Group B2
Aidan Meharg 20112579
Buyun Wang 59447243
Florence Wang 39333620
Chenyang Mao 89445100

Investigation of Social and Demographic Factors Leading to Mathematical Success

Introduction:
Math is one of the most important subjects in high school. A good conceptual understanding of math is essential to study engineering, computer science, statistics etc in the future. Previous research has shown that family household income has some effect on math ability at a young age (Lombardi & Dearing, 2021). Therefore, we are interested in if other demographic and social factors affect students' math grades. In this project, we use data collected by Cortez and Silva (2008) to investigate how demographic and social features affect secondary school math performance. We also investigate the explanatory power of previous math grades on final grade.

In this project, we will fit three models, the first two are against the demographic and social features, while the last one was based on previous math grades for the same year.

Explanatory variables

- Model I (full model): sex, age, family size, mother education, father education, mother job, father job, travel time and the quantity of family relationship.
- Model II (only sex and age): sex, age
- Model III: (just to investigate the effect of previous performance) G1, G2

Response variable

- For all three models, the response variable is G3.

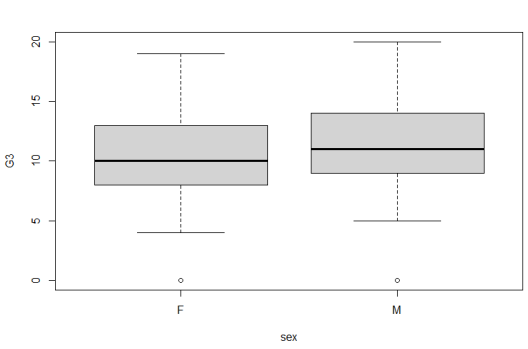
The variables (both explanatory and Response) in the data set are listed below:

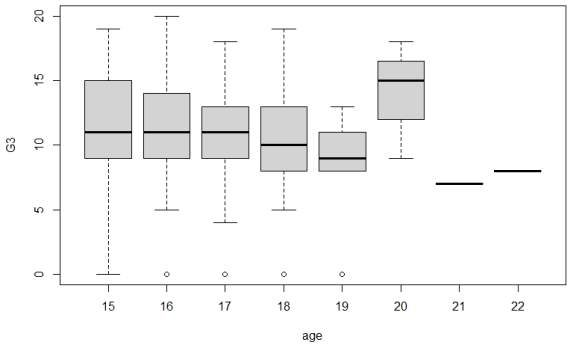
Attribute	Description
sex	student sex (categorical: Female or Male)
age	student age (numeric: from 15 to 22)
Medu	mother's education (categorical: 0 or 1 or 2 or 3 or 4, 0 is none, 1 is primary education (4th grade) , 2 is 5th to 9th grade, 3 is secondary education or 4 is higher education)
Fedu	father's education (categorical: 0 or 1 or 2 or 3 or 4, 0 is none, 1 is primary education (4th grade) , 2 is 5th to 9th grade, 3 is secondary education or 4 is higher education)
Fjob	father's job (categorical: teacher, healthcare related, civil services (e.g. administrative or police), at home or other)
Mjob	mother's job (categorical: teacher, healthcare related, civil services (e.g. administrative or police), at home or other)
famsize	Family size (categorical: smaller or equal to 3 or greater than 3)
famrel	quality of family relationships (categorical: 1 or 2 or 3 or 4 or 5, 1 is really bad while 5 is excellent)
traveltime	home to school travel time (categorical: 1 or 2 or 3 or 4, 1 is < 15 min., 2 is 15 to 30 min, 3 is 30 min - 1 hr, 4 is > 1 hr)
G1	first period grade (numeric: 0 to 20)
G2	second period grade (numeric: 0 to 20)
G3	final grade (numeric: 0 to 20)

