S1. We subsequently assessed whether vocabulary explained variance over and above these variables.

Sociodemographic confounding variables improved model fit compared to the null model (Dm(36,3418.41)=89.91, p<.001). Further, caregiver vocabulary predicted variance in performance above sociodemographic variables (Dm(1,78.57)=254.34, p=<.001). Finally, adding age-5 vocabulary improved model fit (Dm(1,61.43)=325.08, p<.001) such that higher vocabulary predicted higher achievement (β = .22, 95% CIs = [.19;.24]; see Table 2). For every standard deviation increase in age-5 vocabulary scores, average performance in the core subjects improved by 0.22 of a standard deviation.

Sensitivity and additional analyses did not change the overall pattern of results, with vocabulary consistently predicting unique variance in attainment (see Supplementary Material, sections 3-6).

The moderating role of SEC

To test whether the predictive value of age-5 vocabulary for age-16 educational attainment was equal across different socioeconomic groups, we ran a moderation analysis (see Table 3 and Figure 1). When a model with the SEC factor score*age-5 vocabulary interaction term was compared to a model without it, the interaction term was found to

explain significant variance (Dm(1, 192.02) =8.61, p = .003). In this model, higher vocabulary (OR =1.64 [1.54;1.74] and SEC (OR = 2.05 [1.92;2.19] predicted increased odds of passing benchmark educational qualifications. Furthermore, SEC moderated the relation between vocabulary and attainment (OR = 1.09 [1.03;1.15]). As visualised in Figure 1, the relationship between early vocabulary and educational outcomes is strongest for those in the middle SEC quintiles and weaker for both the most disadvantaged and the most advantaged children. The overall pattern of results persisted when the average grade in core subjects was the outcome (see Figure 2a): higher vocabulary scores (β = 0.22 [0.19; 0.24]) and SEC (β = 0.33 [0.31;0.36]) predicted higher educational achievement. SEC moderated the relation between vocabulary and attainment (β = 0.04 [0.02;0.05]), with the predicted average values increasing as vocabulary increases in all SEC groups, and the biggest increases seen in the middle SEC groups.

Since the average grade in core subjects excludes cohort members who did not take the core subjects (as they took other subjects or had no qualifications), Figure 2b shows the pattern of results when the outcome considers whether cohort members took the core subjects (regardless of whether the grade met the benchmark for the subject taken or not). Again, the likelihood of having the core subjects at GCSE increases with vocabulary skill, and this increase is greater in the middle SEC groups. In sum, regardless of how educational attainment is conceptualised, it is clear that the relation with early vocabulary is not equal across socioeconomic groups.

Sensitivity analyses

1) *Country specific analyses*

Including Welsh as a core subject for those living in Wales did not reveal a moderating effect of SEC (see Supplementary File, Section 5). Similarly, analyses by each UK country separately revealed a significant moderation effect of SEC in England only. (see

Supplementary File, Section 6). As can be seen from Figure S3, effect sizes are similar in the four countries, therefore this pattern of findings may be a result of analyses in the smaller UK countries being underpowered.

2) Investigating each SEC component in the moderation analyses

Parent education (Dm(5, 865.4)=2.41, p = .035), household income (Dm(4, 511.18) =2.57, p=.037), and occupational status (Dm(3, 379.31)=5.08, p=.002) all moderated the relation between age-5 vocabulary and achieving benchmark qualifications, reflecting the pattern observed in the main analysis (see Figure 3 and Supplementary File, Section 9). Wealth (Dm(4, 584.49) = 1.67, p=.155) and relative neighbourhood deprivation (Dm(9, 1102.32)=0.8, p = .621) did not moderate the relation (see Supplementary File, Section 9).

3) Adjusting for potential confounding on the interaction term

Extremely conservative sensitivity analyses (Keller, 2014) found no moderation effects for individual SEC indicators (see Supplementary File, Section 10). This suggests that the individual indicators of SEC are not separable in their interaction effects on the relation between vocabulary and educational attainment. The moderation of SEC is likely an additive effect of each SEC indicator, which likely have shared variance in their interaction terms.

4) Addressing the possibility of regression to the mean

Finally, we explored the possibility that the observed moderation effect could be attributed to 'regression to the mean', in an addition to the pre-registered analysis plan. Previous work has made the claim that the advantage of early cognitive abilities for later educational attainment is much greater in higher than lower SEC groups, only for a reanalysis to show that the effect was spurious (Jerrim & Vignoles, 2013). We therefore ran a series of

simulations based on the approach proposed by Krause and Pinheiro (2007). These simulation analyses indicate that a moderation effect persists beyond any presence of regression to the mean (pooled adjusted p value = .018; see supplementary file, section 11).

