Task 1

Introduction

This analysis is based on a study focusing on the influence of openness on the subjective wellbeing of college students. The researchers carried out a correlational study design in which they collected responses from 211 students who filled out a questionnaire containing 11 items, four of which related to openness, four to self-compassion, and four to wellbeing. This study may be relevant because self-compassion can be raised by each individual and if its mediating role is found, this study would add a tool to the disposal of individuals to increase their wellbeing.

Research question

RQ. 1) Does the hypothesised mediation model fit the data?
RQ. 2) Is the relationship between openness and subjective well-being mediated by self-compassion in college students?

Hypotheses

H0. Openness is not predictive of subjective wellbeing, and the hypothesised mediation model does not fit the data.
H1. Openness positively predicts subjective wellbeing.
H2. The relationship between openness and subjective wellbeing is mediated by self-compassion.

Data analysis method and justification

Preliminary analysis

Data accuracy will be verified to ensure the data has been entered correctly, addressing one of the assumptions of path analysis. Data entry errors will have to be coded as missing values. The location and pattern of missing data will then be ascertained. Then, I will test for a potential mechanism in the missing data (Jamshidian et al.'s, 2014; Rubin, 1976) to verify equality of variance. Cases with more than 10% of missing values will be removed, remaining missing values imputed using an adequate method such as mean imputation and outliers, both univariate and multivariate, dealt with before analysis because they can lead to non-normality and bias.Normality of distribution, skewness and kurtosis will be verified to address the normality assumption.

Primary analysis

Multiple regression can only test a single dependent variable at a time and focuses on direct paths, which makes it a poor method for evaluating indirect effects. Path analysis on the other hand can include multiple outcomes to test theoretical models. It tests direct, indirect and total effects at the same time and assesses the model’s fit. It is suited for testing mediation models, which makes it suitable for this research.

Preliminary analyses

The describe() function was used to verify the conformity of data to the appropriate range, 1-7, imposed by the Likert scale-type prompts. This revealed anomalies in the range of the variables open2 and swb1. These data entry errors were replaced with "NA" for a missing value.

The missing data pattern indicated the number of variables that are incomplete in that pattern and the total number of missing values for each variable.

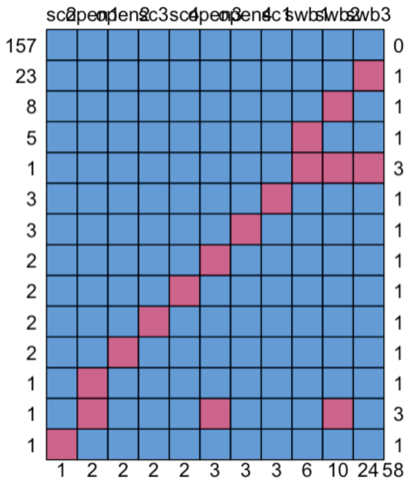


Figure 1. *Output of the md.pattern() function indicating the missing data patterns*

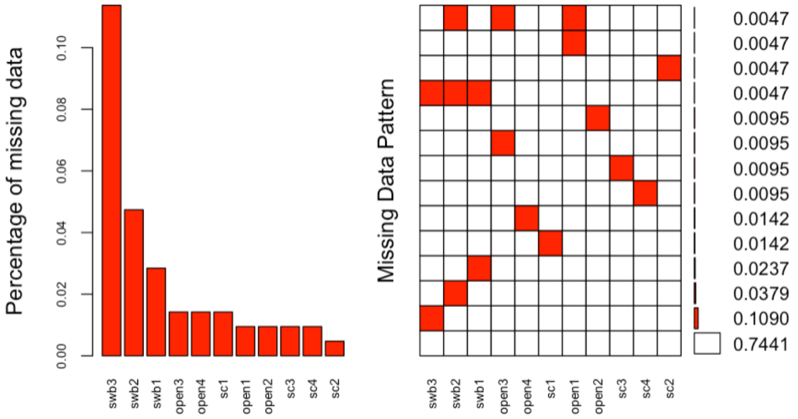


Figure 2. *Output of aggr() indicating the percentage and pattern of missing data for each variable of the dataset. The red cells in the patterns of missing data indicate variables containing missing values.*

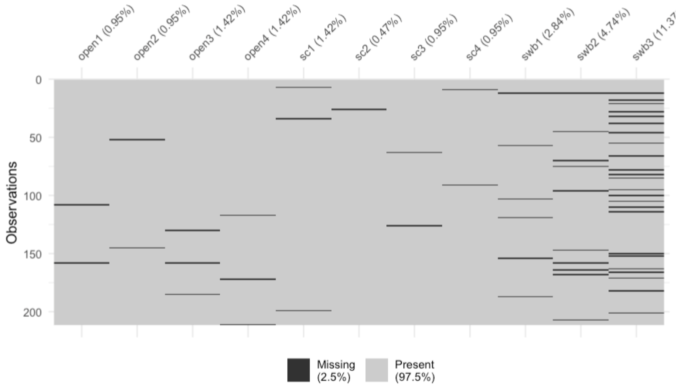


Figure 3. *Output of the vis_miss() function indicating the location of missing data for each variable of the dataset. The dataset only contains 2.5% missing values and is therefore 97.5% complete.*

Image truncated in the output by RStudio.