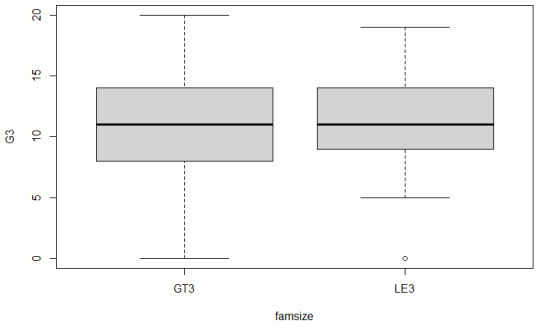
Data visualization

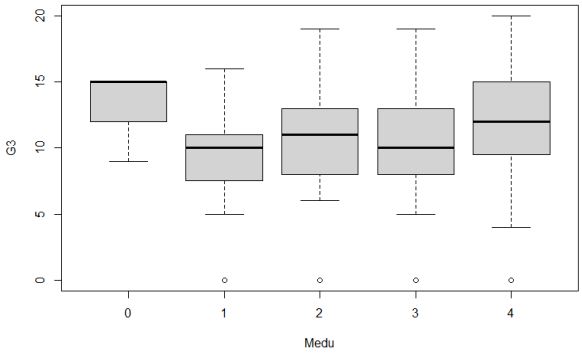
To quickly extract the possible relationships between the response and each of the explanatory variables, we created boxplots on G3 against sex, age, famsize, Medu, Fedu, Mjob, Fjob, traveltime, and famrel respectively as follows:

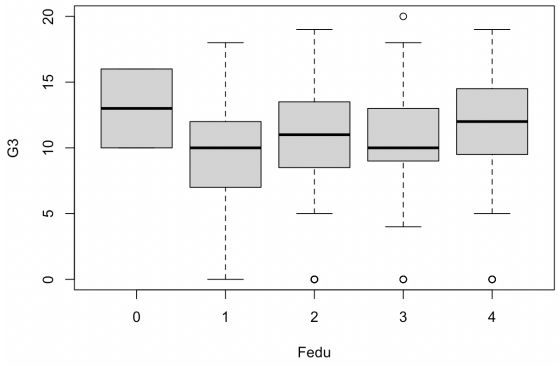
- **Boxplots for each of the chosen demographic and social factors:**