5) *Using a structural equation model framework to test the moderating role of SEC*

In an addition to our analysis plan, we used a structural equation model framework as a sensitivity analysis to address the potential for measurement error, using full information maximum likelihood to handle missing data. The findings from this set of sensitivity analyses did not yield meaningfully different results.

INSERT TABLE 2 HERE

INSERT TABLE 3 HERE

INSERT FIGURE 1 HERE

INSERT FIGURE 2 HERE

INSERT FIGURE 3 HERE

Discussion

In a set of pre-registered analyses, we aimed to unpack the relationship between early vocabulary, a common measure of cognitive ability, and age-16 educational attainment in a nationally representative and contemporary British birth cohort. First, we asked if the relationship between early vocabulary and educational outcomes holds when controlling for other variables including socio-economic circumstances. We found that age-5 vocabulary was indeed a strong predictor of educational attainment above and beyond family

SEC and caregiver vocabulary (and this relation held no matter how we operationalised education outcomes). Second, we asked whether the relationship between early vocabulary and educational outcomes holds across socio-economic strata equally. Here we found that the predictive value of age-5 vocabulary varied across SEC groups and was strongest in the middle SEC groups.

The role of early vocabulary in predicting educational attainment most likely reflects the fact that so many learning activities involve understanding and producing spoken and written language. Vocabulary skills lay the foundations for reading and mathematics (Ricketts, Lervåg, Dawson, Taylor, & Hulme, 2020; Slusser, Ribner, & Shusterman, 2019; Schuth, Köhne, & Weinert, 2017). Moreover, a rich vocabulary feeds itself in that, when a child comes across a new word, knowing the words that surround it helps them infer its meaning (Elleman, 2019; Larsen & Nippold, 2007). Although our analyses focus specifically on the predictive value of vocabulary at the beginning of formal education, we acknowledge vocabulary develops throughout childhood and adolescence (Ricketts et al., 2020), and the concurrent relationship between age-16 vocabulary and educational is likely to be even stronger.

We also found that *caregiver* vocabulary ability predicted child educational attainment, even after accounting for SEC. Previous studies have found caregiver vocabulary mediates the relation between SEC and offspring vocabulary (Sullivan et al., 2021), presumably because caregiver vocabulary is a proxy for both genetic and environmental factors. There is a heritable component of both childhood vocabulary and educational attainment: Chow & Wong, 2021; Selzam et al., 2017). Yet, a recent adoption study found that a beneficial impact of the quality of maternal language input on child vocabulary remained even in the absence of genetic effects (Coffey et al., 2021). The gene*environment hypothesis suggests that the influence of genetics may be suppressed among children

experiencing socio-economic disadvantage (Gottschling et al., 2019; Scarr & McCartney, 1983) and this may offer some explanation for the moderating effect of SEC.

SEC was found to moderate the relationship between early vocabulary and attainment. Three SEC measures, parent education, income and occupational status, appear to drive this. SEC can be conceptualised as a distal factor in a nested set of influences on child development (Bronfenbrenner, 1979). But what more proximal causal factors does SEC reflect and which of these explain why early language is a stronger predictor for some children than others? It is possible to sketch out several pathways via which we might explain the observed moderation. Children with limited language at school entry are more likely to access specialist help if their family is more advantaged; indeed, more advantaged parents are known to have more resources, capacity, and support in navigating the SEND system (Roy et al., 2014; O'Regan & Latham, 2025). Availability of intensive and stimulating educational support and experiences more generally likely explains some of the moderation we see at the top end of the socio-economic quintiles (Jerrim, 2017; Rakesh, Lee, Gaikwad & McLaughlin, 2025). Conversely, it is possible that non-academic outcomes may be more valued than academic qualifications among caregivers in the lowest SEC quintile (House of Commons, 2021) and academic expectations can influence educational outcomes (Rakesh et al., 2025). However, it should be recognised that we currently have very little grasp on why the observed moderating relation holds and whether the moderating role of SEC changes over developmental time (Black et al., 2025). Careful mixed methods research will be needed to unpick this in a way that yields tangible action points. This might be achieved alongside the staged introduction of new policies such that their impact can be assessed and thereby allow more rapid improvements of services for families.