A potential mechanism in the missing data was then researched. This was done using the TestMCARNormality() function, testing whether the correlations between the variables in the cases with missing data are the same as the correlations between the variables in cases with no missing data. The p-value for the non-parametric test of homoscedasticity was 0.064, larger than 0.05, which allows me to accept the null hypothesis: that the covariances are equal. This indicates that the data are missing completely at random (MCAR): there is no evidence of systematic patterning of the missing data. This is expected as it is typical for correlational studies. I removed subjects with more than 10% of missing data. Two cases were removed.

Mean imputation, which imputes the mean of the available missing items, can place an upper limit on standard errors and assumes that the missing values have the same mean as the observed values. However, it is reasonable to assume the missing values would closely resemble other values from the same scale. In addition, the dataset has data missing completely at random, and the total proportion of missing data in the whole dataset is lower than 10%, so mean imputation is appropriate (Cole, 2008). The scale means were calculated for each variable in the dataset, and mean scale scores were imputed to the missing values.

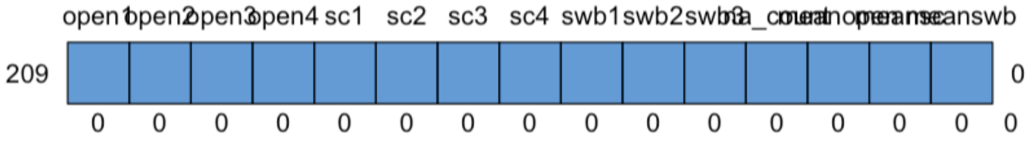
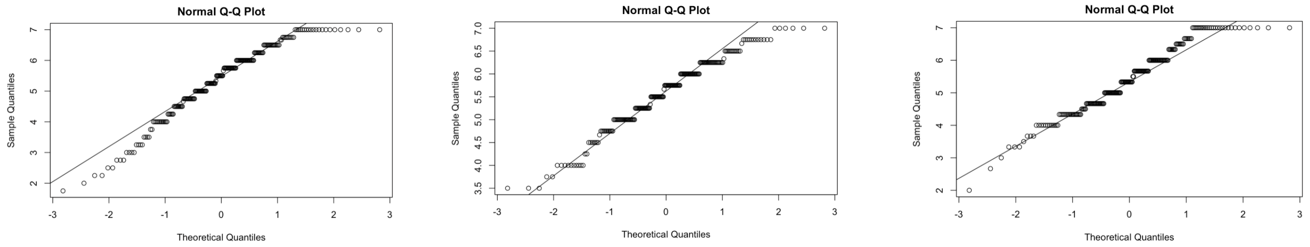


Figure 4. *Output of the md.pattern() function indicating the lack of missing data.*

Univariate outliers were detected by standardising the variables and removing cases with Z-scores larger than 3.29 (Tabachnick et al., 2013). Two cases were removed, leaving a sample of 207 participants. There were no multivariate outliers, as identified using Mahalanobis distances. Skewness: -0.65 (meanopen), -.054 (meansc), and -0.22 (meanswb) and kurtosis: -0.15 (meanopen), -0.01 (meansc), and -0.40 (meanswb) are small values, attesting to the normality of our variables after removing outliers. QQplot verified approximate normality.



Figures 5-7. *QQplot of, in order, mean openness, mean self-compassion and mean wellbeing.*

Primary analyses

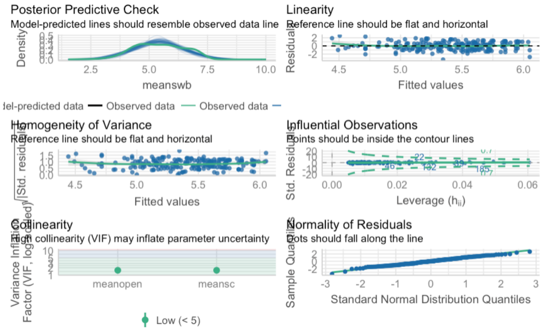


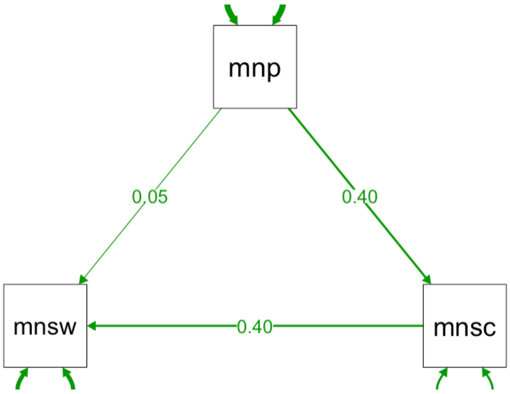
Figure 8. *Posterior check on the assumptions of path analysis. Results of the linear model's diagnostics look good.*

The model's estimated coefficient of meanopen (0.04868) shows 0.04868 increase in average subjective wellbeing for a 1-unit increase in mean openness. This coefficient isn't significant (p =

0.496), meaning that the mean openness variable is not significantly correlated with mean subjective wellbeing. The estimated coefficient of meansc (0.39636) is also positive but is significant (p = 0.000309): increases in mean self-compassion are related to increases in mean subjective wellbeing.

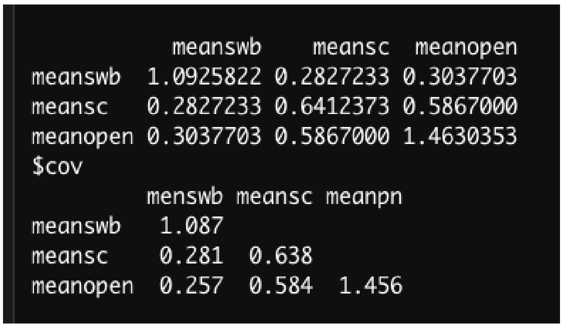
There is a positive correlation between all three variable averages: meanopen and meansc (r = 0.61, a strong correlation), meanopen and meanswb (r = 0.24, weak correlation) and meansc and meanswb (r = 0.34, a moderate correlation). All p-values are lower than 0.05: these correlations are significant.