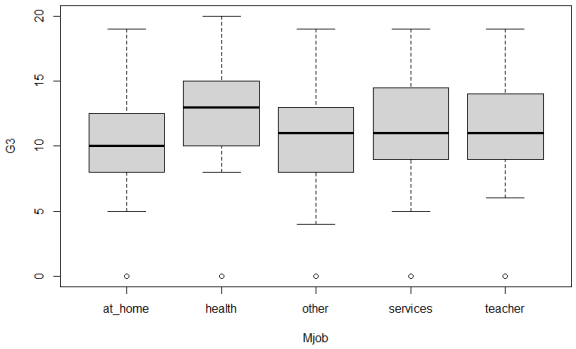


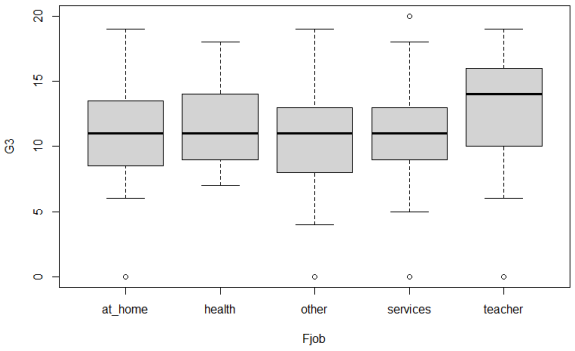


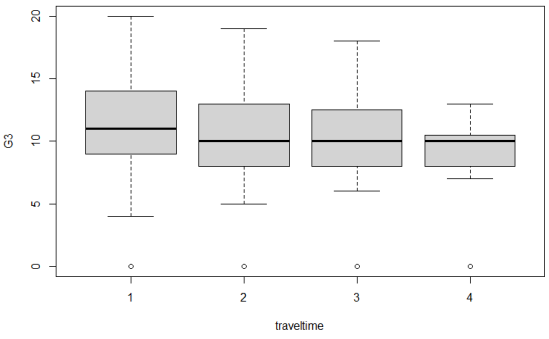


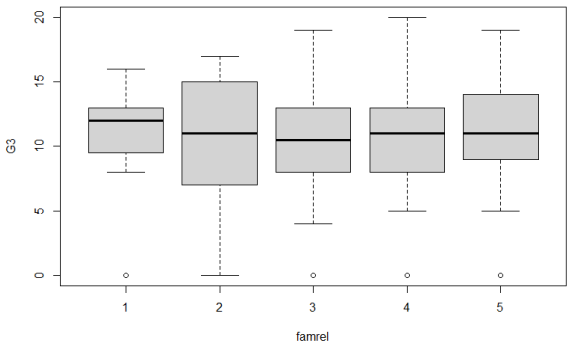




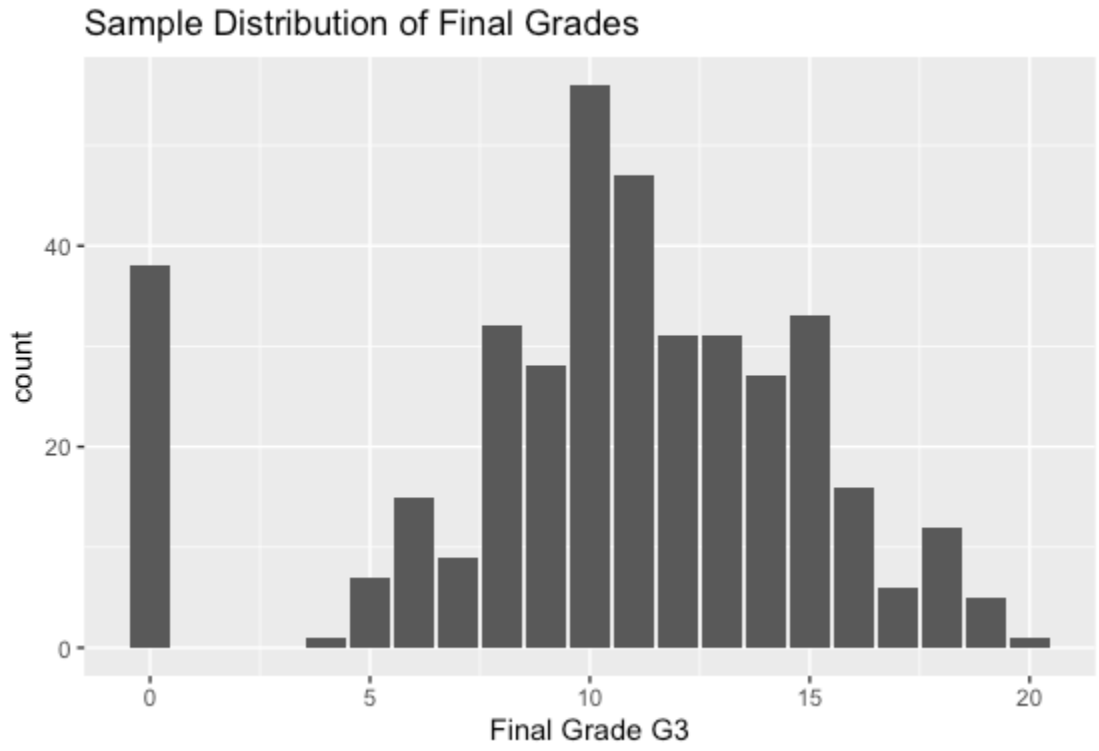








The boxplots do not show significant differences on the effect between each level of most of the demographic and social factors except for the age. But in order to get precise results, we still need to test it by the multiple linear regression test.