Nonetheless, the current findings already lend weight to some practical approaches. The fact that there is a unique role of vocabulary in predicting attainment suggests that

interventions and policies to support early vocabulary development may well improve age-16 educational outcomes. The fact that this predictive relationship is moderated by SEC suggests that programmes and policies to mitigate a range of educational disadvantages associated with SEC are well founded in principle. These would likely need to follow right through to adulthood to have meaningful long-term effects since other modelling work suggests that as children enter adolescence there is a widening of disparity in vocabulary as a function of socio-economic circumstances (Thornton et al., 2024).

There are some limitations to this research that need to be kept in mind when interpreting findings. First, although we used a measure of vocabulary as a proxy for wider language ability, the two constructs are not synonymous, and this should be kept in mind when interpreting results. However, a wealth of evidence suggests that language ability can usefully be described as a unidimensional construct, with multiple dimensions loading onto a single factor early in development (Tomblin and Zhang, 2006; Hulme et al. 2024) making vocabulary a useful proxy. Second, different qualifications are used across the UK at the end of secondary school. We harmonised educational attainment in each country to the best of our ability. Third, our educational attainment outcomes are based on self-reported qualifications: cohort members were asked to indicate the subjects they studied and the grade they achieved, with subjects being presented to them in a list format. Future research could use the linked National Pupil Database (although this is only available for England). Nonetheless, work with the Millenium Cohort Study has reported a correlation of 0.89 between self-report and NPD data, despite some under/over estimation at either end of the distribution (Anders, Green, Henderson & Henseke, 2024). A correlation of .99 between self-report and NPD qualification data has been reported in another dataset (Smith-Woolley et al., 2018). Together, these findings suggest that self-report qualification data is reliable. Finally, as with any longitudinal data analysis, missing data had to be accounted for. Families experiencing lower SEC tend to

be underrepresented in data collection sweeps of cohort studies as time goes on (Mostafa & Wiggins, 2014), and it is also possible that those with lower vocabulary may have been less likely to respond to later sweeps, meaning effect sizes could be underestimates. However, analyses were sample and attrition weighted and we used multiple imputations with a rich set of auxiliary indicators to account for missing data, which is a 'best effort' approach (Little & Rubin, 2019).

Despite these limitations, the strengths of this research lie in the large, nationally representative longitudinal birth cohort used, which has rich data using gold-standard, researcher-administered, standardised tests of both caregiver and child vocabulary and a broad range of SEC indicators across all four nations of the UK. The strengths of this dataset allowed us to robustly estimate the relation between early vocabulary and the attainment of important gateway qualifications. Further, we can be confident that our results are generalisable across the United Kingdom, given the nationally representative sample. Being able to include caregiver vocabulary in analyses, also allowed us to account indirectly for likely effects of both the home environment and genetics. Finally, we can have confidence that there is a true moderating effect of SEC in the relation between early vocabulary and education outcomes, due to simulation analyses included to account for regression to the mean.

Conclusion

Overall, we found that in a large, nationally representative cohort, age-5 vocabulary predicted unique variance in age-16 educational qualifications. This was the case even after adjusting for SEC and caregiver vocabulary. However, we also found SEC to moderate the relation between age-5 vocabulary and age-16 educational outcomes such that the benefits of a larger early vocabulary were most stark for children in middle SEC bands. This suggests

that supporting vocabulary development as a means of improving educational attainment is well founded but not sufficient in isolation. Support for children and their caregivers across the lifespan is necessary to better help all children thrive in education.

Ethical information

Ethical approval was not necessary for this paper, which consisted of secondary data analysis of the UK Millennium Cohort Study. However, end user license agreements were agreed for data access. Ethical approval information of the Millennium Cohort Study is available here: https://cls.ucl.ac.uk/data_documentation/mcs-birth-age-7-ethical-review-and-consent/

References

Anders, J., Green, F., Henderson, M., & Henseke, G. (2024). Private school pupils' performance in GCSEs (and IGCSEs). Cambridge Journal of Education, 1-19.

Black, M., Akanni, L., Adjei, N.K., Melendez-Torres, G J., Hargreaves, D., & Taylor-Robinson, D. Impact of child socioemotional and cognitive development on exam results in adolescence: findings from the UK Millennium Cohort Study *Archives of Disease in Childhood* 2025;**110:**645-650.

