The Bifactor Model of Psychopathology:

Methodological Issues and Clinical Applications

A thesis submitted for the degree of Doctor of Philosophy by

Matthew Paul Constantinou

University College London

September 2019

Declaration

I, Matthew Paul Constantinou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

.... 04/09/19

Abstract

For decades, clinicians have debated whether psychiatric problems should be ‘lumped’ into broad dimensions or ‘split’ into discrete entities. The bifactor model provides a potential solution to this debate by including both a general dimension of psychopathological severity known as the *p* factor, and specific dimensions reflecting specific problem areas such as internalizing and externalizing. This thesis evaluates the methodological properties and clinical utility of the bifactor model.

Chapter 3 is a reliability review of bifactor studies and demonstrates that while self-report measures capture both general and specific domains of psychopathology, the total and subscale scores derived mainly reflect a general *p* factor. Chapter 4 investigates whether the general and specific psychopathology factors are products of response biases (i.e. tendencies in the way people fill out questionnaires), rather than variation in people’s experiences of psychiatric problems. Less than 4% of the variance in the general and specific psychopathology factors was explained by response biases, demonstrating their substantive validity.

Chapter 5 analyzes clinical outcomes assessed over a psychosocial intervention for antisocial youth with a bifactor model and demonstrates more nuanced changes in disorder-specific factors after accounting for changes in the *p* factor (e.g., antisociality declines but anxiety increases over time). Similarly, Chapter 6 demonstrates the prognostic value of specific personality disorders for predicting depression outcomes assessed over an inpatient intervention only after accounting for the prognostic effect of a general personality disorder factor (e.g., borderline personality disorder predicts slower recovery).

These findings demonstrate the substantive nature of the general and specific psychopathology factors, but also the difficulties in reliably measuring specific domains beyond general psychopathology. They also support the bifactor model’s utility in untangling clinically relevant effects that are otherwise masked by the shared variance among psychiatric problems.

Impact Statement

This thesis adds to a new wave of research showing that common psychiatric disorders, such as depression, substance abuse, and psychosis, can be measured on a spectrum known as the general psychopathology factor or the *p* factor. In other words, a single dimension describes the severity of people's problems across a range of psychiatric disorders that are characteristically thought to be distinct. The unique qualities of specific psychiatric disorders are still important to understand, but not without also considering their shared characteristics.

Impact for Psychiatric Research.

Psychiatric research is focused on identifying what makes psychiatric disorders different, e.g., identifying unique biomarkers and environmental risk factors. However, this thesis suggests that it is also important to consider what makes psychiatric disorders similar. It demonstrates that findings associated with specific disorders, such as their amenability to psychosocial interventions, might in fact reflect characteristics shared by all disorders, i.e. general psychopathology. A tool is presented, the bifactor factor analytic model, that allows researchers to control for these shared characteristics to isolate the unique effects of specific disorders. By publishing tutorial papers and hosting academic workshops, the bifactor model could become more widespread in the scientific community.

Impact for Clinical Practice

Much like psychiatric research, clinical practice is organized around specific disorders. Clinicians aim to identify the most fitting psychiatric diagnosis using disorder-specific interviews or questionnaires. A diagnosis then guides the choice of disorder-specific interventions that are designed to target a disorder's unique mechanisms. This thesis suggests that it is important for clinicians to assess both the shared and specific characteristics of psychiatric disorders.

The characteristics shared by all disorders (i.e. general psychopathology) might index the overall severity of a service user's impairment and could guide the intensity of an intervention (e.g., self-help, outpatient, or inpatient services). The unique characteristics of disorders could inform the type of intervention delivered (e.g., modality, format). This alternative approach to clinical assessment could be implemented by tasking clinicians, clinical scientists, and experts by experience, to develop reliable, valid, and clinically sensitive measures that can be used to assess these complementary aspects of a service user's presentation.

Impact for Mental Health in Society

A single dimensional measure of psychopathological severity aligns with the growing view in society that mental health, like physical health, varies on a spectrum: we all have it and it changes for better or worse depending on our circumstances. The findings in this thesis can be used in public engagement activities that aim to educate about mental health and reduce stigma, as well as public policy interventions that aim to prevent poor mental health (e.g., by targeting general psychopathology rather than specific disorders).

Table of Contents

List of Tables

List of Figures

Note to Examiners

The following are published and prospective manuscripts that use content from this thesis:

Constantinou M. P., & Fonagy P. (pre-print). Evaluating Bifactor Models of Psychopathology with Model-Based Reliability Statistics. *PsyArXiv.* doi:10.31234/osf.io/6tf7j

Constantinou M. P., Allison, E., & Fonagy P. (under review). Fact or Artifact? Testing the Response Bias Hypothesis of the General Psychopathology *p* factor through Crowdsourcing. *Psychological Methods.*

Constantinou, M. P., Goodyer, I. M., Eisler, I., Butler, S., Kraam, A., Scott, S., ... & Fonagy, P. (2019). Changes in General and Specific Psychopathology Factors Over a Psychosocial Intervention. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(8), 776-786. doi:10.1016/j.jaac.2018.11.011

Constantinou, M. P., & Fonagy, P. (under review). Response to "Common Factors and Interpretation of the P Factor of Psychopathology". *Journal of the American Academy of Child & Adolescent Psychiatry*.

Constantinou, M. P., Frueh, B. C., Fowler, J. C., Allen, J. G., Madan, A., Oldham, J. M., & Fonagy, P. (in prep). Predicting Responses to Inpatient Treatment for Depression Using the General and Specific Personality Disorder Factors.

Acknowledgements

Behind every great work is a person, and behind every person is usually a group of inspiring and loving people. I am fortunate to have received an unparalleled level of intellectual and emotional support throughout my studies, so I take this opportunity to thank those who kept me feeling inspired and loved.

First and foremost, I thank my supervisors, Peter Fonagy and Liz Allison, for providing the perfect balance of intellectual and emotional support (with many a time Liz providing the intellectual support and Peter the emotional). Peter, it goes without saying that you have shaped the psychological scientist I consider myself to have become. But it is your implicit teachings that, upon reflection, have been most formative. First, you have taught me the importance of silence and patience. To quote Plato, "Wise men speak because they have something to say; fools because they have to say something". Ideas need time to develop and should not be rushed. This is not to say that our meetings were spent in silence. On the contrary, we spent most of the time laughing. Hence, you have taught me to be passionate but never to take myself too seriously. In turn, one can remain humble. You have always said, "it doesn't pay to be nasty, and it doesn't cost anything to be nice". By seeing the best in others, one can listen with an open ear. It is no coincidence that the theories of mentalizing and epistemic trust were pioneered by someone who listens without judgement and recognises with integrity. These are just some of the many teachings that have helped me become a better scientist and all-round human being, and while I would not say that I have mastered them all, I hope to one day pass them on to the next generation of scientists, implicitly, of course.

Liz, if Peter was the sail, then you were the anchor. You have been the person I could turn to with my deepest concern and supported me unconditionally through my hardest times. I am always struck and inspired by how someone so knowledgeable about the mental condition can put it all to one side and be fully present in what someone else is thinking and feeling. People, including myself, can be so quick to judge or attribute meaning, and, as a consequence, lose sight of the lived experience. You, on the other hand, have never redirected one of my issues or 'filed away' a troubling experience. It is truly an understatement to say that you have been my sanctuary of mind over the past three years and I am indebted to you for that. You have also shaped my aspirations in life and how I want to live it. You helped me see that while our work is incredibly stimulating, one should not to lose sight of the things that are irreplaceable, such as our loved ones (and training!). Thank you for giving me the confidence to lead a healthier, but no less fulfilling, life.

I thank my colleagues who have made this journey both fascinating and fun. Chloé, for our stimulating discussions, Laura and Tobi, for our countless exchanges about *p*, scholarship, and life in general, Jilly, for taking the stress out of meetings (and sharing a passion for fitness), Sophie Bennet and Helen King, for making student life as smooth as possible, and everyone at the unit who endured my random stretching routines, strange sitting positions, and constant eating habits (Alex, Angela, Christine, Clare, Daniela, Lila, Michal, Nicola, Tamy, and Thomas). I also thank Jon Roiser, for first having confidence in my ability and for your patience and support throughout the programme, my rotation supervisors, Himanshu Tyagi and Eileen Joyce, Eamon McCrory and Essi Viding, and Roz Shafran and Amy Coughtrey, for enriching me with personal and professional experiences that have shaped me to this day, my colleagues within UCL (Martin Debbané, Pasco Fearon, Patrick Luyten, Steve Pilling, and David Tuckett) and outside of UCL (Jon Allen, Chris Fowler, Chris Frueh, Alok Madan, and Jon Oldham; Stephen Butler, Ivan Eisler, Ian Goodyer, Abdullah Kraam, and Stephen Scott) for sharing their time, insight, and expertise.

I thank my friends and family for their continued love and support. I am particularly grateful to the following people:

Niki and Mario, for being the two people I can always count on,
Yiayia, Σ’ευχαριστὠ πολὐ,
Alex and Sive, for helping me stay curious,
Chris and Tilly, for helping me live,
Mattia and Fran, for being my inspiration,
Jack and Adel, for always being there,
Jonny and Petros, for all the laughs (osu!).

Finally, I thank my parents–to whom I dedicate this thesis–for always encouraging me to pursue my interests and for putting my needs before their own without hesitation.

Chapter 1 Introduction to Factor Analysis and Hierarchical Models

In this chapter, I aim to provide the reader with a basic understanding of factor analysis–the main statistical method used in this thesis. I also aim to present a broad historical and statistical overview of the main factor analytic models used in this thesis and beyond, including the single factor, common factor, bifactor, and higher-order factor models. I then introduce the bifactor model in the context of psychopathology research–which preoccupies the content of this thesis–and conclude by outlining the thesis aims.

1.1 What is Factor Analysis?

Factor analysis is a statistical method used to describe the relationships among a set of observed variables from a smaller set of latent variables known as factors (Olkin & Sampson, 2001). The method was first introduced by Charles Spearman (1904), who demonstrated that the correlations among intelligence test scores could be explained by a single general intelligence factor (the '*g*' factor), as well factors specific to each type of test. In practical terms, one could design a questionnaire to assess depression. While the items might capture various aspects of depression, including low mood, reduced motivation and poor concentration, they all contribute to measuring the latent construct of 'depression', which can be represented as a factor.

Factor analysis assumes that the observed variables have at least one common influence (e.g., a factor representing a psychological construct) which can be deduced from the relationships among the variables (Wright, in press). After reverse-engineering the variance associated with a latent factor from the covariance among a set of observed variables, the factor is used to predict variation in each observed item. There is an implicit assumption that the factor represents a psychological construct that exists beyond the data used to estimate it, and which causes variation in the observed variables (Harman, 1960). However, factor analysis is built on regression analysis which relies on statistical dependencies rather than causal inferences. Like in regression, other psychological constructs ('third variables') might contribute to the variation in a factor (e.g., motivation or test-taking ability; Coan, 1964). Factors might even reflect non-substantive aspects of the measure (e.g., item-wording effects; Podsakoff, MacKenzie, & Podsakoff, 2012).

Statistically speaking, the aim of factor analysis is to recreate a matrix of correlation coefficients among all combinations of observed variables–the **R** matrix (Cattell, 1965). The observed variables can be performances on intelligence tests, scores on psychological questionnaires, or even biological measures such as cell culture densities–any variables that are hypothesised to covary due to broader common influences. One recreates the **R** matrix using an $j \times k$ factor matrix, or **V** matrix, where k is the number of factors ($k = 1, \dots, K$), which is typically smaller than the number of observed variables, j, where $j = 1, \dots, J$. The **V** matrix includes factor loadings (λ) which reflect the strength and direction of the prediction between the factors and each observed variable (i.e. how much an observed variable is predicted to increase or decrease with a one-unit increase in a given factor). Factor loadings are calculated by taking the mean correlation between a given variable with all other variables (i.e. the column sums in the **R** matrix). The **R** matrix is then resolved by multiplying the **V** matrix with its transpose ($\mathbf{V}'$), which reproduces the correlations between variables using the factor estimates. In more concrete terms, each cell in the **R** matrix is restored by multiplying each row of the **V** matrix with each column of the **V′** matrix (see Figure 1.1).

$$\mathbf{R}\;\begin{array}{c|cccccc|}
 & y_1 & y_2 & y_3 & y_4 & \cdots & y_J \\ \hline
y_1 & v_1 & & & & & \\
y_2 & r_{12} & v_2 & & & & \\
y_3 & r_{13} & r_{23} & v_3 & & & \\
y_4 & r_{14} & r_{24} & r_{34} & v_4 & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\
y_J & r_{1J} & r_{2J} & r_{3J} & r_{4J} & r_{5J} & v_J
\end{array} = \mathbf{V}\;\begin{array}{c|cccc|}
 & f_1 & f_2 & \cdots & f_K \\ \hline
y_1 & \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1K} \\
y_2 & \lambda_{21} & \lambda_{22} & & \\
y_3 & \lambda_{31} & \lambda_{32} & & \\
y_4 & \lambda_{41} & \lambda_{42} & & \\
\vdots & \vdots & \vdots & \ddots & \\
y_J & \lambda_{J1} & \lambda_{J2} & \lambda_{J3} & \lambda_{JK}
\end{array} \times \mathbf{V}'\;\begin{array}{c|cccccc|}
 & y_1 & y_2 & y_3 & y_4 & \cdots & y_J \\ \hline
f_1 & \lambda_{11} & \lambda_{12} & \lambda_{13} & \lambda_{14} & \cdots & \lambda_{1J} \\
f_2 & \lambda_{21} & \lambda_{22} & \lambda_{23} & \lambda_{24} & \cdots & \lambda_{2J} \\
\vdots & \vdots & & & & \ddots & \\
f_K & \lambda_{K1} & & & & & \lambda_{KJ}
\end{array}$$

Figure 1.1. Visual representation of the relationship between the **R, V**, and **V'** matrices. y_j and v_j refer to a given observed variable and its variance, respectively, $r_{j_1 j_2}$is the correlation coefficient among two observed variables, f_k is the variance for a given factor, λ_{jk} and λ_{kj} are the factor loadings for variable j on factor k. The circled regions in each matrix illustrate how multiplying a row of the **V** matrix with a column of the **V′** matrix reproduces the coefficients in the **R** matrix (e.g., $\lambda_{11} \times \lambda_{11} = v_1, \lambda_{12} \times \lambda_{21} = v_2$, etc).

