ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Economics and Business Economics

Major Health Economics

The relationship of genetic risk of mental health issues and childhood socio-economic status on years of education

This paper will investigate if a relationship between the genetic risk of mental illness can be found by studying the genetic risk of neuroticism, depression, and subjective well-being, and years of education. It will also include factors such as socio-economic status during childhood, and if this interacts with the influence of the genetic risk of neuroticism, which influences years of education. The results will be obtained via multiple regression models, using data from the Health and Retirement Study. The study has found a negative association between the genetic risk of neuroticism and years of education, as well as a positive association between childhood socio-economic status and years of education. It could not establish a difference in association in different childhood socio-economic groups, when observing the genetic risk of neuroticism on years of education.

Name student: Daan van Oeveren

Student ID number: 501103

Supervisor: S.F.W. Meddens

Second assessor: R.D. Pereira

Date final version: 12-7-2021

Table of Contents

1. **Introduction** 3
2. **Literary Review** 5
3. **Methodology and Analysis** 7
4. **Results** 12
5. **Discussion** 18
6. **Conclusion** 21
7. **References** 23
8. **Appendix** 25

A1 Acronyms 25

A2 Tables 25

1. Introduction

In recent years, the prevalence of mental health issues among students has been on the rise. These issues consist mostly of depression, anxiety or suicidal tendencies (Liu et al., 2019). Part of the risk of these mental health issues developing can be lead back to someone their genes. There are specific bits of your genetic code which decide the likelihood of someone developing certain traits. These traits in itself, can be affected by environmental sources (Nugent et al. 2011).

One of these environmental sources can be attributed to the surroundings someone grew up in. These surroundings could consist of living in a well-off family, or close-knit parental relationships. These factors could be considered a part of the socio-economic status someone grew up in (Vable et al., 2017). Taking into account how gene expression can be affected by the environment someone grew up in, one could expect different outcomes. One could consider if a difference in socio-economic status while someone was attending any form of education, could influence the genetic risk of a mental health issue developing. If this is the case, one could also consider arguing that if this socio-economic status has such an effect influencing genetic risks, it could increase or decrease the total amount of education someone obtained, depending on which socio-economic environment they grew up in. This raises the following research question:

How does a genetic risk of mental health issues, affected by childhood socio-economic status, influence years of education?

To answer parts of this research questions, three hypotheses have been brought forward. Each of these hypothesis will deal with a different part of the research question. Each hypothesis will be stated next, but they will be argued for in the literary review later on. The first two hypothesis focussing on each individual effect of genetic risk and socio-economic status on years of education, and the third hypothesis investigating an interaction between these two.

The first hypothesis suggests that a genetic risk of mental health issues, will correlate with a decrease in the amount of years studied by someone. This taking into account the research of Fletcher (2010), which showed the decrease in years of schooling when faced with depressive symptoms.

For the second hypothesis, it will be suggested that a higher socio-economic status will correlate with an increase in the amount of years studied by someone. A study by Walpole (2003) finds

differences in students among the experiences they have in college, when differentiating by socio-economic status.

For the third hypothesis, it will be suggested that an interaction between the genetic risk of mental health could have a relationship with the years studied by someone. This relationship is suspected to demonstrate itself in a reduction in years of education lost with a high childhood socio-economic status. A previous study by Papageorge & Thom (2018) has shown a connection between genes predicting educational attainment and the socio-economic status a kid grows up in, with these two factors interacting. It will be argued that genes influencing mental health will have a negative relationship, with a higher socio-economic status, it will be less likely that someone will suffer from depressive disorder, thus increasing years studied.

This research is an addition to current literature, while a lot of papers mention the influences of mental health on education, or socio-economic status on mental health or education, no specific research has looked at the interaction between socio-economic status and the genetic risk for mental health issues. Some previous studies, as mentioned before, have already looked at genes predicting certain characteristics (Krapohl et al., 2018), and how they can influence the outcome of some people their lives. With this paper also investigating the genetic risk of certain traits, as well as an interaction effect with the socio-economic environment during childhood, this would add on to this already existing literature.

The results could also influence policy making at schools. If any results are significant on the interaction between socio-economic status and mental health having an effect on education, schools could focus on students from a low socio-economic background and offer them extra psychological support. Even if no significant result can be found between these two variables interacting, it could still be possible that a significant result for a genetic risk for mental disorder could be found which could correlate with a decrease in years of education. If these results deem to be influential enough for people with a high genetic risk for mental health problems, they could take this into account by paying attention to any early signs of mental illnesses, taking action against it, thus not obtaining a reduction in years of education.

First, a short discussion of some current literature will be given, next an introduction to gene risk scores will be discussed. It will detail how the genetic scores are calculated, and which methods are used. After this, the data used will be introduced. It will be discussed how each variable was obtained or constructed. Next, the method of analysis will be discussed, as well as which variables will be used. Next, the results for each analysis will be shown, as well as

discussed. It will detail if any significant results have been found, as well as how this answers the above named hypothesis. Finally, the results will be summarized, and suggestions for future research will be given.

2. Literary Review

To first define part of the research question, when mentioning the genetic risk for mental health issues, the genetic score for neuroticism is intended. Neuroticism can be linked to less beneficial living situations, like a lower overall happiness, or a less healthy way of living (Lahey, 2009). People who portray signs of neuroticism tend to overreact often to small issues. Additionally, it also mentioned that neuroticism has some causal links predicting mental health disorders, like depression and schizophrenia, due to these traits sharing the same genetic variants.

Defining the second part of the research question, when talking about childhood socio-economic status (SES), the factors contributing to the score in this paper are discussed by Vable et al. (2017). The factors being referenced when talking about childhood SES consist of financial capital during childhood, as well as human capital, which contains parental educational attainment.