Based on this distribution figure of G3, we notice that there is a large subset of students who received a final grade of zero. This suggests that these may be students who did not attend school at all, which is quite a different scenario than attending and receiving a failing grade. Therefore, we also want to compare the results for including/excluding these observations.

Model fitting & Analysis:

Full Model (Model I)

Upon fitting the full model, which includes all the demographic and social factors named in the introduction (all predictors excluding G1 and G2), we found most of the terms to be insignificant. Age, sex, and family size appeared to be significant at the 10% level using the full dataset, so we chose to explore their effects further in a reduced model.
Although some terms were statistically significant, the full model has extremely low R-squared and adjusted R-squared values, indicating that it has very little explanatory power over student final grades.

Full model: including students with a final grade of 0

```
> fit = lm(G3 ~ ., data = d_clean)
> summary(fit)

Call:
lm(formula = G3 ~ ., data = d_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-12.6393  -1.8667   0.5552   2.8388  10.2865

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.00637    5.45853   4.398 1.43e-05 ***
sexM          0.80702    0.46879   1.721   0.0860 .
age          -0.42542    0.18364  -2.317   0.0211 *
famsizeLE3    1.05295    0.51189   2.057   0.0404 *
Medu1        -4.79627    2.67966  -1.790   0.0743 .
Medu2        -3.95651    2.65932  -1.488   0.1377
Medu3        -3.54866    2.70005  -1.314   0.1896
Medu4        -2.05499    2.76129  -0.744   0.4572
Fedu1        -3.43974    3.26664  -1.053   0.2930
Fedu2        -3.41999    3.25896  -1.049   0.2947
Fedu3        -3.39065    3.26849  -1.037   0.3002
Fedu4        -3.61409    3.29571  -1.097   0.2735
Mjobhealth    0.79303    1.15784   0.685   0.4938
Mjobother    -0.12925    0.74346  -0.174   0.8621
Mjobservices  0.57354    0.82947   0.691   0.4897
Mjobteacher  -1.09010    1.09953  -0.991   0.3221
Fjobhealth    0.62317    1.52259   0.409   0.6826
Fjobother    -0.01961    1.08421  -0.018   0.9856
Fjobservices  0.03789    1.12524   0.034   0.9732
Fjobteacher   1.17965    1.41201   0.835   0.4040
traveltime2  -0.69667    0.53347  -1.306   0.1924
traveltime3  -0.52300    1.00271  -0.522   0.6023
traveltime4  -2.05914    1.65475  -1.244   0.2142
famrel2      -1.45370    1.93966  -0.749   0.4541
famrel3      -0.67678    1.69651  -0.399   0.6902
famrel4      -0.25881    1.63056  -0.159   0.8740
famrel5       0.13503    1.65753   0.081   0.9351
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.447 on 368 degrees of freedom
Multiple R-squared:  0.1201,    Adjusted R-squared:  0.05795
F-statistic: 1.932 on 26 and 368 DF,  p-value: 0.004598
```

We also fit this model excluding students with G3 equals zero (summary seen below). It provides a better fit since the adjusted R-square is higher. However, none of the explanatory variables in this model have a p-value that is smaller than 0.05. Thus, it seems that none of these explanatory variables significantly affect the G3 (response variable).