Figure 9. *Output of the semPaths() function for the just-identified model. The relationships between the variables are sufficiently represented by the indirect effect so I will use Ockham's razor and remove the direct path from openness to wellbeing. Image truncated in the output by RStudio.*



The over-identified model fits the data well: the Tucker-Lewis Index (1.014) and Comparative Fit Index (1) surpass 0.9. The Root Mean Square Error of Approximation is equal to 0 and the Standardised Root Mean Square Residual is 0.015, indicating a low difference from a perfect fit.

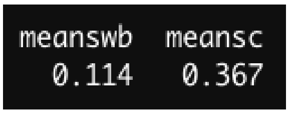
The c path is therefore not needed in the model and can be removed for the sake of parsimony. This fits the finding that the c path had a small effect and was non-significant (b = .05, p > .05) in the just-identified model.



```
            meanswb    meansc  meanopen
meanswb  1.0925822 0.2827233 0.3037703
meansc   0.2827233 0.6412373 0.5867000
meanopen 0.3037703 0.5867000 1.4630353
$cov
         menswb meansc meanpn
meanswb   1.087
meansc    0.281  0.638
meanopen  0.257  0.584  1.456
```

Figure 10. *The model implied and actual correlations are quite similar due to the lack of full information in the over-identified model. The openness-wellbeing pair is the only one that is slightly less similar, which is the case because this represents the c path which was left out, and which the model has to estimate. Overall, this again confirms the strength of this model.*

The 95% CI for the 0.18 indirect effect runs from 0.09 to 0.3, excluding zero. I therefore reject the null hypothesis: self-compassion does indeed mediate the relationship between openness and subjective wellbeing, even when leaving out the direct effect.



```
meanswb  meansc
  0.114   0.367
```

Figure 11. *Around 11% of variance in wellbeing and about 37% of the variance in self-compassion is explained by the path model.*

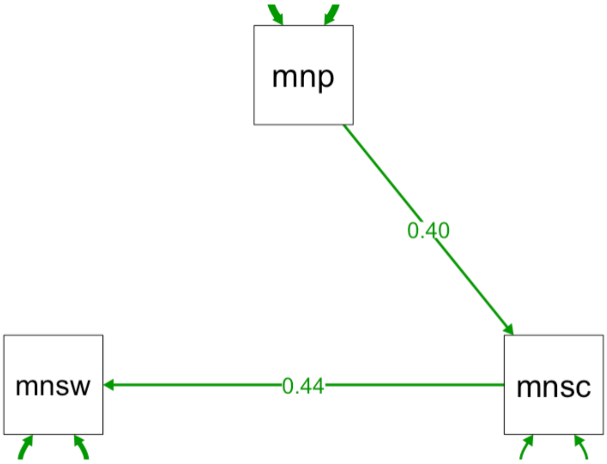


Figure 9. *Output of the semPaths() function for the over-identified model. Image truncated in the output by RStudio.*

Discussion

These results allow me to reject H0: the hypothesised mediation model fits the data and openness is positively predictive of subjective wellbeing. Self-compassion mediates the relationship between openness and subjective wellbeing in college students. The dataset includes a lot of missing data for swb3, “I feel I am flourishing.” which may be because it’s vague and hard to give a clear answer to. As it’s the last question of the survey, some may have missed it or skipped it due to participant fatigue. The methodology of the study and analysis presents limits: self-report data can have issues of reliability. Path analysis can only prove correlation, not causation, and the same goes for studies with a correlational design.

Task 2

Introduction

This analysis is based on a study focusing on the influence of perfectionism on exhaustion in a sample of 214 academics. The researchers carried out a correlational study design in which they collected responses to a questionnaire with 10 items, three of which related to perfectionism, three to rumination, and four to exhaustion. Research findings on exhaustion are important as they may help guide changes on an individual and policy level towards increasing wellbeing. In particular, this study may be relevant because rumination can be decreased by each individual and if its mediating role is found, this study would add a tool to the disposal of individuals to increase their wellbeing.

Research question

RQ. 1) Does the measurement model for perfectionism, rumination, and exhaustion fit the data?
RQ. 2) a. Does the hypothesised mediation model fit the data?
b. What are the size, direction, and statistical significance of the paths in the hypothesised mediation model?
c. Does rumination significantly mediate the relationship between perfectionism and exhaustion?

Hypotheses

H0. Perfectionism is not predictive of exhaustion, and the hypothesised mediation model does not fit the data.
H1. The relationship between perfectionism and exhaustion is mediated by rumination.
H2. Perfectionism positively predicts exhaustion.