Suppose a researcher believes that different psychiatric symptoms measure a single thing in common. She may develop a questionnaire covering a range of psychiatric symptoms, from low mood to grandiose delusions, and collect responses from a large outpatient population. She then decides to run a confirmatory model to test whether a single factor adequately explains the relationships among symptom responses, and how well variation in each symptom is predicted by a ʻgeneral psychopathology' factor. For patient i, where $i = 1 \ldots, N$, responses on a given symptom indicator can be expressed as:

$$y_{ij} = \mu_j + \lambda_{jk}\eta_{ik} + \varepsilon_{ij},$$

where μ_j is the intercept for indicator j (i.e. the predicted score when the factor equals zero), λ_{jk} is the factor loading of item j on factor k, η_{ik} are factor scores for a given participant i on factor k, and ε_{ij} reflects person-specific errors for each indicator (i.e. variance in symptom responses unaccounted for by the factor. Errors are assumed to have a mean of zero and should be uncorrelated with the factor). In matrix form, the model can be expressed as:

$$\mathbf{y}_j = \boldsymbol{\mu}_j + \boldsymbol{\Lambda}_{jk}\boldsymbol{\eta}_k + \boldsymbol{\varepsilon}_j,$$

where $\mathbf{y}_j$ is a vector of symptom responses for J indicators across the sample, $\boldsymbol{\mu}_j$ is a vector of J indicator intercepts, $\boldsymbol{\Lambda}_{jk}$ is a $j \times k$ matrix of factor loadings, $\boldsymbol{\eta}_k$ is a vector of factor values, and $\boldsymbol{\varepsilon}_j$ is a vector of indicator-specific errors. We can expand the matrix formula as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta},$$

where $\boldsymbol{\Sigma}$ is a $j \times j$ square correlation matrix of J indicators, $\boldsymbol{\Lambda}$ **is a** $j \times k$ factor loading matrix, $\boldsymbol{\Psi}$ is a $k \times k$ matrix of factor correlations (in this case a 1×1 matrix), and $\boldsymbol{\Theta}$ is a $j \times j$ diagonal matrix of errors for each indicator. This expanded equation is essentially the solution that we used to recreate the **R** matrix described above (e.g., **R** = **VV'**), but for a population correlation matrix rather than a sample correlation matrix.

The single-factor model is misleadingly known as Spearman's (1904) two-factor model because the variance in each indicator is partitioned into that which is accounted for by a common factor of interest, as well as that which is accounted for by a unique factor (see Figure 1.2). The unique factor is represented by a latent error term that is a mix of systematic error variance (e.g., reliable variance in the indicator unaccounted for by the factor) and random error variance (e.g., measurement error). In practice, these two sources of error are not separable, leaving us with a single substantive source of variance plus error (hence a 'single' factor).

Thurstone (1947) adapted Spearman's (1904) two-factor model to include more than one factor of interest. Each factor accounts for the correlations among different subsets of indicators. In other words, there is no general factor in Thurstone's common factor model (despite what the name suggests), just a set of common factors that account for commonalities among subsets of indicators (see Figure 1.2). The common factor model, more recently referred to as the correlated factors model, can be expressed as:

$$y_{ij} = \mu_j + \lambda_{j1}\eta_{i1} + \lambda_{j2}\eta_{i2} + \lambda_{j3}\eta_{i3} + \ldots + \lambda_{JK}\eta_{JK} + \varepsilon_{ij}.$$

The variance in each indicator is predicted to be a linear function of multiple common factors, as well as a unique error factor. Provided that the model follows a simple structure (i.e. each indicator loads on one and only one factor), a common factor model with two factors and four indicators can be summarized as:

$$y_{i1} = \mu_1 + \lambda_{11}\eta_{i1} + 0\eta_{i2} + \varepsilon_{i1},$$

$$y_{i2} = \mu_2 + \lambda_{21}\eta_{i1} + 0\eta_{i2} + \varepsilon_{i2},$$

$$y_{i3} = \mu_3 + 0\eta_{i1} + \lambda_{32}\eta_{i2} + \varepsilon_{i3},$$

$$y_{i4} = \mu_4 + 0\eta_{i1} + \lambda_{42}\eta_{i2} + \varepsilon_{i4}.$$

Factor loadings have been replaced by zeros to show that an indicator does not load on a given common factor. The common factors can be correlated or uncorrelated with each other; if strong enough, common factor correlations can be further analysed to estimate a higher-order factor (see section 1.3).

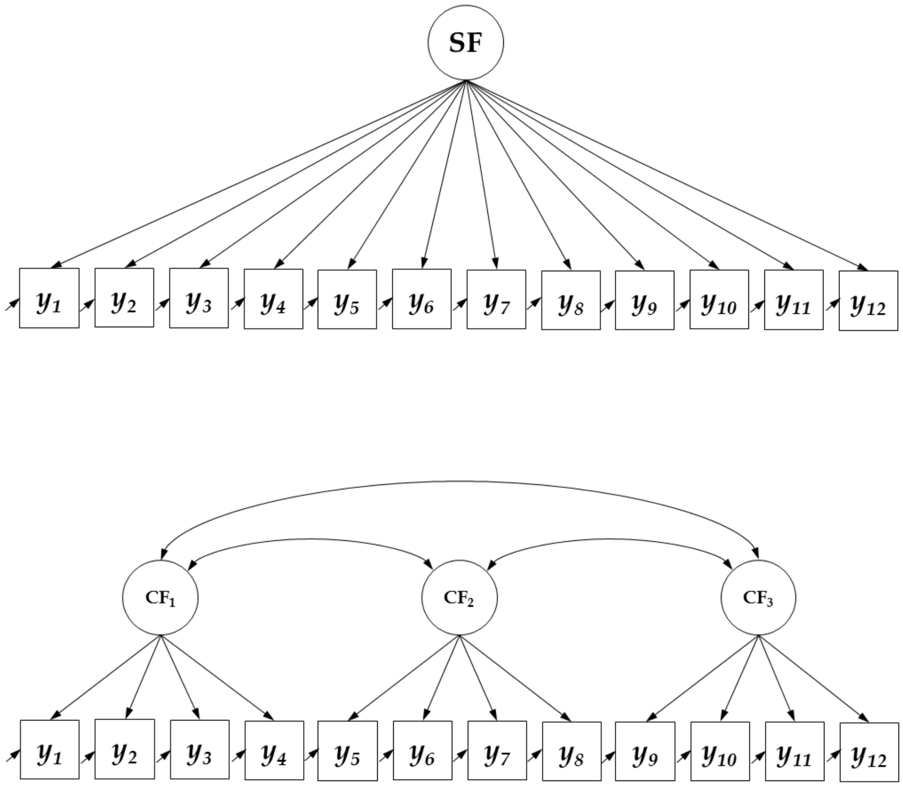

Figure 1.2. Schematic of the single-factor ('SF') model or Spearman's two-factor model (top) and the correlated factors model or Thurstone's common factor ('CF') model (bottom). Latent variables are represented by circles; observed variables are represented by squares. The diagonal arrows intersecting each observed variable represent the residual terms, which include systematic and random components. The common factors are correlated (represented by the bidirectional arrows between factors), but common factors can also be orthogonal (i.e. uncorrelated).

Factor models can either be estimated with exploratory or confirmatory factor analysis. Using Exploratory Factor Analysis (EFA), one makes no prior assumptions about the optimal number of factors that reproduce the variance-covariance matrix, or the way in which indicators relate to these factors. In other words, all possible factor loadings are estimated; no restrictions are specified (each item loads onto each factor). By contrast, in Confirmatory Factor Analysis (CFA), one has an idea of the number and nature of the factors, and restricts certain factor loadings and model parameters in line with the hypothesised model (e.g., the common factor model example above where certain loadings were set to zero is an example of setting restrictions). Both EFA and CFA aim to reproduce the R matrix using a smaller set of hypothetical variables but differ in whether they have been prespecified or not.

As I mentioned earlier, error variances in factor analysis have a systematic and random component. CFA allows one to specify an error structure among residuals to incorporate systematic influences unaccounted for the factors of interest. For example, the residuals of indicators that are similarly worded or come from the same scale can be correlated or predicted with a method factor. By contrast, EFA assumes that the indicator errors are random, i.e. once the common influences have been estimated, the remaining variance is assumed to be due to noise–there are no further systematic influences. This highlights a main disadvantage of EFA: there is little control over the factors estimated. It is therefore typical for optimal EFA solutions to include method factors (Brown, 2014).

As I described above, one can control for method effects in CFA by specifying correlations among residuals or estimating a method factor. While the high level of control offered by CFA makes it a preferred choice over EFA, it is also a double-edged sword. One can unknowingly restrict important sources of influence, leading to model misspecification and biased estimates. Ultimately, both CFA and EFA complement each other’s weaknesses and should be used in a guided fashion. For example, in scale development, EFA may be used to determine an optimal structure that is tested by CFA, but when testing theory, CFA may be used to test a hypothesised model followed by EFA to explore sample-specific deviations in fit (Brown, 2014).

1.2 Limitations of Factor Analysis

The statistical groundwork of factor analysis has been developed extensively over the last century, but its theoretical assumptions remain debatable. In a paper that has now become a (young) classic, Borsboom, Mellenbergh, and van Heerden (2003) summarized the indefensible assumptions of factor analysis. As described above, factor analysis involves the regression of observed variables (e.g., item response data) on a latent variable that represents a construct of interest (e.g., neuroticism). In turn, the latent factor is thought to cause the variation in item responses. Not making this assumption would question the use of factor analysis; other methods, such as principal components analysis or weighted sum scores, could be used to summarize the covariation in item responses without assuming causality on the part of the summary variable.

The main issue associated with the causal assumptions behind latent factors is that they are based on a tautology, which, by definition, is circular not causal. In other words, factors predict (and potentially cause) the variation in item responses, but they themselves are estimated from the covariances among item responses. Therefore, the variable that supposedly causes the covariation is in a sense a product of it. To overcome this issue, one needs to assume that the factor represents something real in the world that is distinct from the dataset it predicts. For example, there must be a real neuroticism trait in the population that causes individuals to respond differently on a questionnaire about emotional experiences, and this trait can be represented statistically. Put differently, we can justify the tautology with a 'gestalt' of sorts, by saying that the variable used to predict variation in item responses is in fact a placeholder for something 'out there' that is more than a summation of the item responses fed into it.

Borsboom et al. (2003) argued that the realist view required of factor analysis is hard to justify, since nowhere in the statistical formulation is this 'gestalt' specified. We are forced to assume that our model parameters are underpinned by a 'real' psychological trait, but our model simply indicates what the parameters would be *if* they were generated by that model. Other data-generating mechanisms could have caused the pattern of item responses (van der Maas et al., 2006). In fact, factor analytic models are subject to 'under-determination', i.e. the pattern of observed responses can be summarized equally well by an infinite number of models (Molenaar & von Eye, 1994). Hence, there is a gap between one's factor model and the forces underlying the pattern of observed responses that is often overlooked due to the realist assumptions of factor analysis.

The current thesis takes a 'pseudo-realist' view as described by Wright (in press). Like the realist view, it is assumed that factors reflect something real that is distinct from, but implied by, the data. However, that 'something' is not fully known and is based on a best guess, which is currently a mix of processes that include the latent construct of interest (e.g., neuroticism), but also overlapping constructs (e.g., alexithymia) and method effects (e.g., response biases). The goal of latent variable modelling should not be to just estimate a construct of interest, but also to isolate it from the other processes that are inherently captured by a factor. As the reader shall see, there is a constant effort in each chapter to validate the factors against explanatory variables (e.g., response biases; see Chapter 4) or clinical outcomes (see Chapters 5 and 6). The take-home message is that these factors should be thought of as proxies, rather than direct measures, of psychopathology constructs that themselves are multi-faceted and not fully understood.

1.3 Hierarchical Models

We have already seen the two main factor models: Spearman's two-factor model (with a single general factor and unique error terms) and Thurstone's common factor model (with multiple common factors and unique error terms). A problem with the two-factor model, and perhaps the greatest criticism Spearman received other than *g* being an artifact of his method, was that by focusing on the general factor of intelligence, more specific factors were 'explained away' (Beaujean, 2015). The common factors model thus took prominence in the 1930s, but ironically, began to 'explain away' the general factor (Holzinger, 1945). Eventually, those who saw the importance of both general and specific factors developed 'hierarchical' models (Gustafsson & Åberg-Bengttson, 2010).

The bifactor model (also known as the nested factor model) is a hierarchical model introduced by Karl Holzinger within an exploratory framework (Holzinger & Swineford, 1937) and extended to a confirmatory framework by Jan-Eric Gustafsson (Gustafsson & Balke, 1993). It includes a general factor with loadings from all indicators, as well as domain-specific factors with loadings from subsets of indicators (see Figure 1.3). The general factor is orthogonal to the specific factors, so that change in the indicators predicted by the general factor is independent from change predicted by the specific factors. Traditionally, specific factors are specified to be orthogonal to each other (Holzinger & Swineford, 1937).

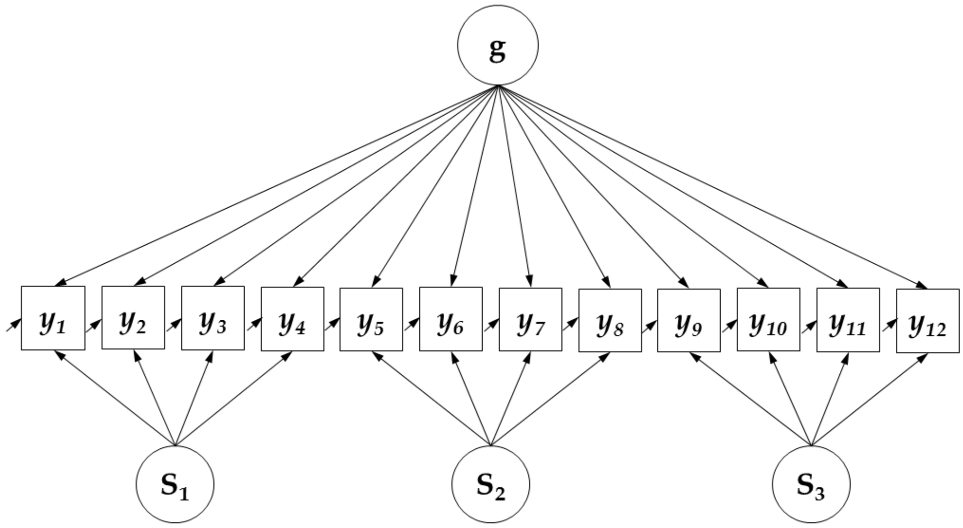

Figure 1.3. Schematic of the bifactor model. 'g' reflects the general factor, which predicts each observed variable directly. 'S' reflects the specific factors that predict clusters of observed variables. Error terms reflect the systematic and random components that are unaccounted for by the general and specific factors.

The basic formula for reproducing the R matrix using a bifactor model is:

$$\mathbf{R} = \mathbf{GG}' + \mathbf{SS}' + \mathbf{u^2},$$

where **G** is a factor loading matrix for the general factor, **S** is a factor loading matrix for the specific factors, and $\mathbf{u^2}$ is a vector of indicator-specific error variances. The variance of a test is thus partitioned into three sources: i) that which is common to all indicators, ii) that which is specific to certain groupings of indicators, and ii) that which is unique to certain items due to error. Within this division of variance, there is a bias towards the general factor because specific factors are estimated from the covariance remaining after accounting for the general variance (i.e. they are residual factors). This follows the British tradition which emphasised the general factor over specific factors (Beaujean, 2015). Note that while Karl Holzinger was American, his work was heavily inspired by Spearman as a mentor and colleague (Holzinger, 1945).