Bringing up the first hypothesis, the aforementioned study by Fletcher (2010) will be taken into account. With a sample of 2400 twin pairs, based in the United States, he finds a negative association of the amount of depressive symptoms on years of schooling. Yet while controlling for a variety of conditions, like tobacco usage or obesity, no significant result can be attributed to the amount of depressive symptoms portrayed. The study also looks at drop-out rates. In this model, no significant results can be found for a relationship between depression and drop-out rates. Fletcher also argues for reverse causality, where an increased effect in mental health issues become apparent because of a low education. He also raises the risk of certain omitted variables which could increase the risk of people dropping out of school at an early age, thus decreasing their total years of education. A study by Jordan et al. (1996) confirms that certain situations, like feeling unwelcome or fearing other students for example, do increase the risk of someone dropping out early. The sample size consisted of 25000 students, situated in the United States, between the ages of 14 and 16 years old. At both these ages, the data was collected. In the study, of the reasons mentioned by the students for dropping out, any mental health issues were not mentioned. But, some of the reasons mentioned could contribute to mental health issues developing. Besides this, the study also dives deeper into different subgroups of the

population, even controlling for SES during high school. It mentions that differences in SES explained some of the drop-out rates, thus the decrease in years of education obtained.

This leads to the second hypothesis. The aforementioned study by Walpole (2003), finds that SES experienced during college can affect the outcome of people attending graduate school. The sample consists of around 12000 students from the United States, over a 4 year college degree. It most notably shows the differences in activities performed between high and low SES classes. These differences are found in time spent studying and amount of hours worked for example. It also shows a lower amount of percentage in the low SES group attending graduate school, or obtaining a degree. One thus could argue that these decreases in educational attainment, also lowers the years of education someone obtained. However, this relationship can not be established with certainty, since the research did not explicitly look at years of education, just the earned degrees.

When investigating the third hypothesis, a study done by Papageorge & Thom (2018), with a sample size of around 19000 people based in the United States, found that both the genetic effects of educational attainment, as well as parental influences have an effect on the highest educational level obtained. Next to genes predicting a statistical significant result for college graduation, it found an even stronger results for people who experienced a higher socio-economic status during this time. This SES was based on the father his income, financial wealth, paternal employment, and if the family had to move for financial reasons. When controlling for another part of childhood SES, like parental education, they found any relationship between the genetic predictor for educational attainment and SES to decrease. Thus the childhood SES used in this paper will also control for parental education. They also found that the interaction effect between SES group and genetic risk varied at different levels of education. For lower levels of education, they found a negative interaction effect, while they found the opposite at higher levels of education. However this paper focusses mostly on the interaction between SES and the genetic predictor for educational attainment, one could argue a similar effect for the genetic risk for mental health issues could be found.

Another research describing environment interactions with genes predicting depressive symptoms, is a paper by Nugent et al. (2011). It describes a relationship between things causing stress in life, like life events or family related causes, and the genetic risk for depression. From these family related causes, it names poverty and maternal interactions as one of the examples. These examples are also included in the earlier mentioned childhood SES. The paper demonstrates these interaction effects of environment and genes when studying predictors for

mental health issues. It is important to note that this study does not directly link to genes predicting the genetic risk of mental health issues. It mostly shows that these interaction effects are existent, and observable. It also argues that multiple different environments can cause the same result when interacting with genes. Thus one could conclude that it is possible to also find an interaction effect between the genetic risk of mental health, and childhood SES when observing years of education obtained.

A study done by Wang & Sheikh-Khalil (2014) looked at how childhood interactions with their parents influenced their educational results, as well as their risk of developing depression. In their analysis, they controlled for different SES. These different SES were divided into three different groups, high, moderate, and low. The SES used consisted of the parental education, and the family income. The sample size consisted of around 1000 students from the United States, spanning over three time periods. They found that children who had parents who engaged with their children at school during events, had a lower chance of developing depression one year later. The same negative result was found for SES affecting the chance of depression developing the next year. When observing educational performances being influenced, parental interactions at home, as well as SES both had a positive influence on the grades obtained. This could support the third hypothesis mentioned, finding that an higher childhood SES could contribute to a smaller chance of depression developing, thus increasing the amount of years of education obtained.

With the aforementioned papers, there should be a basis to investigate an interaction effect between the PGS of neuroticism, and childhood SES. With no current literature describing or mentioning this interaction effect, it could be an addition to current scientific literature.

3. Methodology and Analysis

When measuring genetic risks or genetic influences, a polygenic score (PGS) can be constructed. This score can then be used to predict if a certain trait has a risk of developing, like certain mental health risks. It can also be used to link genetic effects to achievements later in life, like educational attainment (Dudbridge, 2013; Krapohl et al. 2018). One way to determine which trait is linked to which genetic marker is done via a genome-wide association study (GWAS). It tries to detect commonalities between genetic markers and their traits on big sample size of a population (Visscher et al. 2012). Dudbridge (2013) defined the PGS to be obtainable via the following formula, making use of GWAS.

$$\hat{S} = \sum_{i=1}^{m} \hat{B}_{il} G_i$$

It demonstrates that the PGS is obtained by analysing a certain trait, taking into account the sum of the weighted genetic markers.

An addition onto these GWAS studies, is a study done via multi-trait analysis of GWAS (MTAG). This method of analysis, takes into account multiple traits when constructing a PGS. One reason why this could increase statistical power is that multiple traits might share some of the same genes. One point of attention mentioned by the researchers is that one might interpret the coefficient for only the one trait, while the effect might be caused by one of the others traits which are also taken into account when using this variable (Turley et al. 2018). These PGS are standardized, meaning that they will have a mean of zero, and a standard deviation of one. This is done to make these scores comparable for different people in the same dataset.

For the analysis, data from the Health and Retirement Study (HRS) will be used. One dataset will be constructed from reported variables from four different datasets and studies from the HRS. Some variables will be obtained from a dataset made available by the HRS, containing the longitudinal data from people between 1992 and 2018. Variables with the genetic score are obtained from a dataset constructed by Turley et al. (2018). This dataset takes includes PGS analysed via the earlier mentioned MTAG, thus taking into account multiple traits. The variable for education attainment is obtained from a dataset, collected between 2006 and 2012 for some HRS respondents. This data is based of European ancestry and genotyped by the Center for Inherited Disease Research. Each PGS was discovered via a GWAS. The final variables describing childhood SES is a dataset, created by Vable et al. (2017).