Bornstein, M. H., Hahn, C. S., & Putnick, D. L. (2016). Stability of core language skill across the first decade of life in children at biological and social risk. Journal of Child Psychology and Psychiatry and Allied Disciplines, 57(12), 1434–1443. https://doi.org/10.1111/jcpp.12632

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University PressChow, B. W., & Wong, S. W. L. (2021). What does genetic research tell us about the origins of language and literacy development? A reflection on Verhoef et al. (2020). Journal of Child Psychology and Psychiatry, 62(6), 739–741. https://doi.org/10.1111/jcpp.13399

Closs, S. J. (1986). APU vocabulary test (multiple choice format, 1986). Kent: Hodder and Stoughton Educational Ltd.

Coffey, J. R., Shafto, C. L., Geren, J. C., & Snedeker, J. (2021). The effects of maternal input on language in the absence of genetic confounds: Vocabulary development in internationally adopted children. Child Development, 1–17. https://doi.org/10.1111/cdev.13688

Connelly, R. (2013). *Millennium Cohort Study data note 2013/1: Interpreting test scores*. *September*, 1–20.

Connelly, R., & Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). International Journal of Epidemiology, 43(6), 1719–1725. https://doi.org/10.1093/ije/dyu001

Elleman, A. M., Oslund, E. L., Griffin, N. M., & Myers, K. E. (2019). A review of middle school vocabulary interventions: Five research-based recommendations for practice. Language, Speech, and Hearing Services in Schools, 50(4), 477–492. https://doi.org/10.1044/2019_LSHSS-VOIA-18-0145

Elliott, C. D., Smith, P., & McCulloch, K. (1996). British Ability Scales Second Edition (BAS II) Early Years. NFER-Nelson.

Gottschling, J., Hahn, E., Beam, C. R., Spinath, F. M., Carroll, S., & Turkheimer, E. (2019). Socioeconomic status amplifies genetic effects in middle childhood in a large German twin sample. Intelligence, 72, 20-27.

House of Commons. (2021). The forgotten: how White working-class pupils have been let down, and how to change it (Issue June).

Hulme, C., Snowling, M. J., West, G., Lervåg, A., & Melby-Lervåg, M. (2020). Children's Language Skills Can Be Improved: Lessons From Psychological Science for Educational Policy. *Current Directions in Psychological Science*, *29*(4), 372–377. https://doi.org/10.1177/0963721420923684

Hulme, C., McGrane, J., Duta, M., West, G., Cripps, D., Dasgupta, A., ... & Snowling, M. (2024). LanguageScreen: The development, validation, and standardization of an automated language assessment app. Language, speech, and hearing services in schools, 55(3), 904-917.

Jerrim, J. (2017). Private tuition and out of school study, new international evidence. In The Sutton Trust (Issue September).

Jerrim, J., & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *176*(4), 887-906

Krause, A., & Pinheiro, J. (2007). Modeling and simulation to adjust p values in presence of a regression to the mean effect. *The American Statistician*, *61*(4), 302-307.

Keller, M. C. (2014). Gene × environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution. Biological Psychiatry, 75(1), 18–24. https://doi.org/10.1016/j.biopsych.2013.09.006

Larsen, J. A., & Nippold, M. A. (2007). Morphological Analysis in School-Age Children: Dynamic Assessment of a Word Learning Strategy. Language, Speech, and Hearing Services in Schools, 38(3), 201–212. https://doi.org/10.1044/0161-1461(2007/021)

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018). Unpicking the developmental relationship between oral language skills and reading comprehension: It's simple, but complex. Child development, 89(5), 1821-1838.

Little, R., & Rubin, D. (2019). *Statistical analysis with missing data*. Wiley-Blackwell.

Lupton, R., Thomson, S., Velthuis, S., & Unwin, L. (2021). Moving on from initial GCSE "failure": Post-16 transitions for "lower attainers" and why the English education system must do better (Vol. 4, Issue February, pp. 1–140). Nuffield Foundation. https://www.research.manchester.ac.uk/portal/files/187105835/FINAL_main_report_for_publishing.pdf

McGillion, M., Pine, J. M., Herbert, J. S., & Matthews, D. (2017). A randomised controlled trial to test the effect of promoting caregiver contingent talk on language development in infants from diverse socioeconomic status backgrounds. Journal of Child Psychology and Psychiatry, 58(10), 1122-1131.