Responses on a given item j for a given participant i in the bifactor model can be summarized as:

$$y_{ij} = \mu_j + \lambda_{jG}\eta_{iG} + \lambda_{jS_1}\eta_{iS_1} + \lambda_{jS_2}\eta_{iS_2} + \ldots + \lambda_{jS_m}\eta_{iS_m} + \varepsilon_{ij},$$

for a general factor (G) and m specific factors $(S = 1, \ldots, m)$ that are orthogonal to each other. This can also be expressed in matrix form as:

$$\mathbf{y}_j = \boldsymbol{\mu}_j + \boldsymbol{\Lambda}_{jk}\boldsymbol{\eta}_k + \boldsymbol{\varepsilon}_j$$

where $\mathbf{y}_j$ is a vector of observed responses on each indicator, $\boldsymbol{\mu}_j$ is a vector of intercepts per indicator, $\boldsymbol{\Lambda}_{jk}$ is a $j \times k$ matrix of factor loadings for the general (G) and specific (S) factors, $\boldsymbol{\eta}_k$ is a vector of factor values for k general and specific factors, and $\boldsymbol{\varepsilon}_j$ is a vector of indicator-specific errors (see Figure 1.4).

$$\mathbf{y}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \boldsymbol{\mu}\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{bmatrix} + \boldsymbol{\Lambda}\begin{bmatrix} \lambda_{1,G} & \lambda_{1,S_1} & 0 & 0 \\ \lambda_{2,G} & \lambda_{2,S_1} & 0 & 0 \\ \lambda_{3,G} & \lambda_{3,S_1} & 0 & 0 \\ \lambda_{4,G} & \lambda_{4,S_1} & 0 & 0 \\ \lambda_{5,G} & 0 & \lambda_{5,S_2} & 0 \\ \lambda_{6,G} & 0 & \lambda_{6,S_2} & 0 \\ \lambda_{7,G} & 0 & \lambda_{7,S_2} & 0 \\ \lambda_{8,G} & 0 & \lambda_{8,S_2} & 0 \\ \lambda_{9,G} & 0 & 0 & \lambda_{9,S} \\ \lambda_{10,G} & 0 & 0 & \lambda_{10,S_3} \\ \lambda_{11,G} & 0 & 0 & \lambda_{11,S_3} \\ \lambda_{12,G} & 0 & 0 & \lambda_{12,S_3} \end{bmatrix} \times \boldsymbol{\eta}\begin{bmatrix} \eta_G \\ \eta_{S_1} \\ \eta_{S_2} \\ \eta_{S_3} \end{bmatrix} + \boldsymbol{\varepsilon}\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}$$

Figure 1.4. Matrix representation of the bifactor model. y_j reflects the observed scores for a given indicator, μ_j reflects the intercept for a given indicator, $\lambda_{j,G}$ reflects the factor loading for a given indicator onto the general factor, λ_{j,S_m} reflects the factor loading for an indicator onto a given specific factor (zeros denote where factor loadings have been constrained), η_k reflects factor scores for the general and specific factors, and ε_j reflects indicator-specific errors.

The other main hierarchical model is the higher-order model (also known as a second-order model) introduced by Thurstone (1944) in an exploratory context, and Jöreskog (1971) in a confirmatory one. The higher-order model is equivalent to the common factors model, but the correlations among the common factors, also called first-order factors, are described by a higher-order or second-order factor (see Figure 1.5). In other words, the higher-order factor predicts the correlations among common factors, in the same way that the common factors predict the correlations among indicators. Rather than decomposing the variance into different sources, the higher-order model identifies a common stream of variance organized in a hierarchy: the higher-order factor describes what is common among first-order factors, which in turn describe what is common among indicators. Nonetheless, common factors are estimated first and are thus prioritised over the general higher-order factor, which, following the American tradition, is thought to be a product of the common factors (Beaujean, 2015).

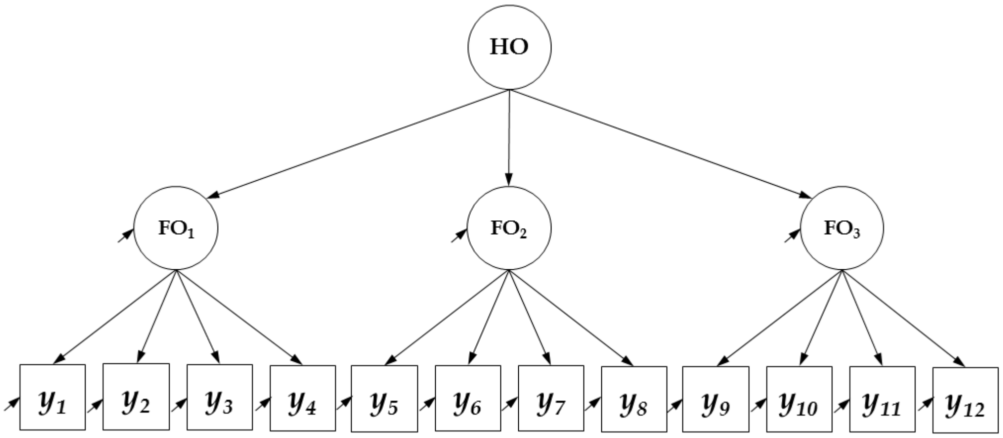

Figure 1.5. Schematic of the higher-order model. ‘FO’ reflects the first-order or common factors, which directly predict each observed variable. ‘HO’ reflects the higher-order factor, which directly predicts the first-order factors and indirectly predicts the observed variables. Error terms for the observed variables reflect the systematic and random components that are unexplained by the first-order factors, and errors terms for the first-order factors reflect the systematic and random components that are unexplained by the higher-order factor.

The higher-order model can be expressed as follows:

$$\mathbf{y}_j = \boldsymbol{\mu}_j + \boldsymbol{\Lambda}_{jk}\boldsymbol{\eta}_k + \boldsymbol{\varepsilon}_j$$

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta},$$

where $\boldsymbol{\Lambda}_{jk}$ is a $j \times k$ matrix of factor loadings of the observed indicators on the common/first-order factors, $\boldsymbol{\eta}_k$ is a vector of factor values for the common/first-order factors, $\boldsymbol{\varepsilon}_j$ is a vector of indicator-specific errors, $\boldsymbol{\Gamma}$ is a matrix of first-order factor loadings on the higher-order factor, $\boldsymbol{\xi}$ is a vector of factor values for the higher-order factor, and $\boldsymbol{\zeta}$ is a vector of residuals for the first-order factors (e.g., variance unaccounted for by the higher-order factor; see Figure 1.6).

$$\mathbf{y}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \boldsymbol{\mu}\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{bmatrix} + \boldsymbol{\Lambda}\begin{bmatrix} \lambda_{1,FO_1} & 0 & 0 \\ \lambda_{2,FO_1} & 0 & 0 \\ \lambda_{3,FO_1} & 0 & 0 \\ \lambda_{4,FO_1} & 0 & 0 \\ 0 & \lambda_{5,FO_2} & 0 \\ 0 & \lambda_{6,FO_2} & 0 \\ 0 & \lambda_{7,FO_2} & 0 \\ 0 & \lambda_{8,FO_2} & 0 \\ 0 & 0 & \lambda_{9,FO_3} \\ 0 & 0 & \lambda_{10,FO_3} \\ 0 & 0 & \lambda_{11,FO_3} \\ 0 & 0 & \lambda_{12,FO_3} \end{bmatrix} \times \boldsymbol{\eta}\begin{bmatrix} \eta_{FO_1} \\ \eta_{FO_2} \\ \eta_{FO_3} \end{bmatrix} + \boldsymbol{\varepsilon}\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}$$

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\begin{bmatrix} \Gamma_{FO_1,SO} \\ \Gamma_{FO_2,SO} \\ \Gamma_{FO_3,SO} \end{bmatrix} \times \boldsymbol{\xi}[\xi] + \boldsymbol{\zeta}\begin{bmatrix} \zeta_{FO_1} \\ \zeta_{FO_2} \\ \zeta_{FO_3} \end{bmatrix}$$

Figure 1.6. Matrix representation of the higher-order factor model. y_j reflects the observed scores for a given indicator, μ_j reflects the intercept for a given indicator, λ_{j,FO_m} reflects the factor loading for a given indicator onto one of the common factors, η_{FO_m} reflects factor scores for a given common factor, ε_j reflects indicator-specific errors, $\Gamma_{FO_m,SO}$ reflects the factor loading for a given common factor onto the higher-order factor, ξ reflects the factor scores for the higher-order factor, and ζ_{FO_m} reflects the errors specific to a given common factor.

The hierarchical and higher-order models are not equivalent as was once thought (Gustafsson & Balke, 1993; Mulaik & Quartetti, 1997). The higher-order model is nested within the bifactor model (Yung, Thissen, & McLeod, 1999). Moreover, the general factors in each model differ in their relationship with the observed variables: the general factor in the bifactor model directly predicts what is common among the indicators, whereas the general factor in the higher-order model indirectly predicts commonalities among indicators via the common/first-order factors (Rindskopf & Rose, 1988).

While the difference may seem negligible at the statistical level, it has rather profound theoretical implications. In the bifactor model, the construct associated with the general factor, be it general intelligence or general psychopathology, does not underpin the specific domains, such as crystal and fluid intelligence or internalizing and externalizing psychopathology, as would be predicted by the higher-order model (Gignac, 2008). Instead, the general factor is assumed to directly cause the covariation in the observed variables, such as intelligence test scores or psychiatric assessment scores. Specific domains are also assumed to directly underpin the covariation in clusters of observed variables, but covariation that is not explained by the general construct (Reise, 2012). In more concrete terms, general psychopathology is an entity distinct from internalizing and externalizing problems.

The one instance where these models are equivalent is when a Schmid-Leiman transformation is used. Briefly, the Schmid-Leiman transformation is a method for deriving the direct effects of a higher-order factor on the observed variables (Schmid & Leiman, 1957). It is not a model per se, but a bifactor-like estimation tool that converts an exploratory correlated factors solution into a second-order solution, and then orthogonalizes the higher-order and first-order factors to derive the unique variances associated with each (Wolf & Preising, 2005).

While the Schmid-Leiman transformation decomposes the variance into general and specific sources, it does so with more constraints than the bifactor model. For instance, the general factor loadings are a product of the first-order and second-order factor loadings, and (hence) the ratio of the general factor loading to the specific factor loading is proportional within each specific factor (i.e. 'proportionality constraints'; Yung et al., 1999). The bifactor model is free from proportionality constraints, which is one reason why it generally fits the data better than the higher-order model (Gignac, 2016).

The main distinction between the hierarchical models described is that in the bifactor model, the general factor directly predicts the observed variables, whereas in the higher-order model, the general factor indirectly predicts the observed variables via the first-order factors. Humphreys (1962) argued that this distinction, based on 'distance' of the latent variable from the observed variables, is superficial. The general factors should instead be compared based on their 'breadth' of influence (i.e. how many variables they ultimately predict). Since both general factors influence a large range of indicators, either directly or indirectly, they should produce similar results (Gustafsson & Balke, 1993).

Gustafsson and Balke (1993) applied the bifactor model to a series of child aptitude tests that were originally analysed with a higher-order model, and found that the general factors in each were equally (most) predictive of grade outcomes (but the bifactor model was more parsimonious). Furthermore, Chen, West, and Sousa (2006) found that the general and specific well-being factors of a bifactor model showed almost identical predictions of external criteria as the first- and higher-order factors of a higher-order model.

Not all studies have shown such equivalence, however. For instance, Beaujean, Parkin, and Parker (2014) found that while the bifactor and higher-order models produced similar general factors of the Wechsler Intelligence Scale for Children, the specific factors and first-order factors differed in their meaning and prediction of language achievement. Moreover, Gignac (2008) reported superior fit of the bifactor model over the higher-order model across several child intelligence datasets, but it should be noted that the frequent superiority of the bifactor model over the higher-order model may, in part, be due to differences in model complexity (see section 3.5.5).

Overall, the 'similar but different' relationship between the bifactor and higher-order models has led some to conclude that "in some ways there is no meaningful distinction to be made between these two models, whereas in other ways, they are vastly different" (Reise, Moore, & Haviland, 2011, p. 547). It may be most helpful to view these models as similar but suited to different theoretical and practical contexts, rather than different representations of the truth in nature (Rodriguez, Reise, & Haviland, 2016b).

1.4 The Bifactor Model of Psychopathology

The bifactor model was overshadowed for many years by the correlated factors and higher-order models in intelligence research (Beaujean, 2015), which would have naturally limited its use in psychopathology research. The first published report applying the bifactor model to psychopathology data was by Gibbons and Hedeker (1992), who estimated a 'primary depression dimension' and four specific subscale factors in an item-level analysis of the Hamilton Depression Rating Scale. Since then, the bifactor model was judiciously applied to assessment scales of specific disorders to capture the multidimensionality caused by sampling a diverse item pool for a single construct (Reise, 2012).

Some researchers also applied the bifactor model to groups of disorders, such as depression, anxiety and somatic problems (Simms, Prisciandaro, Krueger, & Goldberg, 2012), antisocial and substance-related problems (Krueger, Markon, Patrick, Benning, & Kramer, 2007), and attention-deficit/hyperactivity disorder and oppositional defiant disorder (Martel, Gremillion, Roberts, von Eye, & Nigg, 2010) to estimate transdiagnostic factors like internalizing (i.e. inwardly-oriented problems, e.g., depression, anxiety) and externalizing (i.e. outwardly-oriented problems, e.g., aggression, substance-misuse). This trend was the impetus for Lahey et al. (2012) to apply the bifactor model to a range of internalizing and externalizing disorders, as these broad domains tended to be positively correlated in the same way that the disorders making up each broad domain were positively correlated (Krueger & Markon, 2006). Lahey et al. hypothesised that the correlation among internalizing and externalizing implies that these factors are influenced by a broader, more 'general' factor that reflects a shared set of aetiological factors distinct from the aetiological factors uniquely associated with internalizing or externalizing problems.

Lahey et al. (2012) estimated a model with three correlated factors, where the prevalence of major depression, dysthymia, and generalized anxiety disorder loaded onto a distress factor, phobias loaded onto a fear factor, and antisocial personality and drug and alcohol dependence loaded onto an externalizing factor. The correlated factors model was compared to a bifactor model with uncorrelated distress, fear, and externalizing factors and a general factor upon which all disorders loaded. This was in a population sample of 35,336 18-65 year-olds who were assessed over two timepoints in the National Epidemiologic Study of Alcohol and Related Conditions.

The bifactor model showed a significant improvement in model fit compared to the correlated factors model according to the chi-square difference tests; differences in other model fit indices, such as Akaike Information Criterion and Bayesian Information Criterion, were less pronounced. Furthermore, experiences of physical, sexual, or emotional abuse were solely predicted by the general factor, supporting Lahey et al.'s (2012) hypothesis that the general factor is associated with a set of broad vulnerability factors. Lahey et al. (2011) also found that a single factor underpinned genetic risk to multiple psychiatric disorders.

Lahey et al. (2012) can be credited with the first bifactor analysis of psychopathology[1], but it was Caspi and colleagues (2014) who provided a theoretical foundation for the notion of 'general psychopathology'. Caspi et al. (2014) analysed symptom counts of internalizing, externalizing, and psychotic disorders across ages 18, 21, 26, 32, and 38 in a representative cohort of 1,037 participants from the Dunedin longitudinal study. They found that a general factor and correlated specific internalizing and externalizing factors fit the data similarly to a model with correlated internalizing, externalizing and thought disorder factors alone.