These different variables, will all be collected into one dataset. People whose scores were not available for any of these variables will be dropped from the dataset. This keeps 8579 observations in the dataset, who have had both their childhood SES reported, as well as their genes analysed, and reported by the data collected by Turley et al. (2018).

Next the data will be filtered for the variables being used to answer the research question. For each analysis containing a PGS, the 10 principal components for European ancestry have to be taken into account, as well as their birth year, gender, birth year multiplied by gender, and their birth year squared. From the variables obtained from Turley et al. (2018) the PGS for neuroticism analysed via MTAG will be kept as a variable, with the reasoning this is the most

complete score for genetic risk to see if someone could develop mental health issues. It will also keep the GWAS analysed PGS for neuroticism, depression and subjective well-being. These scores will be kept to look at the differences in R^2. Additionally, the PGS for educational attainment from the obtained from the HRS PGS variable list will be kept as a control variable in the analysis.

From the variables obtained by Vable et al. (2017), the variable for the average childhood SES will not be used. This variable includes financial capital during childhood, social capital taking into parental relationship and household, and human capital taking into account the parental years of education. Since the social capital score was calculated based of mostly parental relationships. These relationships in itself could be negatively affected due to the parents also showing signs of mental illnesses (Downey & Coyne, 1990). It does not have to be the case that these parental relationships are only affected by parental signs of depression, but could also be influenced by other traits present in the parents, or the children. To reduce the effect this could have via endogeneity, a different childhood SES score was constructed. This childhood SES variable was constructed from data obtained by Vable et al. (2017). The new variable was constructed with the average score of the financial capital during childhood, as well as the human capital. This score was then standardized for this group.

The constructed average childhood SES variable is a continuous variable, this could leave interpretability to be desired. Thus to make these values better to interpret, they are assigned a categoric variable for high, middle and low SES. These three categories will be based of the tertiles of the dataset containing these childhood SES variables. Thus the top one third of the dataset, will be considered having had a high childhood SES. For the analysis, the birthyear will also be centred, meaning the average birthyear of the entire dataset will be subtracted from someone their actual birthyear, hoping to increase clarity when doing the analysis. Below in table 1, the descriptive statistics of the dataset are displayed.

Table 1. Descriptive Statistics.

	Mean	**Standard Deviation**	**Min**	**Max**	**Observations**
Years of Education	13.161	2.539	0	17	8,565
PGS Neuroticism MTAG			-3.856	3.875	8,579
PGS neuroticism GWAS			-3.876	3.988	8,579
PGS depression GWAS			-3.551	3.741	8,579
PGS subjective well-being GWAS			-3.833	3.592	8,579
PGS Educational Attainment GWAS			-3.380	3.610	8,579
Gender	1.584	0.493	1	2	8,579
Birthyear	1937.821	10.420	1905	1974	8,579
Childhood SES	0.274	0.868	-3.322	2.809	8,579
Low	-0.794	0.506	-2.989	-0.119	2,860
Medium	0.307	0.231	-0.119	0.713	2,867
High	1.274	0.489	0.714	3.020	2,852

The PGS values are standardized, meaning their mean equals 0 and their standard deviation 1. Childhood SES being displayed in both all the values of the entire dataset, as well as the statistics of the three individual childhood SES groups, consisting of low, medium, and high. Gender is a binary valuable being able to take the value 1 for a male participant, and the value of 2 for a female participant.

The table with descriptive statistics above still displays the actual birthyear for each participant. In the regression models, the centred birthyear will be used. This is done by taking the average of the birthyear of all the participants, and the subtracting this average birthyear of people their actual birth year to centre it. Another thing that is also shown is that there is a slight majority of females in this dataset, with 58,4% of people in the dataset being female.

With these variables, multiple OLS regression models will be run to establish any significant associations between these variables. Since there will not be made use of Mendelian randomization (Emdin et al., 2017), there can only be spoken of a correlation or association,

and not a causal effect. Each model will be representative of each hypothesis, being able to confirm it, or reject it based on the outcome. When constructing a model for the first hypothesis, years of education will be determined by a constant, the coefficients for different PGS of different traits, as well as the covariates and principal components. The PGS for educational attainment will be used as a control variable. The expected result, based on the first hypothesis, is that a negative coefficient will be found for the PGS of neuroticism, analysed via MTAG.

1) $\boldsymbol{Years\ of\ Education}$

$$= \boldsymbol{Constant} + \boldsymbol{\beta_{PGS_I}} \times \boldsymbol{PGS_I} + \boldsymbol{\beta_{Covariates_I}} \times \boldsymbol{Covariates_I} + \boldsymbol{\varepsilon}$$

For the next model, an OLS regression model will be constructed based on people their childhood SES. It will look at if there is an association between years of education and the childhood SES group someone was in. As mentioned before, the childhood SES groups are divided into three groups, based on tertiles. This means that the constant in the model already includes the coefficient for the low childhood SES group. The only two childhood SES groups left with a coefficient represented are the medium and high childhood SES groups. As a control variable, the PGS for educational attainment is used. This means the 10 PC, as well as the beforementioned covariates will also be taken into account in this regression model.. The expected result, based on the second hypothesis, is that a positive coefficient will be found for each of the higher childhood SES groups. Meaning that when someone is part of a higher childhood SES group, they will have an increase in years of education.

2) $\boldsymbol{Years\ of\ Education}$

$$= \boldsymbol{Constant} + \boldsymbol{\beta_{SES_I}} \times \boldsymbol{SES_I} + \boldsymbol{\beta_{Covariates_I}} \times \boldsymbol{Covariates_I} + \boldsymbol{\varepsilon}$$

For the next OLS regression model, the focus will be on the interaction effect variable. This will also include the earlier genetic risk score for neuroticism, and the three different childhood SES groups. Besides this, it will also take into account the earlier mentioned covariates and PCs. It is expected for the interaction effect to obtain a positive coefficient in line with the third hypothesis. This interaction effect is displayed as IE in the regression formula. The interaction effect is obtained by multiplying the childhood SES scores with the PGS for neuroticism, thus creating a continuous variable.