Full model: excluding students with a final grade of 0

```
Call:
lm(formula = G3 ~ ., data = d_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7779 -2.0894 -0.2311  2.0099  7.6183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.94185    3.94119   5.314 1.98e-07 ***
sexM          0.58624    0.34566   1.696   0.0908 .
age          -0.26031    0.13741  -1.894   0.0590 .
famsizeLE3    0.38633    0.37532   1.029   0.3041
Medu1        -2.82238    1.89401  -1.490   0.1371
Medu2        -2.07822    1.88371  -1.103   0.2707
Medu3        -2.05472    1.91474  -1.073   0.2840
Medu4        -1.60695    1.95429  -0.822   0.4115
Fedu1        -2.53403    2.30941  -1.097   0.2733
Fedu2        -2.18000    2.29768  -0.949   0.3434
Fedu3        -2.46949    2.30734  -1.070   0.2853
Fedu4        -2.08889    2.32803  -0.897   0.3702
Mjobhealth    1.19589    0.85578   1.397   0.1632
Mjobother    -0.35991    0.56887  -0.633   0.5274
Mjobservices  0.52975    0.63507   0.834   0.4048
Mjobteacher  -0.44494    0.81339  -0.547   0.5847
Fjobhealth   -1.23044    1.11781  -1.101   0.2718
Fjobother    -0.79967    0.82505  -0.969   0.3331
Fjobservices -0.83932    0.85675  -0.980   0.3280
Fjobteacher   0.81153    1.06852   0.759   0.4481
traveltime2  -0.43918    0.39710  -1.106   0.2695
traveltime3   0.10574    0.76894   0.138   0.8907
traveltime4  -1.40288    1.25481  -1.118   0.2644
famrel2      -0.74851    1.47427  -0.508   0.6120
famrel3      -0.69491    1.28338  -0.541   0.5886
famrel4      -0.57985    1.23140  -0.471   0.6380
famrel5       0.03368    1.25028   0.027   0.9785
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.127 on 330 degrees of freedom
Multiple R-squared:  0.1303,    Adjusted R-squared:  0.06174
F-statistic: 1.901 on 26 and 330 DF,  p-value: 0.005852
```

Only sex and age (Model II):

In model II, both sex and age appear to have a statistically significantly effect on final grade. There may be some collinearity for both two variables since both p-value are decreased according to model I. However, the adjusted R square is lower than model I (full model), which indicates an even worse fit than the full model.

Model II is of the form:

$$G3 = \beta_0 + \beta_1 male + \beta_2 age + \varepsilon$$

The obtained least squares coefficients (seen below) indicate that final grade (G3) tends to decrease with increasing student age. Male students tended to obtain grades 0.9065 points higher on average than female students in the same age category.

Model II: including students with a final grade of 0

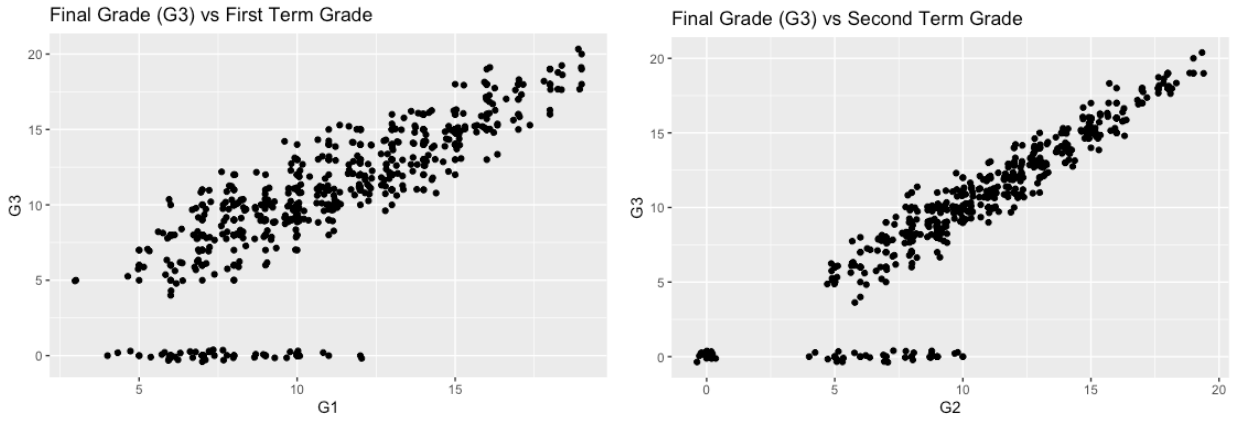
```
> fit = lm(G3 ~ sex + age, data = d_clean)
> summary(fit)

Call:
lm(formula = G3 ~ sex + age, data = d_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-11.8593  -1.8593   0.2806   3.1407   9.7571

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.5024     2.9966   6.508 2.33e-10 ***
sexM          0.9065     0.4547   1.994  0.04688 *
age          -0.5700     0.1781  -3.200  0.00149 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.51 on 392 degrees of freedom
Multiple R-squared:  0.03588,   Adjusted R-squared:  0.03097
F-statistic: 7.295 on 2 and 392 DF,  p-value: 0.0007751
```