While the bifactor and correlated factor models could not be distinguished on the grounds of model fit, several additional analyses demonstrated the value of

[1] Technically, Gibbons, Rush and Immekus (2009) were first to apply the bifactor model to a range of internalizing, externalizing, and thought disorder problems, but with the aim of investigating multidimensionality in the Psychiatric Diagnostic Screening Questionnaire, rather than demonstrating a general psychopathology factor.

the bifactor model. For example, the positive correlation between internalizing and externalizing factors that is commonly seen in correlated factor models was negative in the bifactor model, suggesting that, in line with Lahey et al.'s (2012) hypothesis, internalizing and externalizing problems are opposing sets of problems after accounting for their shared aetiology. Furthermore, Caspi et al. (2014) found that experiences of childhood maltreatment and parental psychopathology were uniquely associated with the general factor, further supporting its role as a shared aetiology factor. There are, of course, issues with Caspi et al.'s model which are discussed in Chapter 2.

Caspi and colleagues (2014) named the general factor of psychopathology the *p* factor, after the *g* factor of general intelligence that describes the positive manifold among intelligence tests. In the same way that the *g* factor summarizes the consistencies in people's performance across a range of intelligence tests, the *p* factor summarizes people's propensity to experience a range of common mental health problems. The notion of a *p* factor is not without criticism (van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017; see Chapter 2), but it challenges current psychiatric nosologies that categorize people's problems into distinct entities or disorders. While it is undeniable that classification systems like the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013) and International Classification of Diseases (ICD; World Health Organization, 2018) have helped organize mental distress in a pragmatic way, we must not forget that disorders have ultimately been constructed, as have psychometric factors, but based on consensus among committee members (who each carry private and political interests) rather than empirical data. As a result, diagnostic cut-offs often lack reliability, people with the same diagnosis can present with different problems, and comorbidity among diagnoses is the rule rather than the exception (Nesse & Stein, 2012).

The bifactor model of psychopathology offers a testable classification of mental health problems that includes people's overall severity or liability to any and all forms of problems (*p* factor), as well as the specific and gendered ways in which people react to stress (e.g., internalizing, externalizing; Caspi & Moffitt, 2018). It is not about 'lumping' or 'splitting' mental health problems but testing the unique contribution of each. The *p* factor may resolve some of the issues associated with current nosologies, such as the difficulty in identifying specific biomarkers or developing tailored treatments for individual disorders (Insel, 2014; Wampold & Imel, 2015).

1.5 Thesis Aims

The bifactor model of psychopathology has received a surge of interestin recent times–Caspi et al.'s (2014) seminal paper has been cited 926 times in past five years (Google Scholar, 31.08.2019). There is hope that the bifactor model, and the quantitative movement in general, will bring a new era of evidence-based research to psychiatry (Kotov et al., 2017; Krueger et al., 2018). However, we still have much to learn about the bifactor model, both in terms of its methodological properties and its clinical applications. The current thesis aims to evaluate the strengths and limitations of the bifactor model as a measurement tool, and to explore its utility when applied to clinical outcomes.

The next chapter provides a review of bifactor studies to date in terms of the nature and development of the *p* factor (Chapter 2). The following two chapters address methodological issues, including predictors of variability in the reliability of *p* factors in different studies (Chapter 3), and how much response biases contribute to the *p* factor (Chapter 4). The final two chapters apply the bifactor model to clinical outcomes data to determine how the bifactor dimensions change over a psychosocial intervention (Chapter 5), and whether the bifactor dimensions of personality disorder have any prognostic value for predicting depression outcomes over an inpatient treatment (Chapter 6). Finally, the results from these chapters are synthesised in Chapter 7, followed by a discussion of their limitations and implications for future research.

Chapter 2 Systematic Review: Examining the Structure and Development of the *p* Factor

Caspi et al. (2014) proposed two hypotheses of the *p* factor: a structural hypothesis, where *p* is seen as a severity dimension defined by the degree to which thoughts are disordered, and a developmental hypothesis, where people differ in the extent to which they progress through the continuum of severity. I will use these two hypotheses to organize a systematic review of bifactor studies of psychopathology to date.

After outlining the search strategy, I will review different structural hypotheses of the *p* factor (i.e. hypotheses about what the *p* factor represents), including the thought disorder continuum hypothesis, emotion dysregulation hypothesis, and universal suffering hypotheses. I will then evaluate Caspi et al.'s (2014) developmental hypothesis against longitudinal studies of the bifactor model to date, based on how mean levels of variability explained by the general and specific psychopathology factors differ between age groups (i.e. absolute stability) and the consistency in people's scores on the general and specific factors over time (i.e. differential stability).

2.1 Method

2.1.1 Search Strategy

A literature search was conducted using PubMed to identify studies that applied the bifactor model to psychiatric symptoms or disorders. Search terms included (*bifactor* OR *bi-factor* OR *nested factor* OR *p factor*) AND (*psychopathology* OR *psychiatr** OR *disorder* OR *symptom* OR *diagnosis* OR *mental health*). The search produced 296 results since 2006, with most of the relevant studies published from 2015 onwards. Studies were also identified with a citation search of Caspi et al.'s (2014) seminal paper using Google Scholar.

Studies were included if: (i) they modelled more than one disorder-domain with a bifactor model (e.g., studies that analysed depression and substance problems include two domains; internalizing and externalizing problems), (ii) they used confirmatory factor analysis, and (iii) they provided a standardized factor loading matrix. Given that the analysis of psychopathology data with bifactor models is relatively new, studies were not excluded based on the type of estimator used (e.g., maximum likelihood vs. weighted least-squares), whether multiple scales were used rather than a single assessment measure, whether their solution included cross-loadings or specific factor correlations, or whether their analysis was at the item- or subscale-level. The bifactor model in each study was close to or surpassed an acceptable fit (e.g., Comparative Fit Index [CFI] and Tucker-Lewis Index [TLI] ⩾ .9 and Root Mean Square Error of Approximation [RMSEA] ⩽ .08; Hu & Bentler, 1999). A total of 49 studies published between 2009 and 2019 that met the above criteria were included in the current review. A full list of studies can be found in Appendix A.

2.2 Structural Review: What is the *p* Factor?

2.2.1 A Continuum of Disordered Thought

Caspi et al. (2014) proposed that the *p* factor reflects a dimension of severity differentiated by the extent to which thoughts become disordered. At the extreme end of the spectrum, individuals may experience uncontrollable and irrational thoughts that are characteristic of, but not limited to, the psychoses. For example, in addition to hallucinations and delusions, uncontrollable worry, intrusive beliefs and images, paranoid and hostile attributions, irrational fears and beliefs, body image disturbances, and dissociative experiences accompany a range of severe presentations (Caspi et al., 2014). These experiences share an involuntary and unfiltered quality to them that distorts reality (Caspi & Moffitt, 2018).

Caspi et al.'s (2014) disordered thought hypothesis emerged from the finding that their specific thought disorder factor was subsumed by the *p* factor. That is, while a thought disorder factor was identified alongside internalizing and externalizing factors in a correlated factors model, it did not retain any reliable variance once the general variance was estimated. Specific factors that load to unity with a general factor are thought to represent or define the general factor in some way (Koch, Holtmann, Bohn, & Eid, 2018). Therefore, the general psychopathology factor may be defined by a continuum of disordered thought.

Further support for the disordered thought hypothesis comes from a study by Laceulle, Vollebergh, and Ormel (2015), who found that a bifactor model (with *p* and uncorrelated internalizing, externalizing, thought disorder specific factors) was not identified, whereas a revised bifactor model (with *p* and correlated internalizing and externalizing specific factors, with thought disorder items loading directly onto *p*) provided an adequate and more parsimonious fit than a three-factor model (with correlated internalizing, externalizing, and thought disorder factors). These findings are based on self-reported symptoms in a community sample of 2,230 Dutch 11-19 year old adolescents, and were replicated with parent-reported symptoms. Despite differences in sample characteristics and measures, Lacuelle et al. replicated the unique contribution of thought disorder items to *p*. Since then, others have also reported that thought disorder items loaded uniquely onto the *p* factor (Bloeman et al., 2018; Calkins et al., 2015; Hankin et al., 2017; Lahey et al., 2017; Rosenström et al., 2018; Urbán, Arrindell, Demetrovics, Unoka, & Timman, 2016; Urbán et al., 2014).

A similar finding has also been reported in the child literature, where problems in social communication, social cognition, and autistic mannerisms load directly onto the *p* factor rather than a separate 'autism' factor (Martel et al., 2017; Neumann et al. 2016). Caspi et al. (2014) noted that disordered thought problems of a psychotic nature are featured in all but childhood disorders; however, disordered thought of an autistic kind may define the continuum of severity in childhood. A key cognitive ability that develops during childhood and is implicated in autism spectrum disorders is theory of mind or mentalization, i.e. the ability to interpret one's own and other's behaviours in terms of intentional mental states (Frith & Frith, 2003; Fonagy, Gergely, Jurist, & Target, 2002). Problems with mentalizing in childhood may not be limited to autism spectrum disorders, as was initially thought (e.g., Baron-Cohen, Leslie, & Frith, 1985). Instead, problems in mentalizing may be most pronounced in autism spectrum disorders, but shared across a range of child emotional and behavioural problems (Gray, Jenkins, Heberlein, & Wegner, 2011). Mentalizing, which broadly falls under meta-cognition, might also break down in psychotic-like experiences during adolescence and adulthood, where an awareness of the self, other, or reality is compromised (Lysaker, Gumley, & Dimaggio, 2011).

While the disordered thought hypothesis is clinically plausible, it emerged from a methodological phenomenon (e.g., a specific factor loading to unity with the general factor) that is subject to measurement differences. Demonstrating this point, Carragher et al. (2016) found that a bifactor model with a specific thought disorder factor in addition to correlated internalizing and externalizing factors converged and fit the data best in a community sample of 2,175 13-year olds. Moreover, thought disorder items loaded healthily onto the specific thought disorder factor and *p* (favouring the former). Carragher et al. (2016) argued that their ability to model the specific thought disorder factor may be due to including a greater number of thought disorder items compared to Caspi et al. (2014), which increased the thought disorder factor's reliability. This may also explain why Laceulle et al. (2016) were unable to identify the specific thought disorder factor, as they only included three disorder-level items.

It is important to point out that there are clear signs of scale effects in Carragher et al.'s (2016) study: items from the scale loaded similarly onto the general and specific factors, while items with similar content but originating from different scales loaded differently onto the general and specific factors. Therefore, it may be that the thought disorder factor's reliability was inflated by the fact that the items were from the same scale. However, most studies that include thought disorder items report that they load onto a specific thought disorder factor as well as the *p* factor (Afzali, Sunderland, Carragher, & Conrod, 2017; Arrindell et al., 2017; Haltigan et al., 2018; Hyland et al., 2018; Jones et al., 2018; Martel et al., 2017; Niarchou et al., 2017; Pettersson et al., 2019; Romer et al., 2017; Schaefer et al., 2018; St Clair et al., 2017; Stochl et al., 2015; White et al., 2017). Still, thought disorder items load strongly onto *p*, supporting the idea that thought problems are discriminative of the severity continuum in some capacity.

The preferential loading of autism-related items on *p* has not always been replicated either. For instance, Noordhof et al. (2016) found that in a sample of 2,230 11 year-olds, a separate autism factor was necessary to preserve the superior fit of a revised bifactor model that featured specific internalizing, externalizing, and attention/orientation factors compared to a correlated factors model. Furthermore, autism items were no more representative of the *p* factor than the other items. Noordhof et al. (2016) used the Child Behavioural Social Questionnaire which includes more items associated with autism than the Social Responsiveness Scale-Short Form used by Neumann et al. (2016). Others have also shown that a specific autism factor is identifiable (Bloeman et al., 2018; Pettersson et al., 2019). Therefore, the ability to model a discrete autism factor that reflects the severity of disordered thought may be an issue of power rather than theory.

2.2.2 A Disposition to Emotional Distress

An alternative hypothesis is that the *p* factor reflects a general disposition to emotional distress or a reactivity to emotions (Carver, Johnson, & Timpano, 2017; Lahey et al., 2012). In personality research, the trait 'neuroticism' and temperament 'negative emotionality/affectivity' refer to an ease in experiencing arousing emotions and a difficulty relieving oneself from them (Eysenck & Eysenck, 1985). Neuroticism and negative affectivity have strong cross-sectional and prospective links to internalizing disorders (Griffith et al., 2010; Jeronimus, Kotov, Kiese, & Ormel, 2016) and externalizing disorders (Eisenberg et al., 2009; Jeronimus et al., 2016), making them potential markers of *p*.

Supporting this hypothesis, Tackett et al. (2013) found significantly stronger correlations between self-reported negative emotionality and *p* compared to other in 1,569 9-17 year-old twin pairs. Furthermore, negative emotionality was more strongly correlated with *p* than it was with specific internalizing or externalizing factors, demonstrating a special link between the two. However, these effects were most pronounced for parent-reported factors: the relationship between *p* and negative emotionality was r = .58 in the parent-reported model, but only r = .20 in youth-reported model. Still, others have replicated the stronger relative relationship between *p* and neuroticism/negative emotionality in children, adolescents and adults (Carragher et al., 2016; Caspi et al., 2014; Geeraerts et al., 2015; Hyland et al., 2018; Miller et al., 2019; Neumann et al., 2016; Olino, Dougherty, Bufferd, Carlson, & Klein, 2014; Weissman et al., 2019).

Nonetheless, the relationship between *p* and neuroticism/negative affectivity may not be as strong as is apparent. For example, some have reported that neuroticism and negative affectivity are more strongly correlated with specific internalizing than *p* (Castellanos-Ryan et al., 2016; Hankin et al., 2017). Furthermore, the relationship between neuroticism and *p* may be limited to parent-reported measures rather than self-reported or observational measures (Olino et al., 2014; Geeraerts et al., 2015). It might be that neuroticism shows relatively stronger relationships with *p* (and sometimes internalizing) because of content overlap rather than substantive overlap. That is, neuroticism scales and depression and anxiety scales share similar items, such as 'is depressed, blue', 'worries a lot' and 'gets nervous easily' (from the Big Five Inventory; John & Srivastava, 1999).

The *p* factor also shows negative associations with the trait 'conscientiousness' and temperament 'effortful control', which reflect self-directedness and self-control over emotions and behaviour (MacDonald, 2008). While conscientiousness and neuroticism are separate traits, they are often negatively associated (Rothbart, Ellis, Rueda, & Posner, 2003). For example, greater emotional reactivity is often accompanied by poorer self-regulation, but someone can show high (or low) regulation in the face of high (or low) reactivity (Rydell, Berlin, & Bohlin, 2003). Caspi et al. (2014) found that in addition to higher levels of neuroticism, *p* was associated with lower levels conscientiousness. Others have also reported a negative relationship between *p* and effortful control in children and adolescents (Deutz et al., 2018; Neumann et al., 2016; Olino et al., 2014; Hankin et al., 2017; Snyder et al., 2017).

Several authors have also reported a moderate negative association between *p* and performance on executive function tasks (Bloeman et al., 2018; Caspi et al., 2014; Castellanos-Ryan et al., 2016; Harden et al., 2019; Martel et al., 2017; Neumann et al. 2016; White et al., 2017). While executive function tasks assess 'cool' cognition that lacks emotional valence, they tap the domain-general ability to control thoughts and actions (Miyake & Friedman, 2012). Indeed, a range of psychiatric problems are associated with poorer executive function performance (Snyder, Miyake, & Hankin, 2015), suggesting that 'cool' executive control processes overlap with 'hot' emotion regulation processes (Zelazo & Cunningham, 2007). A similar case can be made for general intelligence and academic attainment, both of which negatively correlate with *p* (Caspi et al., 2014; Castellanos-Ryan et al., 2016; Constantinou et al., 2019; Harden et al., 2019; Lahey et al., 2015; Patalay et al. 2015; Pettersson et al., 2018; Sallis et al., 2019).