3) $\boldsymbol{Years\ of\ Education}$

$$= \boldsymbol{Constant} + \boldsymbol{\beta_{SES_I}} \times \boldsymbol{SES_I} + \boldsymbol{\beta_{PGS_I}} \times \boldsymbol{PGS_I} + \boldsymbol{\beta_{IE}} \times \boldsymbol{IE}$$
$$+ \boldsymbol{\beta_{Covariates_I}} \times \boldsymbol{Covariates_I} + \boldsymbol{\varepsilon}$$

For the final OLS regression model, years of education will be divided into different groups, based on the earlier mentioned childhood SES groups. This should make the results easier to put into perspective. Thus in total, 3 models will be shown. One for the low childhood SES group, one for the medium childhood SES group, and finally one for the high childhood SES group. This is to see if there is any significant difference between the coefficients of the genetic risk of neuroticism. This will be shown by examining the 95% confidence intervals of the variables. If these intervals were to overlap, it means the results obtained are not statistically significantly different enough. Thus it would mean that the coefficients for the genetic risk of neuroticism, based on childhood SES groups can not be called statistically significantly different. With the third hypothesis in mind, it is expected to find a statistically significant difference between these coefficients. Besides this, the genetic score for education attainment will be used as a control variable. Additionally, the 10 principal components, as well as the earlier named covariates of gender, centred birthyear, centred birthyear times gender, and centred birthyear squared, will also be added into the model. The expected results, based on the third hypothesis, is that people who experienced a higher childhood SES, have a lower influence from the genes predicting mental illness, compared to those experiencing a lower childhood SES.

$$\mathbf{4)}\ \boldsymbol{Years\ of\ Education_{SES}} = \boldsymbol{Constant} + \boldsymbol{\beta_{PGS_I}} \times \boldsymbol{PGS_I} + \boldsymbol{\beta_{Covariates_I}} \times \boldsymbol{Covariates_I} + \boldsymbol{\varepsilon}$$

After each model has been analysed, a conclusion can be made about each hypothesis.

4. Results

First, the association between years of education and the PGS score for neuroticism, obtained via MTAG, will be looked at. These coefficients are obtained via an OLS regression analysis. The results of this are displayed in column one of table 2. In column two, the regression coefficients of the association between the PGS of neuroticism, depression, and subjective well-being, obtained via GWAS are displayed.

Table 2. OLS-regression results displaying the association between years of education and the PGS for neuroticism.

	Years of Education
PGS Neuroticism MTAG	-0.080***
	(0.027)
PGS Educational Attainment GWAS	0.696***
	(0.026)
PC	Y
Covariates	Y
Constant	13.798
	(0.094)
Observations	8,565
R^2	0.112

*Standard error is reported between brackets. The entire table is reported in the appendix in table A1. With significance being displayed as: * p < 0.05, ** p < 0.01, *** p < 0.005.*

Inspecting the results, it is apparent that there is a statistically significant association between the PGS of neuroticism, obtained via MTAG. It is also shown that this relationship appears to be negative, thus someone having a higher PGS score for mental health problems, will obtain less years of education, while controlling for the PGS of educational attainment. The difference in R^2 obtained consists of 0.1% compared to the base model, which is described in Table A2 in the appendix.

Yet, one of the main problems which might occur, described by Turley et al. (2018), is that this effect can be caused by one of the other PGS, analysed via GWAS, which could be represented in the PGS score analysed via MTAG. Thus to be sure this effect is not caused by any of the other traits analysed, a separate model has been made in column two, representing the individual PGS, analysed via GWAS, which were included in the PGS of the MTAG. These results can be found in table 3.

Table 3. OLS-regression results displaying the association between years of education and the PGS for three different traits, neuroticism, depression, subjective well-being.

	Years of Education (1)	**Years of Education (2)**	**Years of Education (3)**
PGS Neuroticism GWAS	-0.057*		
	(0.026)		
PGS Depression GWAS		-0.054**	
		(0.027)	
PGS Subjective well-being GWAS			0.065**
			(0.026)
PGS Educational Attainment GWAS	0.702***	0.704***	0.703***
	(0.026)	(0.026)	(0.026)
PC	Y	Y	Y
Covariates	Y	Y	Y
Constant	13.799	13.797	13.796
	(0.094)	(0.094)	(0.094)
Observations	8,565	8,565	8,565
R^2	0.111	0.111	0.111

*Standard error is reported between brackets. The entire table is reported in the appendix in table A3. With significance being displayed as: * p < 0.05, ** p < 0.01, *** p < 0.005.*

Observing these results it is important to note the smaller R^2 obtained, as well as a lower significance for different PGS traits. In the rest of this paper, the PGS of neuroticism obtained via MTAG will be used, since this value represents real life better, as well has slightly more statistical power (Hyman, 2000).

Next, the second hypothesis will be looked at. This hypothesis mentions a relationship between the childhood SES someone grew up in. To make this analysis more representable, the continuous childhood SES variable was divided up into three categories, based on the tertiles

of this variables. These three tertiles were then renamed to: low, medium, or high. So the high childhood SES variable, represents the top third of the dataset. To analyse this relationship, an OLS regression analysis was used. These results are displayed in table 3 below.

Table 4. OLS-regression results displaying the association between years of education and different levels of childhood SES.

	Years of Education
Childhood SES	
Medium	0.743***
	(0.063)
High	1.859***
	(0.064)
PGS Educational Attainment GWAS	0.602***
	(0.025)
PC	Y
Covariates	Y
Constant	12.908
	(0.100)
Observations	8,565
R^2	0.193

*Standard error is reported between brackets. The entire table is reported in the appendix in table A4. With significance being displayed as: * p < 0.05, ** p < 0.01, *** p < 0.005.*

To begin interpreting these results, it is important to note that the constant represents the value of the low childhood SES. Thus the coefficients shows for medium and high childhood SES represent these values, compared to the low childhood SES group. It shows a statistical significant result for each different group of childhood SES, with an increase of 0.743 in years, or about nine months of education for medium childhood SES, and an increase of 1.859 in years, or about 22 months of education for high childhood SES, added onto the constant of 12.908.