Data analysis method and justification

Preliminary analysis

Data accuracy will be verified, then the location and pattern of missing data will then be ascertained. I will test for a missing data mechanism to determine whether the data is missing completely at random (MCAR) or follows another missing data mechanism (Rubin, 1976) to verify the assumption of equality of variance in structural equation modelling

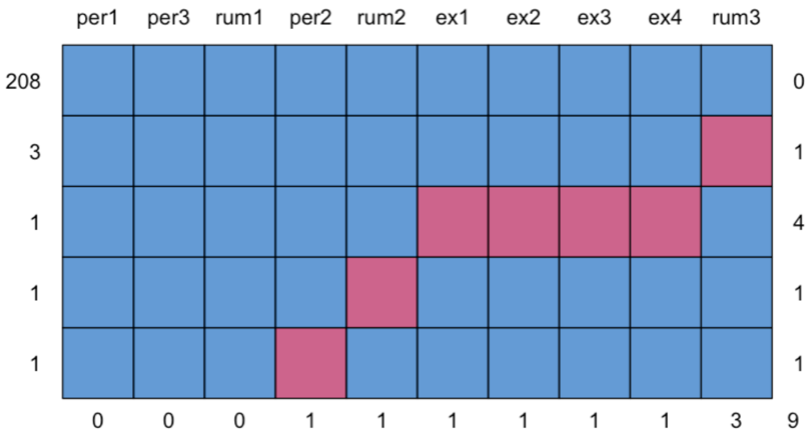
(SEM). Cases with more than 10% of missing values will be removed. The remaining missing values will be imputed using an adequate method. Outliers, both univariate and multivariate, will be dealt with using Z-scores and Mahalanobis distances as reference values. Skewness and kurtosis will be checked to address the normality assumption.

Primary analysis

Confirmatory factor analysis (CFA) will be used to get at the error-free common variance underlying the questionnaire items, comparing the model and actual covariance matrices to ascertain the model fit. Composite reliability will be calculated from factor loadings. A full SEM model, which tests causal relationships and measurements with latent variables, will be built and tested for fit.

Preliminary analyses

There were no anomalies in the range of the variables. I obtained the pattern, location and univariate proportion of missing data.



	per1	per3	rum1	per2	rum2	ex1	ex2	ex3	ex4	rum3	
208	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	0
3	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	1
1	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟦	4
1	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	1
1	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	1
	0	0	0	1	1	1	1	1	1	3	9

Figure 1. *Output of the md.pattern() function indicating the missing data patterns*

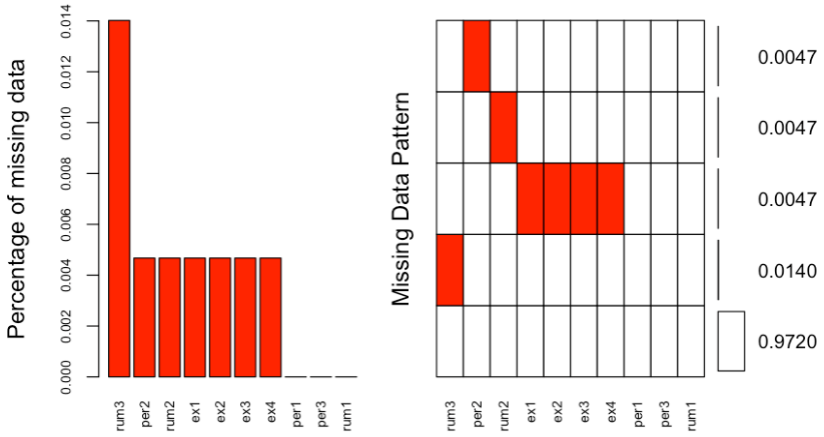


Figure 2. *Output of aggr() indicating the percentage and pattern of missing data for each variable of the dataset. The red cells in the patterns of missing*

data indicate variables containing missing values.

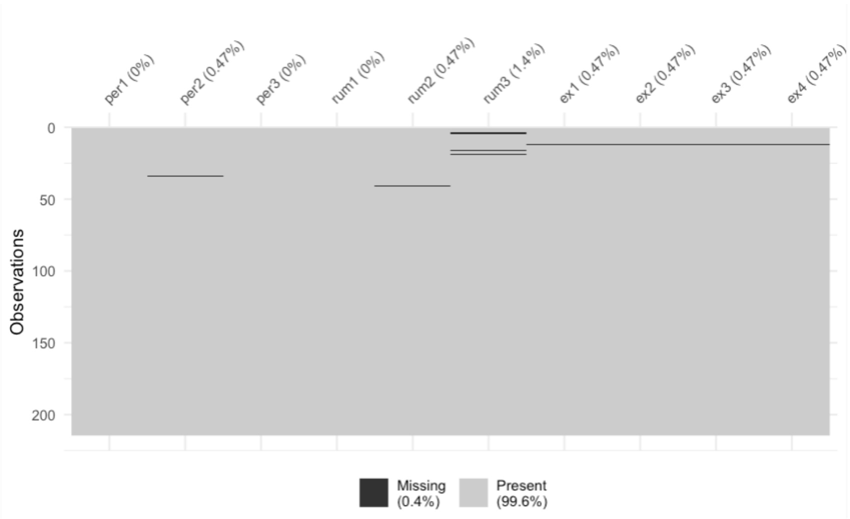


Figure 3. *Output of the vis_miss() function indicating the location of missing data for each variable of the dataset. The third figure shows that the dataset only contains 0.4% missing values, and is therefore 99.6% complete.*

The MCAR test's p-value is 0.59, larger than .05 so I accept the null hypothesis: the covariances are equal. This indicates that the data are missing completely at random (MCAR): there is no evidence of systematic patterning of the missing data across groupings. The data is missing completely at random, so I removed subjects listwise with large amounts of missing data (>10%), without introducing bias into the sample since the number of cases with missing data is small (< 5%).

It is reasonable to assume the missing values would closely resemble other values from the same scale, so mean imputation is appropriate (Cole, 2008). Accordingly, the scale means were calculated for each variable in the dataset, and mean scale scores were imputed to the missing values.