Collectively, the personality and neuropsychology studies reviewed suggest that *p* is associated with various measures that would all contribute to a disposition towards emotional distress. They may also contribute to general dysregulation across domains (e.g., cognitive, affective, behavioural). It is important to note that these studies did not separate out the general and specific variance in trait measures. Therefore, the *p* factor's associations with neuroticism and effortful control may in fact be driven by 'general personality', the '*g*' factor of the personality domain (Musek, 2007). This may also account for the breadth of associations observed between *p* and other, non-affective domains. It may be that general psychopathology will no longer correlate with specific personality and performance traits once the general factors of personality, executive function, and intelligence are controlled for. While the association between general psychopathology and personality has been hypothesised, it has yet to be tested (Widiger & Oltmanns, 2017).

2.2.3 A Universal Expression of Human Suffering

A final hypothesis is that the *p* factor is a consequence (rather than cause) of mental health problems. That is, *p* captures individual differences in the level of suffering caused by the problems people face, each of which has a separate but interrelated cause[2] (van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017). Variation in certain behaviours and experiences (e.g., hearing voices or worrying) are not problematic in and of themselves, but rather, cause impairment when they prevent the fulfilment of fundamental life tasks, such as holding down a job or maintaining steady relationships with family, friends, and romantic partners (Livesley, 2011). People who are susceptible to the impact of life stressors may express impairment in a similar way, which we often refer to as 'depression' (Monroe & Reid, 2009).

Consistent with this hypothesis, major depressive disorder has among the highest comorbidity rates with other disorders: 72% of respondents with a lifetime history of depression in the National Comorbidity Survey met the criteria for at least one other DSM-IV disorder, which was not limited internalizing disorders (e.g.,

[2] The underling theory behind this argument, network theory, is described more fully in Chapter 3 (see section 3.5.2).

generalized anxiety disorder, social anxiety disorder, panic disorder, agoraphobia, specific phobia, obsessive-compulsive disorder, post-traumatic stress disorder) but included externalizing problems, such as alcohol dependence, drug dependence, gambling dependence, antisocial personality disorder, and bulimia (Kessler et al., 2003). These high co-occurrence rates are also found in children and adolescents (Birmaher et al., 1996). Comorbidity rates increase monotonically with depression severity (Kessler, Zhao, Blazer, & Swartz, 1997), and depression is more likely to emerge after the comorbid disorder rather than before (Rohde, Lewinsohn, & Seeley, 1991). There may still be cases where depression is a primary condition with an early onset, and for which comorbid problems develop after its onset, but such cases appear to be a minority (Alpert et al., 1999).

There are currently no studies investigating the role of depression in defining the *p* factor. However, a common finding is that depression items load most strongly onto the *p* factor[3], and show a large drop in loading strength on the internalizing factor (Arrindell et al., 2017; Black, Panayiotou, & Humphrey, 2019; Brodbeck et al., 2014; Calkins et al., 2015; Carragher et al., 2016; Caspi et al., 2014; Constantinou et al., 2019; Conway et al., 2019; Gomez, Stavropoulos, Vance, & Griffiths, 2019; Haltigan et al., 2018; Hyland et al., 2018; Jones et al., 2018; Lahey et al., 2012; Lahey et al., 2015; Lahey et al., 2017; Liu, Mustanski, Dick, Bolland, & Kertes, 2017; Martel et al., 2017; Miller et al., 2019; Olino et al., 2014; Rytila-Manninen et al., 2016; Pettersson et al., 2019; Pezzoli et al., 2017; Preti, Carta, & Petretto, 2019; Romer et al., 2017; Schaefer et al., 2018; Snyder et al., 2017; St Clair et al., 2017; Tackett et al., 2013; Urbán et al., 2016; Urbán et al., 2014; Wade, Fox,

[3] Differences in the strength with which indicators load onto the *p* factor are discussed in Chapter 3 (see section 3.5.7).

Zeanah, & Nelson, 2018). These findings suggest that depression is representative of the *p* factor, and hence may be better thought of as a general distress factor (Kim & Eaton, 2015). The reader may recall that thought disorder problems were also representative of *p*, but this fits the universal suffering hypothesis in that psychotic experiences may also be a universal expression of moderate-to-severe distress (but one that is moderated by societal norms; Wüsten et al., 2018).

To conclude, the *p* factor may reflect a continuum of severity defined by the degree of disordered thought, a disposition to emotional distress, or a universal response to human suffering. There is much overlap between the first two hypotheses, as they both refer to a fundamental difficulty in regulating cognitive, emotional, and behavioural states. The real challenge will be in resolving the third hypothesis and determining whether the positive manifold in psychiatric problems is a cause or consequence of a general latent factor. Nonetheless, the bifactor model would not be rendered useless, even if the *p* factor reflects a common consequence of distress. The finding that human suffering manifests in universal and dimensional ways is an important one, forcing us to think about mental health and wellbeing in a broader light (Antonovsky, 1987). Moreover, quantifying human suffering is no small feat, and offers the chance to study disorder-specific mechanisms free from general distress. Ultimately, further studies are necessary to tease out these hypotheses, knowing that both outcomes are fruitful.

2.3 Developmental Review: How Does the *p* Factor Change Over Time?

Caspi et al. (2014) extended their disordered thought hypothesis to include a developmental component. To recap, the continuum of severity includes single internalizing and externalizing disorders that are relatively non-impairing at the low end of the spectrum, and multiple co-occurring internalizing and externalizing problems accompanied by disordered thought at the high end. Caspi et al. argued that most people in the population will have experienced a brief episode of single internalizing or externalizing disorder at some point in their lives (e.g., Moffitt et al., 2010). However, some will go on to develop more widespread and persistent internalizing or externalizing disorders, and a minority will progress to develop a psychotic condition in adolescence or adulthood. Put simply, people differ in the extent that they progress through the continuum of severity over time.

Two testable predictions emerge from Caspi et al.'s (2014) developmental extension of the disordered thought hypothesis. First, comorbidity will increase with age. That is, internalizing and externalizing problems will be experienced as isolated entities in early childhood, but variation in the extent that people accumulate internalizing and externalizing disorders will increase in later childhood into adolescence, until people vary in the extent to which they are susceptible to the full range of problems in adolescence and adulthood.

The second prediction is that higher *p* scores, hence severity, will be associated with stronger heterotypic development of psychopathology, i.e. more severe cases will move in and out of different disorders over time rather than present with the same problem, the latter known as homotypic development. Mental health problems do not just co-occur simultaneously, but also sequentially. For instance, mood disorders subsequently predict substance abuse, although the causal mechanisms may be mixed (Cerdá, Sagdeo, & Galea, 2008). Sequential comorbidity is a marker of severity (Moffitt et al., 2007). Therefore, people with higher *p* scores are predicted to experience a greater number of disorders over time, or report a broader history of problems, compared to people with lower *p* scores.

These predictions reflect two forms of longitudinal stability: absolute and differential stability (Morey & Hopwood, 2013). Absolute stability refers to average changes in a construct within a sample over time, whereas differential stability refers to consistency in people's rank ordering over time. For example, if a group of depressed patients showed large declines in depression scores on average between two time-points, then the sample would show low absolute stability, in that the overall level of the construct (e.g., depression) changes over time. However, if the patients who scored highest at baseline also scored highest at the end-point, and patients who scored lowest at baseline also scored lowest at the end-point, then the sample would show high differential stability because the rank ordering in patients' scores remained consistent over time.

We can examine absolute stability in the general and specific factors by comparing changes in the explained common variance (ECV) over different age groups (see Murray, Eisner, & Ribeaud, 2016, for a similar analysis using omega hierarchical, another reliability index). The ECV reflects the amount of variance in a measure explained by the general (*p*) factor, relative to the total amount of variance explained by all factors modelled (e.g., both general and specific psychopathology factors). It can be used to index the strength of the general (*p*) factor in explaining individual differences in a measure (Rodriguez, Reise, & Haviland, 2016b). ECV values were computed from the standardized factor loading matrices in each study reviewed, the results of which are covered more fully in Chapter 3. I am simply burrowing from that chapter to address a specific question about differences in *p* factor strength across different age groups.

According to Caspi et al.'s (2014) hypothesis, we would expect that the amount of variance in a measure explained by the *p* factor would be lowest in early childhood and increase into adulthood. Consequently, the variance explained by specific factors should be highest in early childhood and decrease into adulthood.

We can assess differential stability in the general and specific factors by comparing test-retest correlations (e.g., correlating factor scores for the same group of people at two different time-points) or autoregressive and cross-lag coefficients (regressing a factor at one time-point onto itself or another factor, respectively, at a preceding time-point). An advantage of the latter method is that the predictive influence of one factor on another can be estimated whilst controlling for spurious dependencies that arise due to autocorrelation within each factor (Jones, Ghannam, Nigg, & Dyer, 1993).

If *p* reflects a vulnerability to psychopathology, as Caspi et al. (2014) suggest, then we would expect it to show moderate-to-strong homotypic (i.e. within-factor) differential stability. In other words, individual differences in the propensity to develop mental health problems will be stable over time. We would also expect *p* and specific factors to show heterotypic (i.e. between-factor) associations, but the direction of prediction should change with age. For example, specific factors in childhood may positively predict *p* at later time-points, but not vice versa, if specific expressions of isolated internalizing and externalizing disorders are risk factors for later comorbid problems. However, in later childhood and adolescence, *p* may predict specific factors at later time-points, but not vice versa, perhaps in sporadic or cyclical ways, if greater severity is a risk factor for heterogeneous expressions of specific problems.

2.3.1 Absolute Stability Across Age Groups

Figure 2.1 plots the *p* factor's ECV against the mean sample age for each bifactor study reviewed.

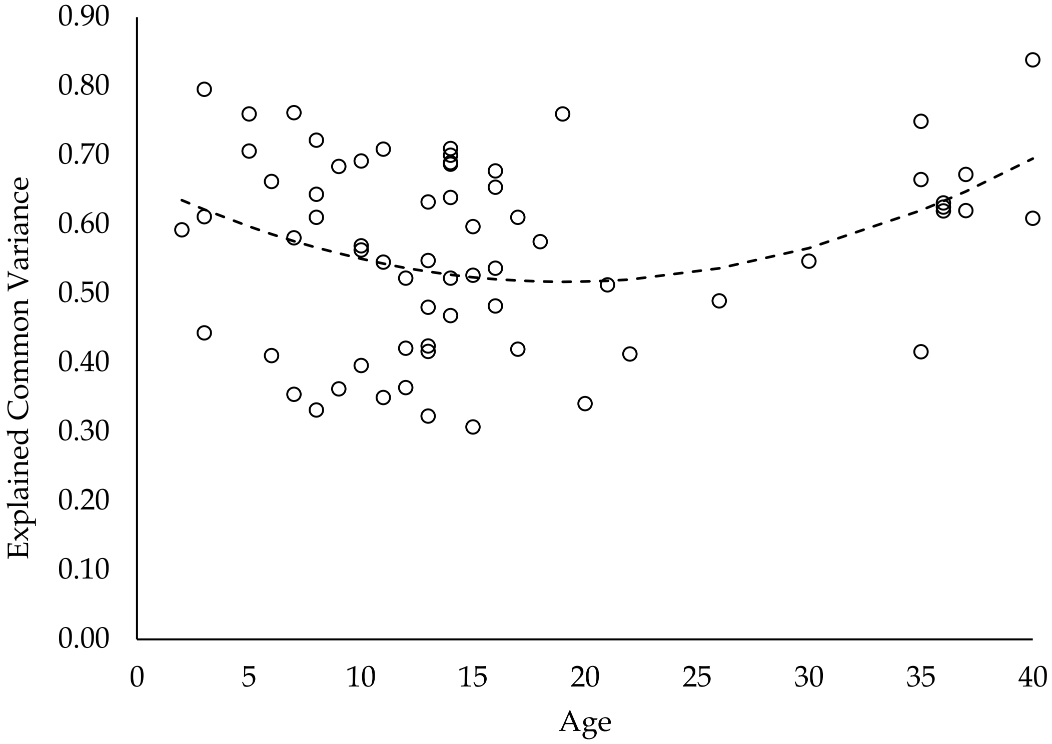

Figure 2.1. Explained Common Variance for the *p* factor plotted against the average sample size in each study reviewed.

The mean ECV in childhood (ages 2-12; *k* = 27) was 0.56 (*SD* = 0.14). Therefore, the *p* factor accounted for 56% of the common or modellable variance in psychopathology measures on average during childhood. This leaves 44% of the common variance explained by specific psychopathology factors, which is a sizeable amount but less than would be expected if variation in specific problems is highest in childhood.

The ECV in adolescence (ages 13-17; *k* = 22) was similar to the estimate in childhood (*M* = 0.54, *SD* = 0.12). However, the pattern of ECVs across studies appeared to subtly decline from infancy to late adolescence/early adulthood (see Figure 2.1), contradicting the hypothesis that the *p* factor strengthens with age, at least in young people. The subtle decline in what is sometimes purported to be a 'general vulnerability' factor is also surprising given the elevated onset of mental health problems in adolescence (Kessler et al., 2007).

The mean ECV in adulthood (ages 18-40; k = 16) was 0.60 (*SD* = 14), meaning that 60% of the variation in psychopathology outcomes during adulthood could be explained by the *p* factor. Furthermore, the ECV appeared to steadily increase over adulthood, partly fitting the notion that the strength of comorbidity increases with age. Nonetheless, ECV values followed a U-shaped pattern across all age groups, initially declining from infancy to late adolescence and then increasing over adulthood, at least until the age of 40. Linear and quadratic slopes of age weakly but significantly predicted ECV values (b_{linear} = -.02, t = -2.23, p = .029, 95% CI [-.03, -.002]; $b_{\text{quadratic}}$ = .0004, t = 2.58, p = .029, 95% CI [.000009, .0007]).

It is improper to analyse these data with a regression analysis since the errors are heteroscedastic, each data-point is unweighted for its sample size, and there are dependencies in the datapoints belonging to the same study. It is also improper to draw conclusions about developmental processes from the analysis of cross-sectional data. However, this preliminary analysis offers some 'food for thought' about age-related changes in the *p* factor. For example, the quadratic change in *p* factor strength might reflect differentiation in the expression of psychopathology. In the same way that the positive correlations among intelligence tests decrease in more intelligent (higher *g*) cases following the law of diminishing returns (Blum & Holling, 2017; Spearman, 1927), the positive correlations among mental health problems may decrease in higher *p* cases during adolescence, as the expression of psychopathology becomes more defined or 'differentiated'. Comorbidity in the specific expression of disorders may then increase into adulthood.

Alternatively, the quadratic change in *p* factor strength may reflect difficulties in assessing the dynamic nature of mental health problems during adolescence, or age-related changes in the design of psychopathology measures (e.g., adult assessment measures may be explicitly or implicitly suited to estimating a stronger *p* factor). I explore these issues further in Chapter 3. Suffice to say, comorbidity does not appear to increase monotonically from childhood to adulthood, as implied by Caspi et al.'s (2014) developmental progression hypothesis.