To find the result for the interaction effect, first a model including a general interaction effect, existing of the childhood SES score, multiplied by the PGS for neuroticism. This creates an continuous variable, This is displayed in table 5 below.

Table 5. OLS-regression results displaying the association between years of education and the PGS for neuroticism, childhood SES and the interaction effect of these two variables.

	Years of Education
PGS Neuroticism MTAG	-0.038
	(0.028)
Childhood SES	
Medium	0.742***
	(0.063)
High	1.854***
	(0.063)
Interaction SES	-0.014
	(0.029)
PGS Educational Attainment GWAS	0.594***
	(0.025)
PC	Y
Covariates	Y
Constant	12.909
	(0.101)
Observations	8,565
R^2	0.194

*Standard error is reported between brackets. The entire table is reported in the appendix in table A5. With significance being displayed as: * p < 0.05, ** p < 0.01, *** p < 0.005.*

It is apparent that the different levels of childhood SES still obtain a statistically significant result, while the genetic risk score for neuroticism is no longer statistically significant. Besides this, the interaction term was not significant.

Finally, to see if there is a statistically significant difference between coefficients of each childhood SES group, three regression analysis have been done. This is displayed in table 6 . Each column represents a different childhood SES, with the first column being the lowest group, and the third column being the highest group. What is apparent is that the observation sizes for these groups are also smaller than those compared in earlier analysis. This is because the childhood SES variable was divided into three groups, and since this was done by inspecting tertiles, the three groups should be of about the same size. The purpose of this table is to compare the results of the coefficients, by observing the 95% confidence intervals. Keeping in mind the results of table 5, it is expected for this not to be statistically significant.

Table 6. OLS-regression results displaying the association between years of education and different PGS for different traits, split by levels of childhood SES.

	Years of Education (Low childhood SES)	**Years of Education (Medium childhood SES)**	**Years of Education (High childhood SES)**
PGS Neuroticism MTAG	-0.010 [-0.104 - 0.083]	-0.082 [-0.166 – 0.002]	-0.047 [-0.127 – 0.033]
PGS Educational Attainment GWAS	0.601*** (0.047)	0.561*** (0.043)	0.620*** (0.039)
PC	Y	Y	Y
Covariates	Y	Y	Y
Constant	12.768 (0.178)	13.408 (0.157)	15.161 (0.142)
Observations	2,854	2,865	2,846
R^2	0.077	0.082	0.102

*Standard error is reported between brackets. The 95% confidence interval is reported in between squared brackets, this is only reported for the PGS for neuroticism. The entire table is reported in the appendix in table A6. With significance being displayed as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$.*

Examining the results, it is apparent that there is no relationship between the genetic risk for neuroticism and years of education, when separating by childhood SES. Observing for any significant differences between coefficients, the 95% confidence intervals for the PGS for neuroticism, analysed via MTAG, are taken into account. This interval for low childhood SES ranges from -0.104 to 0.083, for medium childhood SES this ranges from -0.166 to 0.002, and for high childhood SES this ranges from -0.127 to 0.033. Looking just at the ranges of these confidence intervals, it is shown that these confidence intervals overlap, thus can be concluded that no statistically significant difference is apparent between the coefficients of these three different SES groups. This demonstrates that there is no interaction effect between the risk of neuroticism and childhood SES, which could be related to years of education.

5. Discussion

To see if a genetic risk of mental health issues, affected by childhood socio-economic status, influences years of education, the obtained results will be discussed taking into account the three proposed hypotheses, with the first two hypothesis firstly focussing on the effect of genetic risk of neuroticism, and secondly socio-economic status on years of education. While the third investigates an interaction between these two.

For the first hypothesis, observing the first table with results, it shows a statistical significant result for the genetic risk for neuroticism, obtained via MTAG. This result is a negative coefficient, thus demonstrating a negative association between this and the years of education someone obtained.

To interpret this result in more detail, someone with a PGS for neuroticism being one, would get 0.080 less of a year of education, compared to someone with a PGS score of zero. This result of 0.080 of a year would equate to 29 days. When comparing the R^2 with a model without the genetic risk score for neuroticism, found in table A2 in the appendix, this would lead to an increase of 0.001, thus give a weak association, given the PGS adds only 0.1% of the variance in years of education obtained. These results would confirm the first hypothesis, which offered that an increase in risk in mental health problems correlates with a decrease in years of education.

When observing the second table, diving deeper into the different traits the score for neuroticism is based on, it yields an statistically significant result. All three PGS, obtained via GWAS, are all statistically significant. With a side note that the significance, as well as the R^2 obtained are lower than in the model which made use of the PGS for neuroticism, obtained via MTAG. It shows that the genetic risk scores for neuroticism, as well as depression have a negative impact on years of education obtained, while the genetic risk for subjective well-being obtained a positive coefficient. It is important to note that a genetic risk for mental illness does not have to be caused by one subset of genes. Multiple genes can influence the chances of a mental disorder developing (Hyman, 2000). Thus it seems more like that the PGS for neuroticism, obtained via MTAG, represents the reality of gene interaction better, thus this score will keep being used in the rest of this paper.

Concluding there is indeed an association between the genetic risk for mental illness, and a reduction in years of education. One important thing to note, is that the PGS for neuroticism, analysed via MTAG, range from -3.856 and 3.875, as shown in the descriptive statistics in table 1. So although there is a negative relationship shown, with low genetic risk scores, and a low negative coefficient, the part of year someone would lose by having the highest genetic risks score in this dataset, would result in a loss of 0.308 part of a year of education, equating to roughly 112 days. On the contrary someone with a low genetic risk score would result in an increase of 0.310 part of a year of education, equating to roughly 113 days. Comparing these values with the mean of 13.520 years of education, these differences almost seem negligible for real world situations.