Meng, X. L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, *79*(1), 103–111. https://doi.org/10.1093/biomet/79.1.103

Mostafa, T., & Wiggins, R. D. (2014). Handling attrition and non-response in the 1970 British Cohort Study. *CLS Working Paper 2014/2*, *June*.

Moulton, V., Goodman, A., Nasim, B., Ploubidis, G. B., & Gambaro, L. (2021). Parental Wealth and Children's Cognitive Ability, Mental, and Physical Health: Evidence From the UK Millennium Cohort Study. *Child Development*, *92*(1), 115–123. https://doi.org/10.1111/cdev.13413

O'Regan, C., & Latham, K. (2025). Double disadvantage? Socio-economic inequalities in the SEND system. The Sutton Trust.

Pace, A., Luo, R., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Identifying Pathways Between Socioeconomic Status and Language Development. Annual Review of Linguistics, 3(1), 285–308. https://doi.org/10.1146/annurev-linguistics-011516-034226

Rakesh D, Lee PA, Gaikwad A, McLaughlin KA. Annual Research Review: Associations of socioeconomic status with cognitive function, language ability, and academic achievement in youth: a systematic review of mechanisms and protective factors. J Child Psychol Psychiatry. 2025 Apr;66(4):417-439. doi: 10.1111/jcpp.14082. Epub 2024 Dec 3. PMID: 39625804; PMCID: PMC11920614.

Ricketts, J., Lervåg, A., Dawson, N., Taylor, L. A., & Hulme, C. (2020). Reading and Oral Vocabulary Development in Early Adolescence. Scientific Studies of Reading, 24(5), 380–396. https://doi.org/10.1080/10888438.2019.1689244

Rosenberg, R. (2012). MCS2 : Guide to Derived Variables. August.

Roy, P., Chiat, S., & Dodd, B. (2014). Language and Socioeconomic Disadvantage: From Research to Practice. London, UK: City University London.

Sandel, M. J. (2020). The tyranny of merit: What's become of the common good?. Penguin UK.

Scarr, S., & McCartney, K. (1983). How People Make Their Own Environments : A Theory of Genotype → Environment Effects. Child Development, 54(2), 424–435.

Schuth, E., Köhne, J., & Weinert, S. (2017). The influence of academic vocabulary knowledge on school performance. Learning and Instruction, 49, 157–165. https://doi.org/10.1016/j.learninstruc.2017.01.005

Selzam, S., Krapohl, E., Von Stumm, S., O'Reilly, P. F., Rimfeld, K., Kovas, Y., Dale, P. S., Lee, J. J., & Plomin, R. (2017). Predicting educational achievement from DNA. Molecular Psychiatry, 22(2), 267–272. https://doi.org/10.1038/mp.2016.107

Slusser, E., Ribner, A., & Shusterman, A. (2019). Language counts : Early language mediates the relationship between parent education and children's math ability. Developmental Science, 22(3). https://doi.org/10.1111/desc.12773

Smith-Woolley E, Pingault JB, Selzam S, Rimfeld K, Krapohl E, von Stumm S, Asbury K, Dale PS, Young T, Allen R, Kovas Y, Plomin R. Differences in exam performance between pupils attending selective and non-selective schools mirror the genetic differences between them. NPJ Sci Learn. 2018 Mar 23;3:3. doi: 10.1038/s41539-018-0019-8. PMID: 30631464; PMCID: PMC6220309.

Sullivan, A., Moulton, V., & Fitzsimons, E. (2021). The intergenerational transmission of language skill. The British Journal of Sociology, 72(2), 207–232. https://doi.org/10.1111/1468-4446.12780

The Children's Commissioner. (2019). Briefing: The Children Leaving School with Nothing. 1–11. https://doi.org/https://www.childrenscommissioner.gov.uk/wp-content/uploads/2019/09/cco-briefing-children-leaving-school-with-nothing.pdf

Thornton, E., Matthews, D., Patalay, P., & Bannard, C. (2024). Investigating how vocabulary relates to different dimensions of family socio-economic circumstance across developmental and historical time. Language Development Research, 4(1), 80-174. https://doi.org/10.34842/mhqh-9g10

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. Journal of Speech, Language, and Hearing Research, 49(6), 1193-1208.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

West, G., Snowling, M. J., LervAag, A., Buchanan-Worster, E., Duta, M., Hall, A., ... & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: evidence from a cluster randomized controlled trial. Journal of Child Psychology and Psychiatry, 62(12), 1425-1434.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377–399. https://doi.org/10.1002/sim.4067