We also fit this model excluding students with G3 equals zero. It also provided a worse fit due to the lower adjusted R-square. P-values for both variables increase, and sex becomes less significant.

Model II: excluding students with a final grade of 0

```
>
> fit = lm(G3 ~ sex + age, data = d_clean)
> summary(fit)

Call:
lm(formula = G3 ~ sex + age, data = d_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0932 -2.0873 -0.4392  2.5549  8.2588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.0761     2.2356   7.638 2.08e-13 ***
sexM          0.6422     0.3378   1.901  0.05813 .
age          -0.3519     0.1333  -2.641  0.00864 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.189 on 354 degrees of freedom
Multiple R-squared:  0.02961,   Adjusted R-squared:  0.02413
F-statistic: 5.401 on 2 and 354 DF,  p-value: 0.004893
```

Model III:



Note: the above scatter plots were constructed using ggplot2 with geom_jitter() in order to better see data points that were previously overlapping in the original scatter plot.

Using the subset of students who obtained higher than a zero final grade, we fit a model of the form: $G3 = \beta_0 + \beta_1 G1 + \beta_2 G2 + \varepsilon$

Model III: excluding students with a final grade of 0

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19482    0.16572   1.176 0.240541
G1           0.11167    0.03133   3.564 0.000415 ***
G2           0.88661    0.03226  27.486  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8272 on 354 degrees of freedom
Multiple R-squared:  0.9347,    Adjusted R-squared:  0.9343
```

As one might expect, the previous math performance of students (G1 and G2) provided a good fitting model for final grade (G3). We see that G2 has much more influence over the final grade than G1. In this case, excluding students who obtained a final grade of zero significantly improved the fit of the model. The summary of the model fitted from the full dataset can be seen below with a drop in adjusted R-squared of over 0.1:

Model III: including students with a final grade of 0

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.83001    0.33531  -5.458 8.57e-08 ***
G1           0.15327    0.05618   2.728  0.00665 **
G2           0.98687    0.04957  19.909  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.937 on 392 degrees of freedom
Multiple R-squared:  0.8222,    Adjusted R-squared:  0.8213
F-statistic: 906.1 on 2 and 392 DF,  p-value: < 2.2e-16
```

Conclusion:

It appears that most of the demographic and social features that we investigated do not appear to have strong effects on secondary school math performance. In the full model, only age and family size appeared to have significant effects on final grade at the 5% level.

The sex of students appeared to have a moderate effect on final grade in our reduced model. Male students achieved on average 0.9481 points higher than female students of the same age with respect to the final grade (p-value 0.0399). That being said, both full model and reduced model had very little explanatory power over final grades due to extremely low (adjusted) R-squared values.

By using previous math grades (in the same year) as our predictors we were able to achieve a very good model fit, although this result is somewhat intuitive.

Mathematical success rates are likely explained well by other factors that were not available to us (ex. Quality of instruction, Family income, early education factors, and possibly genetic factors).

Limitations:

The students in this study were volunteers and also most of the demographic and social variables were collected by means of a questionnaire. The responses may not accurately reflect the real conditions due to both response bias (results based on questionnaire) and non-response bias (voluntary to attend this study). Also note that this study sampled students from Portugal, which may very well not be representative of students in other areas of the world.

References:

Cortez, P., Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Lombardi, C. M., & Dearing, E. (2021). Maternal support of Children's math learning in associations between family income and math school readiness. *Child Development, 92*(1), e39-e55. https://doi.org/10.1111/cdev.13436