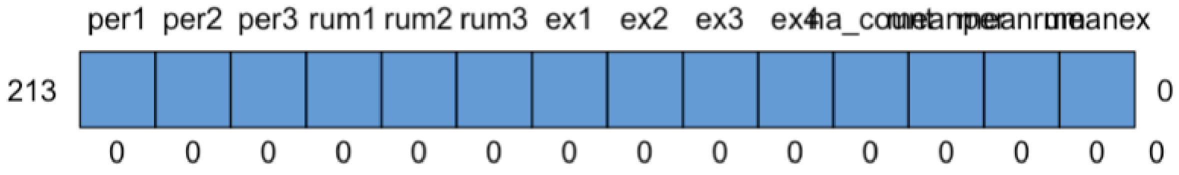
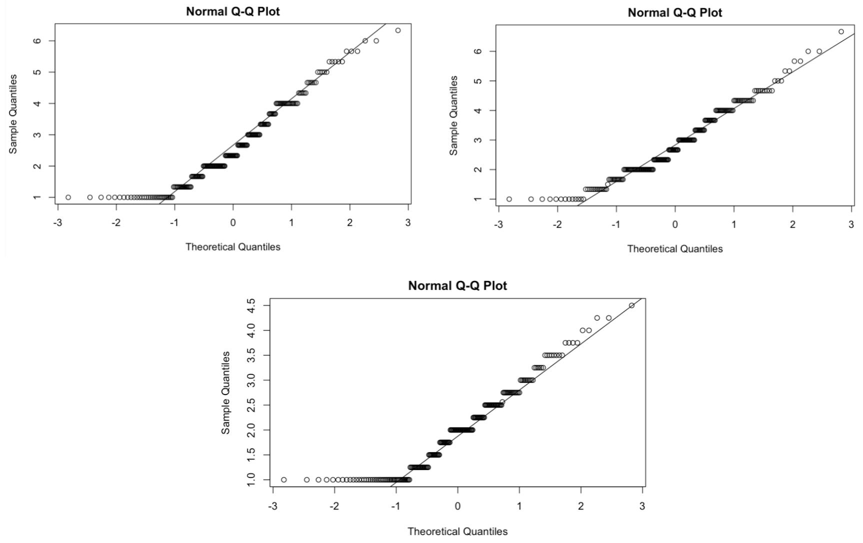


Figure 4. *Output of the md.pattern() function indicating the lack of missing data.*

Univariate outliers were detected by standardising the variables and removing cases with Z-scores larger than 3.29 (Tabachnick et al., 2013). Two cases were removed, leaving a sample of 211 participants. There were no multivariate outliers, as identified using Mahalanobis distances. Skewness: 0.6 (meanper), 0.48 (meanrum), and 0.57 (meanex) and kurtosis: -0.5 (meanper), -0.36 (meanrum), and -0.42 (meanex) are small values, attesting to the normality of our variables after removing outliers. QQplot verified approximate normality.



Figures 5-7. *QQplot of, in order, mean perfectionism, mean rumination and mean exhaustion.*

Primary analyses

The measurement model consisted of three latent variables: three questionnaire items were used as the measured variables for perfectionism, another three for rumination, and the final four for the exhaustion dimension.

All error-free correlations between latent factors were positive, statistically significant, and ranged in magnitude from moderate-to-large according to conventional effect size criteria (i.e., small ≥ .10, moderate ≥ .30, large ≥ .50; Cohen 1988).

The measurement model exhibited an acceptable fit to the data: χ^2 = 59.615 (32), $p < .05$; TLI = 0.96; CFI = 0.972; SRMSRl = 0.050; RMSEA = 0.064 (90% CI = 0.038 to 0.089). The questionnaire passed CFA, fitting the data adequately.

All standardised factor loadings for the measured variables on their latent factors were significant (perfectionism β range = 0.74 to 0.8; rumination β range = 0.8 to 0.9; exhaustion β range = 0.7 to 0.79). Furthermore, each of these latent factors demonstrated acceptable composite reliability (perfectionism ρ = 0.851525; rumination ρ = 0.8980367; exhaustion ρ = 0.9049322).

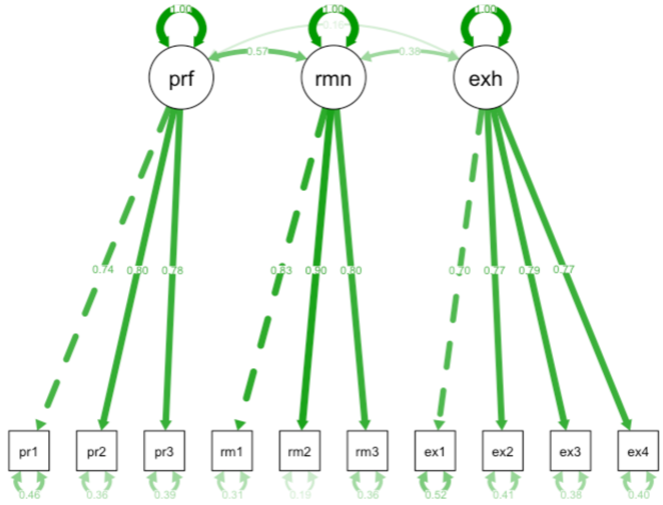


Figure 8. *Visualisation of the model with the standardised factor loadings and covariances.*

A full model was built with the predictive components as well as the measurement components, to test the hypothesised causal relationships. This is also done using the marker variable approach, to which is added the structural element.