One important thing to take into account, is the possibility of omitted variable bias. It could be the case that there are reasons not taken into account, which could decrease drop-out rate, thus decreasing total years of education. One beforementioned research by Jordan et al. (1996) mentioned some of these causes. These causes existed mostly of being bullied at school, or feeling no connection to the institution someone is attending their education at. These causes are not taken into account in this analysis, which could yield a slightly different result.

For the second hypothesis, examining the second part of the results, the parts describing the influence of childhood SES on years of education obtained. Observing the statistical significance of these results, it shows that people who enjoyed a higher SES during childhood, also had more years of education in total. This while keeping in mind that childhood SES, created by Vable et al. (2017), was based of financial capital during childhood, and human capital, taking into account parental years of education. Because this childhood SES coefficient

is scored by taking into account all these factors, it is difficult to link a direct association between one specific factor and years of education.

A study done by Chevalier et al. (2013) also demonstrated an effect between parental education and financial constraints on drop-out rates in the United Kingdom. It showed that maternal education had a bigger effect on kids staying in school, compared to paternal education. So however which specific childhood SES attribute contributing to an increase in years of education, a strong global association between childhood SES and years of education can be made. This confirms the second hypothesis stating that growing up in a higher SES increases the amount of years of education. One could argue that the average years of education obtained are in the group consistent of the medium childhood SES group, which would mean that the average years of education consists of 13.574 years of education. With being part of the low childhood SES group, obtaining only 12.685 years of education, and being part of the high childhood SES group obtaining 14.835 years of education. Thus showing a difference of almost 11 months with the low childhood SES group, and about 13 months with the high childhood SES group.

By having demonstrated that there is both a relationship between years of education and the PGS for neuroticism, obtained via MTAG, and an association between years of education and the SES someone grew up in, and thus confirming both the first and second hypothesis. Considering the third hypothesis, it argues that for different levels of childhood SES, different coefficients will be found for the genetic risk for neuroticism. Besides this, it argues that with a higher childhood SES the genetic risk for neuroticism would be lower, compared to a child growing up in a lower childhood SES.

When discussing childhood SES, one could argue for endogeneity between the variables used to construct the childhood SES score, and unobserved parental characteristics which could influence this SES score. Thus decreasing a proper causal interpretation.

For the third hypothesis, observing the last section of the results, no statistically significant result can be found for the interaction effect between childhood SES and the PGS for neuroticism. It is also interesting to note that the coefficient for the genetic risk for neuroticism also is no longer statistically significant. When observing the R^2 of this model, the difference is only 0.001 higher when comparing it with the model which just has the childhood SES included. When examining the models split up by childhood SES group, the coefficients obtained for the PGS for neuroticism are also not statistically significant. Additionally, all models also obtaining

a lower R^2 than the control model. And finally, the 95% confidence intervals also overlap between the three coefficients of the model, thus there is no statistically significant difference of the effect of the genetic risk for neuroticism in these three different childhood SES groups.

This would mean the third hypothesis is rejected, since no statistically significant results could be found. Yet, an association still exists between the genetic risk of neuroticism and a reduction in years of education. As well as a association between childhood SES and years of education.

6. Conclusion

Looking back at the research question if a genetic risk of mental health issues, affected by childhood socio-economic status, does influence years of education. One could answer positively to the first part of the question. As established by the first hypothesis being accepted, a genetic risk of mental health does correlate with years of education by a negative amount, even if it is a minor amount compared as an absolute value. Also considering the second hypothesis being accepted, one could also find a relationship between the childhood SES someone grew up in, and the amount of years of education. Yet when answering the second part of the question, talking about the interaction of childhood SES, and the genetic risk of mental health, no statistically significant results can be found. This result also rejects the proposed third hypothesis, where mentioned that being in a higher SES during childhood could reduce the effect of the influence of the genetic risk for mental illness on years of education. Thus overall, it can be concluded that the genetic risk of mental health issues developing is not affected by childhood socio-economic status, and thus does not have a relationship with years of education obtained. While observing both the genetic risk for neuroticism and childhood SES separately, a relationship can be found. It is possible that that for future research, better polygenic scores for neuroticism can be obtained, increasing statistical power obtained.

One of the main limitations of this study is the existence of omitted variable bias. However lots of factors included in the childhood SES variable are taken into account, one aspect is not controlled for. As mentioned by Jordan et al. (1996), where they discuss factors which increase the amount of people dropping out. Next to childhood SES, these factors also include bullying and not feeling connected with the educational institution. These factors could negatively affect years of education followed by someone. Thus have an positive effect on the found coefficients for the genetic risk of neuroticism. For future research, it is recommend to also include factors like these to decrease omitted variable bias.

Another limitation that could be argued is that no causality can be concluded from this research, since it has only been based of associations. For future research it would be recommend to set-up a Mendelian randomized model (Emdin et al., 2017). Thus being able to establish proper causality over the associations currently obtained.

When considering any policy implications, taking into the results obtained, a suggested policy implication could be for people who have a genetic risk for neuroticism, could be to pay extra attention to any signs of any mental illness. On the other hand, the effect of the genetic risk for neuroticism is almost negligible. So even if people were not to take this into account, they years of education obtained would not differ by much. Next to this, discussing childhood SES, a higher childhood SES also equates to a higher years of education obtained. Schools could focus on these children in lower SES groups, providing them with perhaps more educational support, or even helping them out financially. Since no interaction effect between the genetic risk of neuroticism and childhood SES was observed, the recommended suggestion mentioned in the introduction, by providing psychologists to children from low SES would not have proper justification.

7. References

Chevalier, A., Harmon, C., O'Sullivan, V., & Walker, I. (2013). The impact of parental income and education on the schooling of their children. *IZA Journal of Labor Economics*, *2*(1), 1-22.

Downey, G., & Coyne, J. C. (1990). Children of depressed parents: an integrative review. *Psychological bulletin*, *108*(1), 50.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, *9*(3), e1003348.

Emdin, C. A., Khera, A. V., & Kathiresan, S. (2017). Mendelian randomization. *Jama*, *318*(19), 1925-1926.