2.3.2 Differential Stability in Childhood

Olino et al. (2018) estimated an orthogonal bifactor model with a general *p* factor and specific internalizing and externalizing factors at ages 3 and 6 based on 541 caregiver's reports of DSM-IV disorders. After establishing partial measurement invariance (the factor models were similar across ages, but items loaded differently on each factor at different ages), they found that the *p* factor and externalizing factor showed moderate differential stability (autoregressive coefficients: B = .51 and B = .50, respectively), while the internalizing factor showed high differential stability (B = .85). Therefore, the *p* factor showed some stability that is perhaps less than expected if it reflects a stable vulnerability to mental health problems. Nonetheless, the autoregressive coefficients for both *p* and externalizing may have been underestimated due to differences in the way that indicators related to the factors over time, rather than changes in the way that children were ranked on each factor.

This may also explain why internalizing showed high stability, as it was associated with a narrow set of fear disorders (e.g., specific phobia, panic disorder) that loaded similarly on the internalizing factor over time.

The internalizing and externalizing factors in Olino et al.'s (2018) study showed weak cross-lagged predictions of the *p* factor (.01 and .08, respectively), while the *p* factor showed a weak but significant cross-lagged prediction of externalizing ($B = .18$, $p < .05$) and a negative prediction of internalizing that did not reach significance ($B = -.13$, $p > .05$). These findings suggest that heterotypic stability stems from the *p* factor rather than specific factors, contrary to our predictions, but is minimal in early childhood.

McElroy, Belsky, Carragher, Fearon, and Patalay (2017) estimated a bifactor model with a general *p* factor and specific internalizing, externalizing, and attention factors at ages 2-14 using 1,253 caregiver's reports on the Child Behavior Checklist (CBCL). The *p* factor showed moderate-to-large autoregressive coefficients across age (median $B = .70$) which subtly increased over time (e.g., ages 2-3 = .66, 3-5 = .52, 6-8 = .69, 8-9 = .75, 10-11 = .76, 11-14 = .64), aside from coefficients between ages 3-5 and 11-14. The medium-sized coefficient between ages 3-5 replicates Olino et al.'s (2018) coefficient for a similar age range (3-6), but McElroy et al.'s finding was more likely a result of using a different form of the CBCL: a change in items would cause changes in factor loadings and hence differences in factor measurement rather than people's rank ordering (a similar drop was observed for all factors). The drop in differential stability between ages 11-14 may be explained by the fact that longer measurement intervals weaken autoregressive coefficients (Roberts & DelVecchio, 2000).

McElroy et al. (2017) also found low-to-moderate differential stability in the specific externalizing factor (median = .38, range = .23-.48 or .29-.48 excluding age 3-5), internalizing factor (median = .48, range = .26-.55 or .39-.55 excluding age 3-5), and attention factor (median = .46, range = .23-.55 or .35-.48 excluding age 3-5). The weaker differential stability in the specific factors compared to the *p* factor may be due to measurement error: specific factors generally show less reliability than the general factor when a measure is essentially unidimensional (Reise et al., 2010).

As for heterotypic differential stability, the specific internalizing and externalizing factors weakly but significantly predicted *p*. For example, the internalizing factor positively predicted *p* between the ages 3-5 (*B* = .09), 8-9 (*B* = .10), and 10-11 (*B* = .06), while the externalizing factor positively predicted *p* between the ages 3-5 (*B* = .13), 5-6 (*B* = .17), 9-10 (*B* = .08), and 11-14 (*B* = .09). Nonetheless, the *p* factor weakly predicted both internalizing and externalizing factors in a cyclical fashion. These findings suggest that specific expressions of internalizing and externalizing problems predicted a higher risk of comorbidity at later time-points, as predicted by Caspi et al.'s (2014) hypothesis, but the reverse was also true. However, the coefficients are too weak to infer theoretical significance.

In sum, the few studies that have investigated differential stability in the bifactor dimensions during childhood demonstrate moderate homotypic stability in the general and specific factors, supporting the notion that *p* reflects a stable vulnerability to psychopathology. However, the weak and bi-directional heterotypic stability between general and specific psychopathology factors does not support the prediction that individual differences in *p* would first be predicted by specific problems in childhood.

2.3.3 Differential Stability in Adolescence

Snyder, Young, and Hankin (2017) modelled an orthogonal bifactor model with *p*, internalizing, and externalizing factors in a community sample of adolescents at ages 13.5 and 15 years, using child and parent reports on a variety of self-report measures (factors were estimated using indicators from each informant to control for response biases). Their *p* factor showed a strong autoregressive coefficient (B = .86, p < .001). That is, *p* factor scores at age 13.5 accounted for roughly 74% of the variance in *p* factor scores at age 15. The internalizing and externalizing factors also showed strong autoregressive coefficients (B = .71, p < .001 and B = .72, p < .001, respectively). However, there were no significant cross-lag relationships between factors: *p* factor scores at age 13 weakly predicted internalizing (B = .04) and externalizing (B = .10) at age 15, while externalizing and internalizing at age 13 weakly predicted *p* (B = .05 and .07, respectively). These findings suggest that the general and specific psychopathology factors are highly stable and trait-like but differentiated between early and mid-adolescence.

McElroy, Shevlin, and Murphy (2017) also observed strong homotypic differential stability in the psychopathology factors during adolescence but using a person-centred analysis. The studies reviewed so far use variable-centred latent modelling techniques, which describe the relationships between variables using latent variables. By contrast, person-centred modelling approaches identify latent groups of people who show similar relationships among variables. McElroy et al. identified four latent groups in a community sample of children at ages 7.5 and 14 years old using parents reports on the DAWBA, including a low symptom (normative) group, high internalizing group, high externalizing group, and comoribd internalizing and externalizing (*p*) group.

Eighty percent of children were predicted to remain within the same latent group between ages 7.5 to 14, but 68% of them were from the normative group. Focusing on the non-normative groups, 56% of children in the comorbid *p* group at age 7.5 stayed within the group at age 14, which mirrors the autoregressive coefficients showing that roughly 60% of the variance in *p* scores can be explained by prior *p* scores. The most frequent heterotypic transition between latent groups was between the normative and internalizing groups, where 74% of children were predicted to move from the normative group at age 7.5 to the internalizing group at age 14, and 88% of children were predicted to move from the internalizing group at age 7.5 to the normative group. This finding mirrors Caspi et al.'s (2014) prediction that most people in the population will experience an isolated internalizing (or externalizing) problem. The percentage of children moving between *p* and internalizing/externalizing groups was minimal, which mirrors variable-centred analyses in childhood and adolescence showing weak heterotypic continuity between general and specific factors.

The high differential stability in general and specific psychopathology dimensions during adolescence may be surprising since this is a period of great biopsychosocial change (Sawyer et al., 2012). A study by Castellanos-Ryan et al. (2016) addresses this issue. The authors estimated the stability in general, internalizing and externalizing factors in a community sample of adolescents at ages 14 and 16 based on adolescent and parent reports on the DAWBA. They too found strong and significant autoregressive correlations (e.g., r_p = .73, *p* < .001; $r_{externalizing}$ .62, *p* < .001; $r_{internalizing}$ = .54, *p* < .001), and weak and non-significant heterotypic-factor correlations (e.g., *r*s < .03; but the *p* factor at age 14 mildly and significantly correlated with internalizing at age 16; r = -.09, p = .033). However, not all indicators loaded consistently onto *p* over time: internalizing diagnoses (e.g., general anxiety disorder, depression, social phobias, panic phobias, and OCD) and substance misuse (e.g., number of drugs used and smoking frequency) increased in loading strength from age 14 to 16. Therefore, whilst adolescents' relative standing on the *p* factor remained stable over time, the symptoms that defined *p* were more unstable, which might be due to differences in measurement properties of the DAWBA across age, but also a function of heterotypic change in psychopathology during adolescence.

It would be wrong to argue that all studies have shown high levels of homotypic continuity in the bifactor dimensions. Murray, Eisner, and Ribeaud (2016) estimated the *p* factor and specific internalizing, aggression, ADHD, and prosociality factors almost annually in a community sample of 1675 children between the ages of 7-15. The *p* and specific factors were estimated using a Schmid-Leiman transformation of an exploratory higher-order factor matrix, which the authors argued would prevent over-estimation of the *p* factor loadings. Autoregressive coefficients for all factors between adjacent years were low, particularly for the *p* factor (average *B* =.29, range = .12-.43), internalizing (average *B* =.32, range = .18-.42), and prosociality factors (average *B* =.32, range =.07-.50). Autoregressive coefficients for the aggression factor (average *B* =.41, range = .20-.56) and ADHD factor (average *B* =.47, range = .23-.61) were moderate. Contrary to the other studies reviewed, Murray et al.'s findings suggest that the general and specific psychopathology factors show low homotypic differential stability over time, even if the variance accounted for by the *p* factor is characteristic of a general latent trait (average omega hierarchical = .69; Gignac, 2014).

Nonetheless, Murray et al.'s (2016) low stability coefficients may reflect changes in the way that symptoms related to each factor over time, rather than instability in adolescents' relative standing on each factor. This is perhaps most apparent between ages 9 and 10, where the stability coefficient drops for all factors (e.g., *p* drops from .43 to .23), and factor loadings show large changes (e.g., some *p* factor loadings increase from.5 to .9). This would not affect the reliability estimates for each factor over time–which suggested that *p* explained the most variance in observed total scores–provided the factor loadings changed in compensatory ways for each factor. Therefore, without demonstrating some aspect of metric invariance, these findings may not be robust and subject to heterotypic differences in the way that *p* is defined over adolescence.

Wade, Fox, Zeanah, and Nelson (2018) also reported weak-to-moderate bivariate correlations among the bifactor dimensions at different time-points. They estimated a *p* factor and specific internalizing and externalizing factors at ages 8, 12, and 16 in 220 young people, half of whom had been abandoned and were randomized to either foster care or child institutions early in life. Homotypic correlations were low-to-moderate for *p* ($r_{8\text{-}12} = .29$, $p < .01$; $r_{12\text{-}16} = .54$, $p < .001$), low for specific internalizing ($r_{8\text{-}12} = .12$, $p > .05$; $r_{12\text{-}16} = .36$, $p < .001$), and low for specific externalizing ($r_{8\text{-}12} = .06$, $p > .05$; $r_{12\text{-}16} = .37$, $p < .001$). Wade et al. demonstrated metric invariance (i.e. equivalent factor loadings over age), so the reasons for their low stability coefficients may not be the same as Murray et al. (2016). The relatively large gap between measurement occasions and the small sample size might explain why homotypic associations were modest between ages 12 and 16 compared to other studies, but this does not explain the weak coefficients observed between ages 8 and 12. It may be that the introduction of parent-reported symptoms at age 12 compared to teacher-reported symptoms alone at age 8 may have reduced the reliability and comparability of the measures between these time-points.

Most of the heterotypic correlations reported by Wade et al. (2018) were weak, with the notable exception of *p* at age 8 and externalizing at age 12 ($r = .37$, $p < .001$), and externalizing at age 12 and *p* at age 16 ($r = .33$, $p < .001$). Others have also reported weak but significant heterotypic interactions between *p* and externalizing that outweigh the associations observed between *p* and internalizing (McElroy et al., 2017; Olino et al., 2018). This may simply be caused by the fact that specific externalizing tends to show reliable variance beyond the *p* factor (Lahey et al., 2017), or it may be that the two factors are co-dependent, at least during childhood and adolescence.

In sum, longitudinal studies of the bifactor dimensions in adolescence show high homotypic (i.e. within-factor) differential stability, particularly for the *p* factor, while heterotypic continuity between general and specific factors is low. Nonetheless, heterotypic continuity in the *p* factor may be expressed in different ways, such as changes in the way that problems relate to *p* over time.

2.3.4 Differential Stability in Adulthood

Greene and Eaton (2017) examined the differential stability of the *p* factor, fear factor, distress factor, and externalizing factor (comprised of substance dependence diagnoses) in a community sample of 34,653 adults aged 18-90+ between two time-points separated 3-4 years apart. The *p* factor showed a moderately large autoregressive coefficient ($B = .65$, p < .001) which was smaller compared to the specific factors ($B_{\text{externalizing}} = .93$, p < .001; $B_{\text{fear}} = .77$, $p < .001$), all of which were large in absolute terms except for the distress factor ($B = .32$, $p > .05$).

Furthermore, the *p* factor weakly but significantly predicted the specific factors at the following time-point (B_{distress} = .17, $p > .05$; B_{fear} = .15, $p > .05$; $B_{\text{externalizing}}$ = -.06, $p > .05$), but was not significantly predicted by specific distress (B = .08, $p > .05$), externalizing[4] (B = -.10, $p > .05$), or internalizing (B = .05, $p > .05$) factors. These findings demonstrate that the strong homotypic stability in the general and specific factors is maintained in adulthood, as well as the weak heterotypic stability. Furthermore, the *p* factor appears to influence future variation in specific factors, albeit weakly, which is consistent with Caspi et al.'s (2014) hypothesis that a higher risk of comorbidity predicts varied expressions of specific problems.

While Greene and Eaton (2017) demonstrated high levels of differential stability within each factor, this was not invariant across age groups. For instance, the *p* factor's autoregressive coefficient was similar for younger adults (18-32; B = .71) and middle-aged adults (33-45; B = .74) but dropped for older adults (46+; B = .62). Furthermore, the distress factor showed low stability in younger adults (B = .33) and middle adults (B = .32), but increased stability for older adults (B = .67). Fear was the only factor to show invariant autoregressive estimates across the age range (younger = .79, middle = .83, older = .79). These results suggest that general psychopathology may become less predictive of individual differences in severity in older adulthood, but more longitudinal studies are needed during this age age. It should be noted that chi-square difference testing has high type 1 error rates, so small changes in coefficients will appear significant (e.g., externalizing was also found to be invariant, but the change in autoregressive coefficients and model fit was negligible, e.g., younger = .92, middle = .92, older = .87).

[4] The near-perfect autoregressive coefficient for externalizing problems was likely a function of its narrow set of indicators.

Kim and Eaton (2017) extended these findings using a person-centred latent class analysis to determine the probability of respondents remaining within the same latent class between the two time-points. The classes mirrored the factors identified by Greene and Eaton (2017) but were hierarchically. For instance, a *p* factor class with two latent groups (high/low *p*) was identified at the top of the hierarchy. The high *p* group subdivided into people with comorbidities but primarily internalizing disorders, and people with comorbidities but primarily externalizing disorders. These sub-classes also subdivided into more narrow internalizing and externalizing subgroups until each part of the severity continuum was represented by a different class at the lowest level, e.g., a highly comorbid internalizing class, highly comorbid externalizing class, distress class, fear class, externalizing class, and low *p* class. The hierarchy is thus a decomposition of *p* into finer-level latent classes.