Fit indexes from this mediation model suggested that it is fitted to the data: TLI= 0.962; CFI = 0.972; SRMSR = 0.052; RMSEA = 0.063 (90% CI runs from 0.037 to 0.087), suggesting that all paths are statistically significant. The model's implied correlation matrix approximates the data's correlation matrix well.

Inspecting the estimates and bootstrap confidence intervals for the regression paths and the indirect effects, which are at the basis of the hypothesised model, indicate that perfectionism positively predicted rumination (b = 0.54, B = 0.57, 95% CI = 0.35, 0.71). In turn, rumination positively predicted exhaustion ((b = 0.24, B = 0.375, 95% CI = 0.095, 0.37). The indirect effect of perfectionism on exhaustion via rumination with 5000 resamples had a standardised estimate of 0.21 and a 95% CI running from 0.05 to 0.21. his interval does not include a zero indirect effect: I can therefore reject the null hypothesis. This model accounted for 32% of the variance in rumination and 14% of the variance in exhaustion.

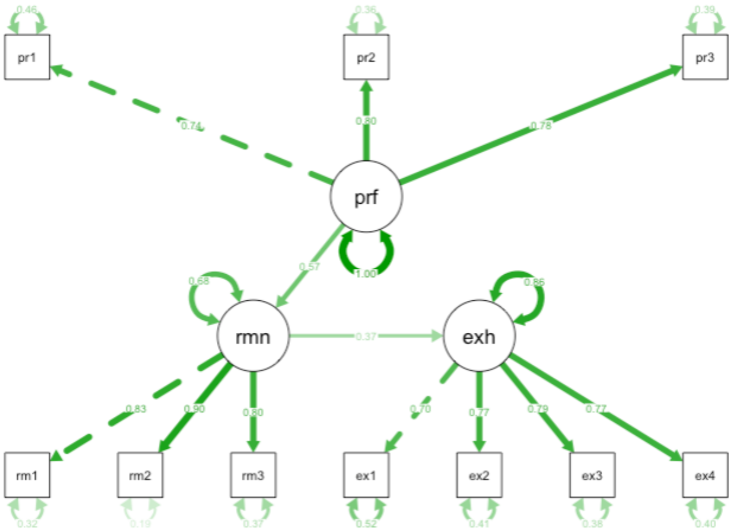


Figure 9. *Visualisation of the SEM path model with the measurement and structural elements, which indicates the size, direction, and statistical significance of the paths*

Discussion

These results allow me to reject H0: perfectionism is positively predictive of exhaustion and the hypothesised mediation model fits the data. Rumination does significantly mediate the relationship between perfectionism and exhaustion.

The methodology of the study is limited by the fact that self-report data can have issues of reliability. In addition, a correlational design can only prove correlation, not causation. Finally, the role of parental upbringing or income may have influenced the relationship between perfectionism and exhaustion, but these latent variables were not included in the dataset.

Task 3

Introduction

This analysis is based on a study focusing on the evolution of aggressive behaviour in 405 children over time. Aggressive behaviour was recorded once a year for four years, and parental neglect was measured once. The researchers used a correlational study design by collecting responses to a 10-item questionnaire, with three items relating to perfectionism, three to rumination, and four to exhaustion. This study has implications for caretakers and may provide important findings for increasing personality outcomes in children.

Research question

RQ. 1) Is child aggression changing over time among children in care?
RQ. 2) Do levels of parental neglect explain variance in the change trajectories of aggression among children in care?

Hypotheses

H0. Child aggression does not change over time among children in care.
Second H0. Levels of parental neglect don't explain variance in the change trajectories of aggression among children in care.
H1. Child aggression increases over time among the sample of children in care.
H2. Levels of parental neglect explain variance in the change trajectories of aggression among children in care, and experiencing more neglect explains steeper increases in aggression compared to those who experienced less.

Data analysis method and justification

Preliminary analysis

Dealing with missing data is necessary because the preliminary analysis uses the cor() function, which returns an error if there are missing values, and the dataset has not been cleaned or screened. Cases with more than 10% of missing values will therefore have to be removed. An empty model will be built to provide information about how much of the total variance in aggression is between-person variance and how much is between-time points variance. The intraclass correlation coefficient indicates the proportion of variation that is attributable to clustering.

Primary analysis

The general linear model assumes independence: the observations from a set of data don't depend on each other. That is almost impossible in psychological science because data can cluster on many levels. These levels have their own intercepts and slopes. In this case, the data is longitudinal so there will likely be within-person variance. The trajectories of aggression are likely nested within individuals. This can be accounted for using multi-level models: instead of using just one general random-effect that captures how each observation deviates from the predicted fixed-effects, there will be multiple random-effects capturing how intercepts and slopes deviate within a cluster, and how each cluster deviates from the overall group.

Preliminary analyses

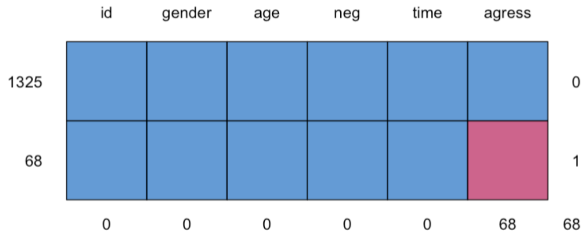


Figure 1. *Output of the md.pattern() function indicating the missing data pattern.*

4% of cases contain missing data, lower than the 5% threshold. I removed the subjects with >10% of missing data. This removed all of the entries containing missing data, leaving a sample of 1325. A faceted plot of the within-person variation across time visualised the relationship between neglect and aggression.