Fletcher, J. M. (2010). Adolescent depression and educational attainment: results using sibling fixed effects. *Health economics*, *19*(7), 855-871.

Gilman, S. E., Kawachi, I., Fitzmaurice, G. M., & Buka, S. L. (2002). Socioeconomic status in childhood and the lifetime risk of major depression. *International journal of epidemiology*, *31*(2), 359-367.

Hyman, S. E. (2000). The genetics of mental illness: implications for practice. *Bulletin of the World Health Organization*, *78*, 455-463.

Jordan, W. J., Lara, J., & McPartland, J. M. (1996). Exploring the causes of early dropout among race-ethnic and gender groups. *Youth & Society*, *28*(1), 62-94.

Krapohl, E., Patel, H., Newhouse, S., Curtis, C. J., von Stumm, S., Dale, P. S., ... & Plomin, R. (2018). Multi-polygenic score approach to trait prediction. *Molecular psychiatry*, *23*(5), 1368-1374.

Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist*, *64*(4), 241.

Liu, C. H., Stevens, C., Wong, S. H., Yasui, M., & Chen, J. A. (2019). The prevalence and predictors of mental health diagnoses and suicide among US college students: Implications for addressing disparities in service use. *Depression and anxiety*, *36*(1), 8-17.

Nugent, N. R., Tyrka, A. R., Carpenter, L. L., & Price, L. H. (2011). Gene–environment interactions: early life stress and risk for depressive and anxiety disorders. *Psychopharmacology*, *214*(1), 175-196.

Papageorge, N. W., & Thom, K. (2020). Genes, education, and labor market outcomes: evidence from the health and retirement study. *Journal of the European Economic Association*, *18*(3), 1351-1399.

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., ... & Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics*, *50*(2), 229-237.

Vable A. M., Gilsanz P., Nguyen T. T., Kawachi I. & Glymour M. M. (2017). Validation of a theoretically motivated approach to measuring childhood socioeconomic circumstances in the Health and Retirement Study. *PLoS ONE, 12*(10): e0185898. https://doi.org/10.1371/journal.pone.0185898

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), 7-24.

Walpole, M. (2003). Socioeconomic status and college: How SES affects college experiences and outcomes. *The review of higher education*, *27*(1), 45-73.

Wang, M. T., & Sheikh-Khalil, S. (2014). Does parental involvement matter for student achievement and mental health in high school?. *Child development*, *85*(2), 610-625.

Health and Retirement Study data accessed via:

Polygenic Score Data. (2021). *Health and Retirement Study*. Accessed via: https://hrsdata.isr.umich.edu/data-products/polygenic-score-data-pgs

RAND HRS Longitudinal File 2018. (2021). *Health and Retirement Study*. Accessed via: https://hrsdata.isr.umich.edu/data-products/rand-hrs-longitudinal-file-2018

8. Appendix

A1 Acronyms

GWAS – Genome Wide Association Study

MTAG – Multi-Trait Analysis of a Genome wide association study

PC – Principle Components

PGS – Polygenic Score

SES – Socio-Economic Status

A2 Tables

Table A1. OLS-regression results displaying the association between years of education and Neuroticism PGS, obtained via MTAG.

	Years of Education
PGS Neuroticism MTAG	-0.080***
	(0.027)
PGS Educational Attainment GWAS	0.696***
	(0.026)
PC	
1	4.099
	(2.314)
2	-2.750
	(2.364)
3	6.838***
	(2.272)
4	2.544
	(2.431)
5	1.007
	(2.371)
6	3.363
	(2.438)
7	-0.210

	(2.439)
8	3.858
	(2.371)
9	-2.989
	(2.454)
10	-2.603
	(2.406)
Covariates	
Birthyear	0.041***
	(0.009)
Gender	-0.402***
	(0.054)
*Birthyear*Gender*	0.001
	(0.005)
Birthyear²	0.000
	(0.000)
Constant	13.798
	(0.094)
Observations	8,565
R^2	0.112

*Standard error is reported between brackets. With gender being a binary variable, 1 for males and 2 for females. Birthyear is the centred birthyear, obtained by subtracting the average birthyear of the entire dataset with someone their actual birthyear. With significance being displayed as: * p < 0.05, ** p < 0.01, *** p < 0.005.*

Table A2. OLS-regression results displaying the association between years of education and the control variables.

	Years of Education
PGS Educational Attainment GWAS	0.712*** (0.026)
PC	
1	4.002 (2.317)
2	-2.564 (2.366)
3	6.619*** (2.274)
4	2.467 (2.429)
5	1.116 (2.372)
6	3.309 (2.438)
7	-0.411 (2.438)
8	3.899 (2.375)
9	-3.035 (2.453)
10	-2.626 (2.408)
Covariates	
Birthyear	0.041*** (0.009)
Gender	-0.401*** (0.054)
*Birthyear*Gender*	0.001 (0.005)
Birthyear²	0.000 (0.000)

Constant	13.798 (0.094)
Observations	8,565
R^2	0.111

*Standard error is reported between brackets. With gender being a binary variable, 1 for males and 2 for females. Birthyear is the centred birthyear, obtained by subtracting the average birthyear of the entire dataset with someone their actual birthyear. With significance being displayed as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$.*

Table A3. OLS-regression results displaying the association between years of education and different PGS for different traits.