The person-centred approach (e.g., changes in class membership) provides clearer insight into what happens to people that show low stability compared to variable-centred approaches (e.g., cross-lagged coefficients). For example, 62% of respondents were expected to remain within the high *p* class between time-points one and two. Hence, 38% were expected to move to the low *p* group. Transition probabilities at the level below the *p* factor class showed that a substantial proportion (46%) of the comorbid-internalizing class transitioned to the low *p* group, compared to a small proportion (2%) in the comorbid-externalizing class. Therefore, people who showed low stability were more likely to display stronger internalizing tendencies among their comorbidities than externalizing tendencies, which mirrors other findings (e.g., McElroy, Shevlin, & Murphy, 2017). Those with a stronger internalizing presentation may be more susceptible to social influence and hence therapeutic intervention than those with a stronger externalizing presentation.

It should be stressed that Kim and Eaton's (2017) latent classes are not independent; the variance in the high *p* class was continuously subdivided into smaller classes. Therefore, it may be that internalizing disorders are more closely related to, or even a subdivision of, *p* than externalizing disorders (Kim & Eaton, 2015), which is why they explained most of the transitions between the high to low *p* groups.

To summarize, the few adult studies that have investigated differential stability in the bifactor dimensions show that homotypic stability is high for both general and specific factors, while heterotypic differential stability is low. However, there may be less stability in older adulthood, and the specific direction of heterotypic continuity–although relatively small–may be developmentally informative.

Across age groups, the *p* factor, and to a lesser extent the specific factors, showed moderate-to-strong autoregressive coefficients that increased slightly from childhood to adolescence and remained consistently high in adulthood. In line with Caspi et al.'s (2014) hypothesis, the *p* factor showed the properties of a stable vulnerability factor for psychopathology. However, the findings do not support the Caspi et al.'s (2014) heterotypic hypothesis, i.e. that people with higher *p* scores will move in and out of various problems, at least when measured with heterotypic associations. It may be that heterotypic continuity is expressed differently, particularly in adolescence, where there may be differences in the way that problems load onto the *p* factor over time, while the rank-ordering among adolescents remains consistent. It may also be that the unreliability of specific factors limited the amount of reliable variance accountable by the *p* factor and vice versa.

Chapter 3 Evaluating Bifactor Studies of Psychopathology Using Model-Based Reliability Indices

In Chapter 2, I reviewed evidence suggesting that the *p* factor represents a continuum of severity in the extent that cognitive, emotional, and behavioural states are dysregulated, and is relatively stable over time. These findings point towards the substantive processes represented by the *p* factor, but they say little about its methodological properties. In other words, they tell us what the *p* factor could be, but not how well it is measured. In fact, bifactor models are typically used to determine the dimensionality in assessment measures or to test the reliability of total and subscale scores (Reise, 2012).

In this chapter, I evaluate bifactor studies of psychopathology for their measurement properties, including (i) the strength of the *p* factor compared to the specific factors in summarizing individual differences in psychopathology measures, and hence their dimensionality; (ii) the extent to which raw total scores and subscale scores accurately represent the *p* factor and specific psychopathology factors, respectively; and (iii) the reliability of *p* factor scores and specific psychopathology factor scores. I will start by describing the practical role of bifactor models in evaluating the psychometric properties of measures. I will then detail the model-based reliability indices used together with bifactor models to evaluate psychometric measures, and aggregate estimates across bifactor studies of psychopathology published to date to determine how well the general and specific psychopathology factors are measured in practice. I end with a review of the methodological issues associated with the bifactor model.

3.1 Introduction

When creating a psychological assessment measure, there is a trade-off between its unidimensionality or reliability (e.g., how consistently the items measure the construct of interest) and its multidimensionality or validity (e.g., how widely sampled the construct is; Bollen, 1989). Researchers often neglect this trade-off when investigating the latent structure of a measure, and either push for a unidimensional (single factor) model or multidimensional (correlated factors) model in an endless struggle to uncover the 'true' population model.

By estimating general and specific factors with a bifactor model, one can determine the relative contribution of domain-general and domain-specific sources of variance to item responses (Reise, Moore, & Haviland, 2011). Some measures will be 'essentially unidimensional', where most of the test variance is attributed to a single factor. Others will be multidimensional with prominent specific factors that are weakly correlated. Most will be multidimensional with a strong general factor that summarizes the overall construct of interest, as well as weaker but meaningful specific factors that summarize particular domains (Reise, 2012).

The dimensionality of a measure reflects the constructs it represents and hence its underlying latent structure (Brown, 2014), but what is assessed might not accurately reflect what is represented. For example, a measure of depression might have a multidimensional latent structure, whereby symptoms associated with cognitive, affective, and physical domains are represented in addition to the overall severity of respondents' depression that cuts across domains. Raw scores on a 'total depression scale' will likely reflect respondents' overall depression severity. By this logic, raw scores on cognitive, affective, and physical subscales should reflect respondents’ problems in specific domains, but they are just as likely, if not more likely, to be influenced by their overall depression severity (Rodriguez, Reise, & Haviland, 2016b). Put simply, a measure can be multidimensional in theory, but in practice its assessment is driven by the overall construct measured that drowns out the specific subconstructs. Recently, Rodriguez, Reise, and Haviland (2016a) showed that bifactor analyses of various psychological measures showed a multidimensional latent structure, but the assessment of total and subscale scores was by and large influenced by the general constructs represented.

Rodriguez et al.’s (2016a) findings suggest that the common practice of creating subscales is misguided because they provide little unique information beyond total scores–even if they reflect subconstructs that map onto specific latent factors. These findings also have implications for latent variable modelling. For example, in the bifactor model, specific factors are residualized for the shared variance; therefore, a specific factor that contributes little beyond the general factor will either show low reliability or might not be identified (Chen, West, & Sousa, 2006). This is often seen with the specific fluid intelligence factor in intelligence research (e.g., Gustafsson & Balke, 1993) and specific psychosis factor in psychopathology research (e.g., Caspi et al., 2014), which both ‘disappear’ when estimated within a bifactor model. In the higher-order model, however, first-order factors that contribute little beyond the shared variance are still identified because their variance is a mixture of both the first-order and higher-order factors (Gignac, 2008). Therefore, researchers who favour higher-order models over bifactor models for their simplicity of interpretation (e.g., Sellbom & Tellegen, 2019) might be purporting first-order factors that are essentially ‘hanging by a thread’.

Rodriguez et al. (2016b) outlined several reliability indices that summarize the structural properties of bifactor models, such as the relative strength of the general to specific factor variances, or the extent that specific factors reliably explain the variance in raw subscale scores beyond the general factor. Given the enthusiasm around a latent bifactor structure of psychopathology (see Chapter 2), it would be important, if not imperative, to determine how well this is measured in practice using these reliability indices. Currently, studies evaluate bifactor models using model fit indices, but these say little about the way in which the variance in multi-domain psychopathology measures is distributed and whether this conforms to a latent bifactor structure of psychopathology. Nor do fit indices tell us how precisely these latent factors can be assessed with raw scores. The current chapter meets the recent calls to evaluate bifactor studies of psychopathology with reliability indices (Greene et al., 2019; Sellbom & Tellegen, 2019; Watts, Poore, & Waldman, 2019) to address the following research questions.

Question 1: Is the latent structure of psychopathology measures multidimensional? A bifactor structure of psychopathology includes two components, a general psychopathology factor and specific psychopathology factors (e.g. internalizing, externalizing, thought disorder; see Chapter 2). The former reflects people's overall severity and perhaps liability to any and all forms of psychopathology; the latter reflects specific and gendered styles of coping (Caspi & Moffitt, 2018; Caspi et al., 2014). Both components are important for describing individual differences in psychopathology. Therefore, the latent structure of psychopathology measures should be multidimensional, with roughly half of the variance attributable to a general *p* factor, and the other half attributable to specific psychopathology factors.

Testing the latent structure of a domain is dependent on the measures used to estimate it (Marsh & Hau, 2007). Therefore, it is important to determine the extent to which an estimated latent structure is influenced by methodological heterogeneity. We can achieve this by predicting between-study variability in the amount of variance explained by the *p* factor or specific factors with methodological variables, such as the assessment method, sample size, age, and population, informant, number of indicators, indicator type, and percent uncontaminated correlations (defined below).

Question 2: Do total and subscale scores reliably reflect variation in the general and specific psychopathology factors, respectively? As described above, the assessment of a multidimensional measure could be driven by a general underlying construct. If general psychopathology is a measurable dimension akin to other constructs such as 'depression' or 'neuroticism', then we would expect the raw total scores from multi-domain psychopathology measures to be influenced by a general source (e.g., Rodriguez et al., 2016a). The extent to which specific problems domains can be precisely assessed with raw subscales among the presence of a general psychopathology dimension remains to be answered. Psychopathology subscales might be subject to the same fate as subconstructs in measures of single-domain constructs, their measurement drowned out by a general psychopathology dimension (e.g., Rodriguez et al., 2016a).

Question 3: Are certain measures best suited to estimating factor scores? The first two research questions concern the reliability of factors in terms of the variance they explain in latent (question 1) or observed (question 2) scores. However, factors are also (random) variables, i.e. they vary across subjects just like scale scores. Therefore, there are indices that reflect the reliability of the factor scores themselves, both when estimated as observed variables or latent variables. Using these indices, we can determine whether certain psychopathology measures produce more reliable factors than others. We can also ask if there are general properties of psychopathology measures that predict more reliable factor scores, such as the number and type of items they include.

3.2 Method

3.2.1 Search Strategy

The current analysis is an extension of the systematic review described in Chapter 2. Therefore, the search terms and inclusion/exclusion criteria for studies are the same as those reported in section 2.1.1. However, the current chapter is based on an analysis of factor loading matrices rather than individual studies. Hence, a total of 75 factor loading matrices reported in 49 studies published between 2009 and 2019 were included in the current analysis. Study characteristics are summarized in Table 3.1. A full list of studies can be found in Appendix A.

Table 3.1

Methodological Characteristics for the Reviewed Study Entries (K = 75)

Study Characteristic	*M* or *n*	SD or %
Age (years; 2-40)	16	10
Childhood (2-12)	8	3
Adolescence (13-17)	15	1
Adulthood (18-40)	30	8
N (201-43,093)	3059	6309
Sample Type		
Community	51	68%
Clinical	18	24%
Population	6	8%
Respondent Type		
Self	34	45%
Caregiver	27	36%

Teacher	8	11%
Multiple	6	8%
Indicator Type		
Item-level	44	59%
Subscale-level	31	41%
Measure Type		
Questionnaire	60	59%
Interview	15	41%

3.2.2 Statistical Analysis

Standardized factor loading matrices were extracted from each study and analysed with Dueber's (2017) bifactor indices calculator, a freely available Excel-based tool for calculating the model-based reliability indices described by Rodriguez et al. (2016a, 2016b). Details of each index are provided below. Statistical indices were recalculated for studies that had already reported them, as there was some variation in their accuracy–Dueber's calculator has been rigorously checked by myself and others (D. M. Dueber, personal communication, March, 2019). Therefore, there may be some differences in the coefficients reported here and those reported in the original papers, if not for the fact that we reproduced the correlation matrix rather than used the original data.

3.2.3 Reliability Coefficients

The statistical properties of each reliability index will now be described, in addition to how they apply to the research questions presented in section 3.1.

Explained Common Variance (ECV). ECV reflects the proportion of common variance (i.e. the total variance in the indicators attributed to all factors modelled, also known as the communality) explained by a given factor (Ten Berge & Sočan, 2004). ECV can be likened to the coefficient of determination in regression analysis (e.g., R^2), which reflects the proportion of predictable variance in an outcome explained by a given predictor variable or model (Nagelkerke, 1991). Similarly, ECV reflects the amount of predictable (or modellable) variance in the indicators explained by a given factor. Both R^2 and ECV range from 0-1, with higher values reflecting a greater proportion of variance accounted for by a given predictor or factor, respectively (Reise, Scheines, Widaman, & Haviland, 2013).

Like R^2, ECV is used to index a factor's success or 'strength' in describing individual differences in a measure compared to other factors (Reise, Moore, & Haviland, 2010). Therefore, it can be used to determine whether the *p* factor and specific psychopathology factors explain an equal amount of variance in psychopathology measures to address research question 1 (e.g., "Is the latent structure of psychopathology measures multidimensional?"). ECV values equal to .50 for the *p* factor would indicate that it explains 50% of the variation in symptom scores as represented by both general and specific factors modelled, and hence that the measure's underlying structure is multidimensional. Higher ECV values (e.g., ⩾ .70) for the *p* factor would suggest that latent structure is 'essentially unidimensional', i.e. that individual differences in psychopathology are mainly explained by a single dimension, despite modelling multiple factors (Rodriguez, Reise, & Haviland, 2016b).

ECV can be calculated from a standardized factor loading matrix by dividing the variance explained by the general factor (i.e. sum of squared standardized general factor loadings) by the total common variance (i.e. the sum of squared general and specific factor loadings; Rodriguez, Reise, & Haviland, 2016b):

$$ECV = \frac{(\sum \lambda_G^2)}{(\sum \lambda_G^2) + (\sum \lambda_{S_1}^2) + (\sum \lambda_{S_2}^2) + (\sum \lambda_{S_3}^2) \ldots (\sum \lambda_{S_m}^2)}.$$

ECV can also be computed for specific factors:

$$ECV_s = \frac{(\sum \lambda_{S_1}^2)}{(\sum \lambda_G^2) + (\sum \lambda_{S_1}^2) + (\sum \lambda_{S_2}^2) + (\sum \lambda_{S_3}^2) \ldots (\sum \lambda_{S_m}^2)},$$

which reflects the proportion of common variance explained a specific factor relative to the variance explained by all factors (Stucky & Edelen, 2015). The higher ECV$_S$ scores are, the more a specific factor explains individual differences in psychopathology relative to the other factors, including the *p* factor.

Coefficient Omega (ω). Omega is a model-based index of a measure's internal consistency, i.e. the extent to which item scores are inter-related (Cortina, 1993). The most common index of internal consistency is Cronbach's alpha (α; Cronbach, 1951). Before describing these reliability indices further, it is important to distinguish between internal consistency and unidimensionality. Strongly correlated or internally consistent items are typically assumed to measure a common or unidimensional construct (Streiner, 2003). However, multiple factors could underpin the inter-relatedness between items, such as in the case of Thurstone's common factors model with positively correlated factors (and hence, positively correlated items between factors). Therefore, internal consistency (or item inter-relatedness) overlaps with, but is distinct from, unidimensionality (or item homogeneity), which is better assessed with ECV.

Omega and alpha coefficients both describe the degree of inter-relatedness between items; while alpha is calculated from the observed variance-covariance matrix, omega is calculated from the model-based factor loading matrix (Rodriguez, Reise, & Haviland, 2016b). In addition to relaxing the assumption of tau equivalence (e.g., equal covariances between items), omega includes multiple sources of variance in explaining the inter-relatedness between items (e.g., general and specific factors), while alpha includes a single source (Reise, Bonifay, & Haviland, 2013). Nonetheless, this does not mean that omega is an index of dimensionality because it explains the degree of inter-relatedness between items not their underlying factor structure (Reise et al., 2010).

To avoid confusion, it is best to think of omega as the proportion of variance explained by all factors in unweighted composite scores (e.g., total or subscale sum scores, also known as unit-weighted composites because each item is treated equally when summed), while ECV is the proportion of variance explained by a given factor in optimally-weighted composite scores (e.g., total or subscale scores weighted by each item's factor loadings; Rodriguez et al., 2016b). The relevance of omega to addressing the second research question will be clear after introducing omega hierarchical below.