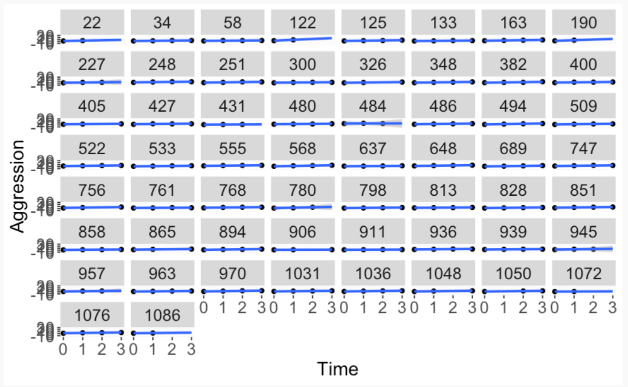


Figure 2. *Output of the ggplot() function. This output is hard to decipher so instead I'll use a different plotting of the relationship.*

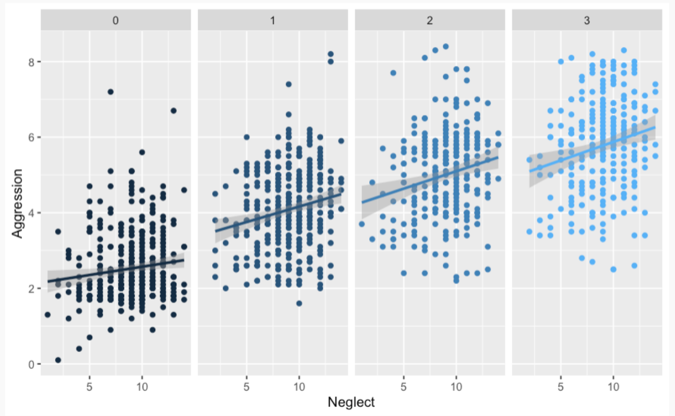
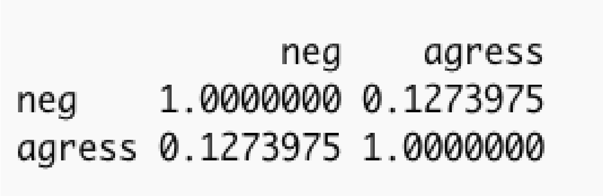


Figure 3. *Output of the ggplot() function showing a positive relationship between neglect and aggression in all of the four within-person time groups. However,*

the slopes vary a little and the intercept progressively increases. This indicates a clustering effect.



```
              neg    agress
neg    1.0000000 0.1273975
agress 0.1273975 1.0000000
```

Figure 4. *Output of the cor() function. The correlation coefficient between neglect and aggression (0.13) verifies the positive relationship between neglect and aggression. However, it does show the clustering effect indicated by the scatter plot. That effect needs to be addressed using multi-level modelling.*

The empty model was specified to only estimate 1 intercept for aggression, and group the data by participant. The variance component (0.3) indicates the amount of variability in aggression from person-to-person. The residual (2.4) represents the variability that is left unexplained in the empty model. This leftover variance is the variability in aggression across time, irrespective of person.

To partition the variance into between-person and between-time group sources, I divided the variance component of the intercept by the sum of the residual variance and the intercept variance. This resulted in an intraclass correlation coefficient (ICC) equal to 0.3 / (0.3 + 2.4) = 0.11. 11% of the variance in aggression can be explained by variation between individuals, which reinforces the necessity of using multi-level modelling.

Primary analyses

I added the Level 1 predictor of time to test whether it explains the within-person variance in aggression, tested the trajectories of aggression for each participant, and got the bootstrap 90% confidence interval for the fixed and random effects for statistical inference.

The expected value of aggression for the average person on the first time point is 2.7 (95% CI = 2.6, 2.8). Aggression increased throughout the study: for every unit increase in time, aggression increased by 1.12 (95% CI = 1.1, 1.17). There is between-person variation in the intercepts of aggression (variance = 0.57, sd = 0.76). The standardised variance of the intercepts has a 95% confidence interval running from 0.7 to 0.8. This does not include zero: although the intercept variance is reduced when time is added to the empty model, there remains a significant amount of random variance in the intercepts to be explained.

There is also between-person variation in the within-person trajectories of aggression over time (variance = 0.07, sd = 0.27). While the mean change in aggression is low, there is a lot of variance around that mean. The low slope per fixed effect hides significant person-to-person variability. This variability has a 95% confidence interval running from 0.2 to 0.3 This does not include zero. So despite the non-significance of the fixed effect of time, there remains a significant amount of random variance in the slopes to be explained.

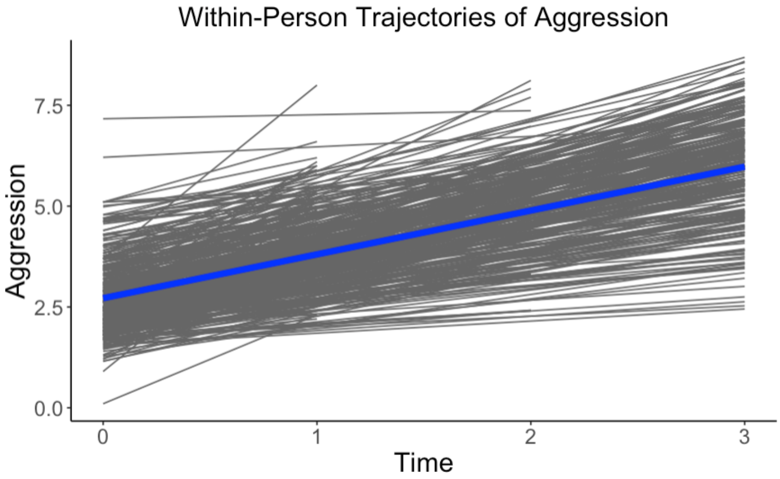


Figure 4. *Plot of the within-person aggression trajectories showing the person-to-person variability.*

The blue line shows the average trajectory. This line has a positive slope and there is significant variability in the intercepts, and a little in the slopes, around the blue line.