	Years of Education (1)	**Years of Education (2)**	**Years of Education (3)**
PGS Neuroticism GWAS	-0.057*		
	(0.026)		
PGS Depression GWAS		-0.054**	
		(0.027)	
PGS Subjective well-being GWAS			0.065**
			(0.026)
PGS Educational Attainment GWAS	0.702***	0.704***	0.703***
	(0.026)	(0.026)	(0.026)
PC			
1	4.164	3.643	4.430
	(2.314)	(2.326)	(2.327)
2	-2.675	-2.697	-2.603
	(2.365)	(2.365)	(2.365)
3	6.738***	6.781***	6.655***
	(2.273)	(2.273)	(2.273)
4	2.475	2.534	2.545
	(2.430)	(2.428)	(2.433)
5	1.087	0.983	1.083
	(2.371)	(2.373)	(2.372)
6	3.345	3.340	3.356
	(2.438)	(2.438)	(2.439)
7	-0.238	-0.338	-0.372
	(2.441)	(2.439)	(2.437)
8	3.861	3.833	3.957
	(2.372)	(2.373)	(2.374)

9	-3.077	-3.021	-2.880
	(2.454)	(2.453)	(2.454)
10	-2.640	-2.543	-2.640
	(2.406)	(2.408)	(2.408)
Covariates			
Birthyear	0.041***	0.041***	0.041***
	(0.009)	(0.009)	(0.009)
Gender	-0.402***	-0.401***	-0.401***
	(0.054)	(0.054)	(0.054)
*Birthyear*Gender*	0.001	0.001	0.001
	(0.005)	(0.005)	(0.005)
Birthyear²	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Constant	13.799	13.797	13.796
	(0.094)	(0.094)	(0.094)
Observations	8,565	8,565	8,565
R^2	0.111	0.111	0.111

*Standard error is reported between brackets. With gender being a binary variable, 1 for males and 2 for females. Birthyear is the centred birthyear, obtained by subtracting the average birthyear of the entire dataset with someone their actual birthyear. With significance being displayed as: * p < 0.05, ** p < 0.01, *** p < 0.005.*

Table A4. OLS-regression results displaying the association between years of education and different PGS for different levels of childhood SES.

	Years of Education
Childhood SES	
Medium	0.743***
	(0.063)
High	1.859***
	(0.064)
PGS Educational Attainment GWAS	0.602***
	(0.025)
PC	
1	7.004***
	(2.239)
2	-3.139
	(2.250)
3	5.973***
	(2.204)
4	1.629
	(2.297)
5	2.346
	(2.280)
6	3.434
	(2.326)
7	-0.007
	(2.322)
8	3.200
	(2.278)
9	-1.149
	(2.325)
10	-3.394
	(2.281)

Covariates	
Birthyear	0.017**
	(0.009)
Gender	-0.388***
	(0.051)
*Birthyear*Gender*	0.005
	(0.005)
Birthyear²	0.000
	(0.000)
Constant	12.908
	(0.100)
Observations	8,565
R^2	0.193

*Standard error is reported between brackets. With gender being a binary variable, 1 for males and 2 for females. Birthyear is the centred birthyear, obtained by subtracting the average birthyear of the entire dataset with someone their actual birthyear. With significance being displayed as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$.*

Table A5. OLS-regression results displaying the association between years of education and different PGS for different traits, split by levels of childhood SES.

	Years of Education
PGS Neuroticism MTAG	-0.038
	(0.028)
Childhood SES	
Medium	0.742***
	(0.063)
High	1.854***
	(0.063)
Interaction SES	-0.014
	(0.029)
PGS Educational Attainment GWAS	0.594***
	(0.025)
PC	
1	7.080***
	(2.236)
2	-3.243
	(2.251)
3	6.079**
	(2.204)
4	1.688
	(2.298)
5	2.280
	(2.280)
6	3.460
	(2.326)
7	0.097
	(2.324)

8	3.182
	(2.276)
9	-1.111
	(2.326)
10	-3.387
	(2.281)
Covariates	
Birthyear	0.017
	(0.009)
Gender	-0.388***
	(0.051)
*Birthyear*Gender*	0.005
	(0.005)
Birthyear²	0.000
	(0.000)
Constant	12.909
	(0.101)
Observations	8,565
R^2	0.194

*Standard error is reported between brackets. With gender being a binary variable, 1 for males and 2 for females. Birthyear is the centred birthyear, obtained by subtracting the average birthyear of the entire dataset with someone their actual birthyear. The interaction SES score is obtained by multiplying the childhood SES score with the PGS of neuroticism, thus creating a continuous variable. With significance being displayed as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$.*

Table A6. OLS-regression results displaying the association between years of education and different PGS for different traits, split by levels of childhood SES.

	Years of Education (Low childhood SES)	**Years of Education (Medium childhood SES)**	**Years of Education (High childhood SES)**
PGS Neuroticism MTAG	-0.010	-0.082	-0.047
	(0.048)	(0.043)	(0.041)
PGS Educational Attainment GWAS	0.601***	0.561***	0.620***
	(0.047)	(0.043)	(0.039)
PC			
1	13.161***	7.135	-1.496
	(3.711)	(3.870)	(4.087)
2	-4.446	-3.457	-0.540
	(4.315)	(3.901)	(3.441)
3	2.815	9.408*	4.602
	(4.447)	(3.637)	(3.480)
4	4.097	1.842	-0.320
	(4.503)	(3.903)	(3.486)
5	-4.453	2.365	9.122*
	(4.307)	(3.770)	(3.700)
6	4.225	7.941*	-1.328
	(4.337)	(3.973)	(3.761)
7	-7.092	5.457	1.804
	(4.415)	(3.919)	(3.738)
8	9.706*	-1.425	1.960
	(4.290)	(3.856)	(3.674)
9	-3.616	-1.812	1.328
	(4.536)	(3.721)	(3.734)
10	1.604	-5.333	-6.702
	(4.295)	(3.836)	(3.665)

Covariates			
Birthyear	0.010	0.039	-0.006
	(0.017)	(0.016)	(0.014)
Gender	-0.300***	-0.263***	-0.595***
	(0.100)	(0.088)	(0.084)
*Birthyear*Gender*	0.016	-0.005	0.012
	(0.010)	(0.009)	(0.008)
Birthyear²	0.000	0.001	0.000
	(0.000)	(0.000)	(0.000)
Constant	12.768	13.408	15.161
	(0.178)	(0.157)	(0.142)
Observations	2,854	2,865	2,846
R^2	0.077	0.082	0.102

*Standard error is reported between brackets. With gender being a binary variable, 1 for males and 2 for females. Birthyear is the centred birthyear, obtained by subtracting the average birthyear of the entire dataset with someone their actual birthyear. With significance being displayed as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$.*