I added the between-person predictor, parental neglect, to the model to test whether the variability of the relationship between time and aggression is explained by parental neglect. I added to the model the main effect of aggression and its interaction with time. I grand-mean-centred neglect before fitting the model so that the unit of measurement for neglect was deviation from the sample mean.

The expected value of aggression for the average person at the first time point is 2.23. Aggression increased with time. For every unit increase in time, aggression increased by 0.93 (95% confidence interval = 0.77, 1.1). This is a significant change.

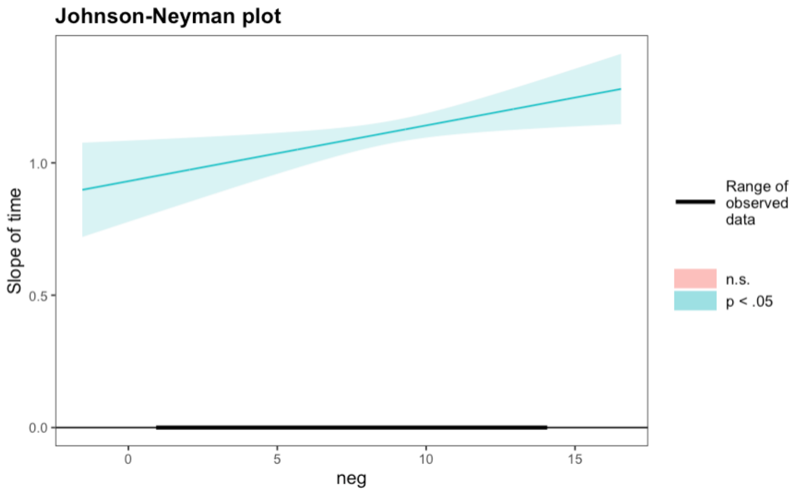
People with higher parental neglect also tended on average to have higher scores in aggression (b = 0.05). For every 1 unit increase in parental neglect away from the group mean, there is a significant (95% CI = 0.02, 0.1) corresponding 0.05 deviation in aggression from the group mean.

Between-person deviations for the grand mean in neglect have a significant (95% CI = 0.003, 0.04), small moderating effect on the relationship between time and aggression (b = 0.02). Variance in the trajectories of aggression is explained, in very small part, by between-person differences in neglect.

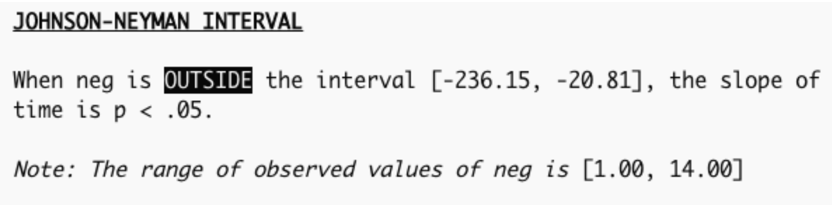
There is significant (95% CI = 0.68, 0.84) between-person variability in the intercepts of aggression (variance = 0.6, sd = 0.7). While the intercept variance is reduced by adding aggression to the time model, some random variance in the intercepts is explained by other factors.

There is significant (95% CI = 0.2, 0.31) between-person variability in the within-person trajectories of aggression over time (variance = 0.07, sd = 0.27). This means that while the mean change in aggression is negligible there is some variance around that mean change from person-to-person. While the interaction of time and neglect is significant, some random variance in the slopes is explained beyond neglect.

The significant (95% CI = 0.06, 0.55) 0.26 correlation between the random intercept slope indicates that higher starting points for aggression correlate with increases in aggression over time.



Figures 5-6. *Output and plot of the interaction term using the Johnson-Neyman technique, plotting the conditional slopes across the range of values of neglect. The within-person trajectories of aggression are significant at all levels of the observed values of neglect (i.e., more than -20.81 units from the grand mean).*



```
JOHNSON-NEYMAN INTERVAL

When neg is OUTSIDE the interval [-236.15, -20.81], the slope of
time is p < .05.

Note: The range of observed values of neg is [1.00, 14.00]
```

This indicates that the children in the sample all had significant increases in aggression over the study period, regardless of their score in parental neglect.

Discussion

For every unit increase in time, aggression increased by 0.93, refuting the null hypothesis and supporting H1. Higher neglect was linked to higher aggression; variation in the trajectories of aggression is explained in small part by levels of neglect (refuting the second null), but this effect may be negligible. This weakly supports H2. Other factors explain a significant amount of the variance.

Bibliography

Cohen, J. (1988). Set correlation and contingency tables. *Applied psychological measurement, 12*(4), 425-434.

Cole, J. C. (2008). How to deal with missing data. *Best practices in quantitative methods*, 214-238.

Jamshidian, M., Jalal, S., & Jansen, C. (2014). Missmech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). Journal of Statistical Software, 56(6), 1–31. https://doi.org/10.18637/jss.v056.i06

Rubin, Donald B. 1976. Inference and missing data. Biometrika 63(3): 581-592.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6, pp. 497-516). Boston, MA: pearson.