Omega can be calculated from a standardized factor loading matrix as follows:

$$\omega = \frac{(\sum \lambda_G)^2 + (\sum \lambda_{S_1})^2 + (\sum \lambda_{S_2})^2 + (\sum \lambda_{S_3})^2 \dots (\sum \lambda_{S_m})^2}{(\sum \lambda_G)^2 + (\sum \lambda_{S_1})^2 + (\sum \lambda_{S_2})^2 + (\sum \lambda_{S_3})^2 \dots (\sum \lambda_{S_m})^2 + (\sum 1 - h^2)},$$

where the squared sum of factor loadings for the general and specific factors is divided by the squared sum of factor loadings for the general and specific factors plus the sum of unique item variances. In other words, omega reflects the proportion of reliable variance in raw total scores attributable to all factors modelled (McDonald, 1999). Omega ranges from 0-1, with higher values indicating that the inter-relatedness among items–which allows them to be summed together into total scores–is mainly attributable to the sources of variance modelled rather than error.

***Omega hierarchical* (*ωH*).** Suppose we wanted to address the first part of research question 2 (e.g., "Do total scores reliably reflect variation in the general psychopathology factor, respectively?"). Coefficient omega provides an estimate of the overall variance in raw total scores explained by all factors modelled. To determine the proportion of variance in raw total scores explained by a single source, such as the p factor, we could calculate omega hierarchical as follows:

$$\omega_H = \frac{(\sum \lambda_G)^2}{(\sum \lambda_G)^2 + (\sum \lambda_{S_1})^2 + (\sum \lambda_{S_2})^2 + (\sum \lambda_{S_3})^2 \dots (\sum \lambda_{S_m})^2 + (\sum 1 - h^2)},$$

where the squared sum of factor loadings for the general factor is divided by the squared sum of factor loadings for the general and specific factors plus the unique variance. In other words, we have removed the variance in total scores explained by specific factors in the numerator of coefficient omega, leaving behind the general factor to explain such variance. Omega hierarchical values range from 0-1, with higher scores indicating that the inter-relatedness or 'summable-ness' of raw item scores is mainly explained by a general factor. While there are currently no accepted cut-offs, omega hierarchical values ⩾. 80 are thought to reflect a high degree of inter-relatedness explained by a factor (Rodriguez, Reise, & Haviland, 2016a). Therefore, an omega hierarchical value of .90 for the p factor would indicate that it explains the majority (e.g., 90%) of the variance in raw total scores. Bear in mind, however, that the latent structure of item responses could still be multidimensional (e.g., equally explained by general and specific factors, which would be indexed by ECV values ~ .50).

We can also determine the proportion of error-free variance in raw total scores attributable to the p factor by dividing ω_H by ω, which is known as relative omega (Dueber, 2017).

These principles can also be used to address the second part of research question 2 (e.g., "Do subscale scores reliably reflect variation in the specific psychopathology factors, respectively?"). For example, we can first calculate the proportion of variance in raw subscale scores attributable to general and specific factors with omega-subscale:

$$\omega_S = \frac{(\sum \lambda_G)^2 + (\sum \lambda_{S_1})^2}{(\sum \lambda_G)^2 + (\sum \lambda_{S_1})^2 + (\sum 1 - h^2)}.$$

We can then determine the proportion of variance in raw subscale scores attributable to a given specific factor while controlling for the general factor using omega hierarchical-subscale (ω_{HS}):

$$\omega_{HS} = \frac{(\sum \lambda_{S_1})^2}{(\sum \lambda_G)^2 + (\sum \lambda_{S_1})^2 + (\sum 1 - h^2)}.$$

Finally, by dividing ω_{HS} by ω_S, we can determine the relative omega-specific, i.e. the proportion of error-free variance in raw subscale scores attributable to a specific factor.

Factor Determinacy (FD) and Construct Reliability (H). The reliability indices presented so far reflect the amount of variability in raw scores (omega hierarchical) or weighted scores (explained common variance) attributed to the general or specific factors. Research question 3 concerns the reliability of the factor scores themselves (e.g., "Are certain measures best suited to estimating factor scores?"). The two main reliability indices for factor scores are factor determinacy (FD) and construct reliability (H). Like how omega and ECV differ based on whether they reflect the explained variability in raw or weighted scores, respectively, FD and H differ in reflecting the variability in observed[5] or latent factor scores, respectively. I now describe FD in more depth followed by the H index.

Where factors cannot be used as latent predictors or outcome variables (e.g., see Chapter 5), researchers can use factor scores, which are observed estimates of individual differences on a latent variable. Factor scores are inherently 'indeterminate', i.e. for any set of factor scores estimated, there exists a different set of scores that could be derived from the same loading matrix (Gutmann, 1955). FD reflects the reliability of observed factor scores, or the extent to which factor scores are accurate estimates of individual differences on a factor given the various possible estimates (Grice, 2001). FD can be calculated from the model-implied correlation matrix with the following formula:

$$FD = diag(\mathbf{\Phi}\mathbf{\Lambda}^{T}\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Phi})^{1/2},$$

where $\mathbf{\Phi}$ is a $k \times k$ matrix of factor correlations, $\mathbf{\Lambda}$ is a $j \times k$ factor loading matrix, and $\mathbf{\Sigma}$ is a $k \times k$ matrix of the model-implied factor correlations. *FD* values range from 0 to 1 and represent the correlation between a latent factor and its observed factor scores. Gorsuch (1983) suggested that *FD* values ⩾ .90 reflect trustworthy factor score estimates. We could address research question 3 by comparing FD estimates across different bifactor studies that use that same psychopathology measures.

H reflects the 'construct reliability' of latent factor scores, or how well they can be replicated across studies (Hancock & Mueller, 2001). The extent to which a

[5] I have intentionally used the term 'observed' rather than 'raw' because factor scores that are estimated as observed variables are still optimally weighted (i.e. the aggregation of item scores is weighted by their factor loadings) while raw composites tend to be unit-weighted (e.g., each item is weighted equally).

factor is replicable will depend on how well it is defined by its indicators. Therefore, H can be estimated using the formula:

$$H = 1/\left[1 + \frac{1}{\sum_{j=1}^{J} \frac{\lambda_j^2}{1 - \lambda_j^2}}\right],$$

which reflects the proportion of variance in each item explained by the factor relative to the proportion of variance in each item not explained by the factor. H ranges from 0-1, with values ⩾ .70 indicating that a factor is well-defined by its indicators and hence replicable in another study using the same indicators (Hancock & Mueller, 2001). Like FD, we can address research question 3 by comparing H values across different bifactor studies that use that same psychopathology measures.

Percent Uncontaminated Correlations (PUC). PUC describes the number of correlations that can be described by the general factor in the absence of specific factors (Rodriguez et al., 2016b). It is not so much about the variance explainable by the general factor, but rather, how the data structure lends itself to measurement that is 'uncontaminated' by the multidimensionality introduced by specific factors. PUC is not directly relevant to any of the research questions outlined in section 3.1, but it influences estimates of ECV and omega hierarchical (see below) and should therefore be considered whenever they are estimated.

PUC is estimated by first calculating the number of unique correlations among indicators, e.g., $p(p-1)/2$, where p is the number of indicators. For example, if there are 12 items, then there will be 66 unique correlations ([12*11]/2). In the single factor model, all 66 correlations would be explained by the general factor. However, in the bifactor model, each item is influenced by both general and specific factors. To determine the number of correlations explained by the general factor alone, one can calculate the PUC:

$$PUC = 1 - \frac{\sum_{m=1}^{K}(p_m(p_m - 1)/2)}{p(p-1)/2},$$

where the numerator reflects the number of unique correlations summed over the specific factors, m, where m = 1, …, K, and the denominator reflects the total number of unique correlations. Continuing our example, if there were three specific factors with four loadings each, then 18 of the 66 unique correlations would be explained by the specific factors ([[4*3]/2]*3). Therefore, 27% of the unique correlations are 'contaminated' by multidimensionality, leaving 48 of the unique correlations (73%) to be explained by the general factor alone.

PUC is largest when there are numerous specific factors each with a small number of loadings. This would increase the number of between-factor item correlations that can be explained by the general factor compared to the number of within-factor items correlations that can be explained by both the general and specific factors. PUC values ⩾. 70 indicate that the correlation matrix is 'essentially unidimensional' (Rodriguez et al., 2016a). In other words, more than 70% of the possible correlations between items can be explained by a single source based on structural properties alone, but this does not indicate how well the correlations will be explained (e.g., the data might lend itself to a strong general factor but its loadings might be weak).

Provided the indicators load well onto the general factor, it will likely explain a large amount of the variance in raw total scores (omega hierarchical) and optimally weighted total scores (ECV) when PUC is high. However, the ECV will start to drop after a certain number of specific factors are modelled, since the amount of common variance attributable to the general factor will be outweighed by specific factors. The PUC also moderates the impact of ECV on parameter bias when fitting a unidimensional model to multidimensional data; fitting a single factor to item responses that are best explained by both general and specific factors will skew parameter estimates to achieve an optimal fit. In instances where the ECV is low-to-moderate (i.e. the common variance is explained by both general and specific factors and is thus multidimensional), parameter bias will be less pronounced if the PUC is high compared to when the PUC is low (Reise, Scheines, Widaman, & Haviland, 2013). Therefore, PUC has important properties that determine the extent that a general factor can predict the covariation among items, as well as protecting against mis-specification that favours a unidimensional model.

3.3 Results

Explained Common Variance. Table 3.2 presents reliability indices summarized across study entries. The mean ECV was 0.57 (SD = 0.14); on average, the *p* factor accounted for 57% of the common variance in psychopathology measures, but there is a sizeable amount of residual variance accounted for by specific factors (e.g., 43%). Therefore, psychopathology measures tend to be multidimensional, with a moderately strong *p* factor and individually weak but collectively meaningful specific factors. The ECV$_s$ for the average specific factor was low but variable (ECVs = .12, SD = .9).

Table 3.2

Reliability Indices for the Reviewed Study Entries (K = 75)

	General (*p*) Factor			Specific Factor		
Reliability Index	*M*	*SD*	Range	*M*	*SD*	Range
No. of items	37	30	8-139	10	7	2-36
Item-level	54	28	12-139	12	8	3-36
Subscale-level	13	8	8-51	5	3	2-17
ECV/ECV_s	0.57	0.14	0.3-0.84	0.12	0.09	0.01-0.35
ω/ω_s	0.93	0.06	0.75-0.99	0.86	0.11	0.26-0.98
ω_H/ω_{Hs}	0.74	0.13	0.45-0.97	0.37	0.22	0-0.91
Relative Omega	0.79	0.11	0.54-0.98	0.43	0.25	0.01-1
H	0.91	0.07	0.74-0.99	0.69	0.17	0.24-0.98
FD	0.93	0.16	0-1	0.85	0.24	0-0.99
PUC	0.67	0.13	0.38-0.92	N/A	N/A	N/A

Note. ECV/ECV_s = explained common variance/explained common variance subscale; FD = factor determinacy; H = construct reliability; ω/ω_s = omega/omega subscale; ω_H/ω_{Hs} = omega hierarchical/omega hierarchical subscale; PUC = percent uncontaminated correlations.

Table 3.3 shows the regression coefficients for the method variables when added separately and simultaneously to regression models predicting ECV values (i.e. variability in *p* factor strength). In the separate regression models, increases in age predicted an initial dip in ECV values followed by an increase. Furthermore, higher PUC values and more items predicted higher ECV values, while teacher- vs. self-reported outcomes predicted lower ECV values. In the simultaneous regression model, only the association between teacher-reported outcomes and lower ECV values remained significant, but also parent- vs. self-reported outcomes predicted significantly higher ECV values.

Collectively, the method variables explained 62% of the variance in ECV values (R^2 = .62; adjusted R^2 = .54), much of which was attributable to the informant variable (B_{parent} = .42; B_{teacher} = -.47). It is, however, likely that amount of variance explained by the teacher- vs. self-reported contrast was driven by the unusually low ECV values for Murray et al. (2016). Nonetheless, the parent- vs. self-reported contrast was still significant, indicating that the informant is an important source of the variability in *p* factor strength.

In Chapter 2, I hypothesised that the U-shaped relationship between age and ECV values could be an effect of differentiation (e.g., there is a gradual shift in the way that psychopathology is expressed between childhood and adolescence which then strengthens into adulthood) or may simply reflect changes in the way that we assess psychopathology across age. We can test this hypothesis by examining the interaction between age and PUC in predicting ECV values, since the PUC is a proxy for the design of psychopathology measures (e.g., higher PUC values reflect more numerous subscales but with a smaller number of items in each).

Table 3.3

Regression Coefficients for the Method Variables When Added Separately (Left) or Simultaneously (Right) to Models Predicting ECV Variability (K = 75)

	Separate Models					Simultaneous Model				
Predictor	*b*	*B*	*t*	*p*	95% CI	*b*	*B*	*t*	*p*	95% CI
N	0	-.07	-0.59	.554	0, 0	0	-.05	-0.49	.628	0, 0
Age	.002	.14	1.16	.248	-.001, .005	.002	.14	0.98	.332	-.002, .006
PUC	**.27**	**.25**	**2.21**	**.030**	**.03, .52**	.18	.16	1.29	.201	-.01, .46
No. of items	**.001**	**.31**	**2.78**	**.007**	**.0004, .003**	.001	.24	1.66	.101	-.0002, .003
Respondent (*v.* self)										
Caregiver	.05	.17	1.68	.098	-.009, .104	**.12**	**.42**	**3.10**	**.003**	**.04, .19**
Teacher	**-.23**	**-.54**	**-5.34**	**< .001**	**-.32, -.14**	**-.20**	**-.47**	**-4.48**	**< .001**	**-.29, -.11**
Multiple	-.01	-.03	-0.31	.761	-.11, .08	.08	.18	1.73	.089	-.01, .18
Item-level analysis (*v.* subscale-level)	-.03	-.09	-0.79	.432	-.09, .04	-.004	-.02	-0.10	.918	-.09, .08
Questionnaire (*v.* interview)	-.05	-.16	-1.39	.170	-.13, .02	-.05	-.14	-1.35	.182	-.12, .02
Sample (*v.* community)										
Population	.04	.13	1.09	.278	-.03, .12	-.007	-.02	-0.21	.836	-.07, .06
Clinical	.09	.20	1.70	.093	-.02, .21	.09	.17	1.58	.120	-.02, .20

Note. b = unstandardized regression coefficient; *B* = standardized regression coefficient; PUC = percentage uncontaminated correlations. Significant results are in bold.

In a model with age, PUC, and an age*PUC interaction, age continued to negatively predict ECV values (*b* = -.02, *B* = -1.67[6], *t* =2.00, *p* = .05, 95% CI [-.05, -.00]), while PUC changed its direction of prediction and negatively predicted ECV values, but this was no longer significant (*b* = -.33, *B* = -0.30, *t* = 1.13, *p* = .262, 95% CI [-.92, .25]). The age*PUC interaction was significant (*b* = .03, *B* = 2.02[8], *t* = 2.09, *p* = .04, 95% CI [.001, .06]). Therefore, the negative prediction of age on ECV values changed with increasing PUC levels. At lower PUC levels, ECV decreased over age groups, whereas at higher PUC levels, ECV increased over age groups (see Figure 3.1).

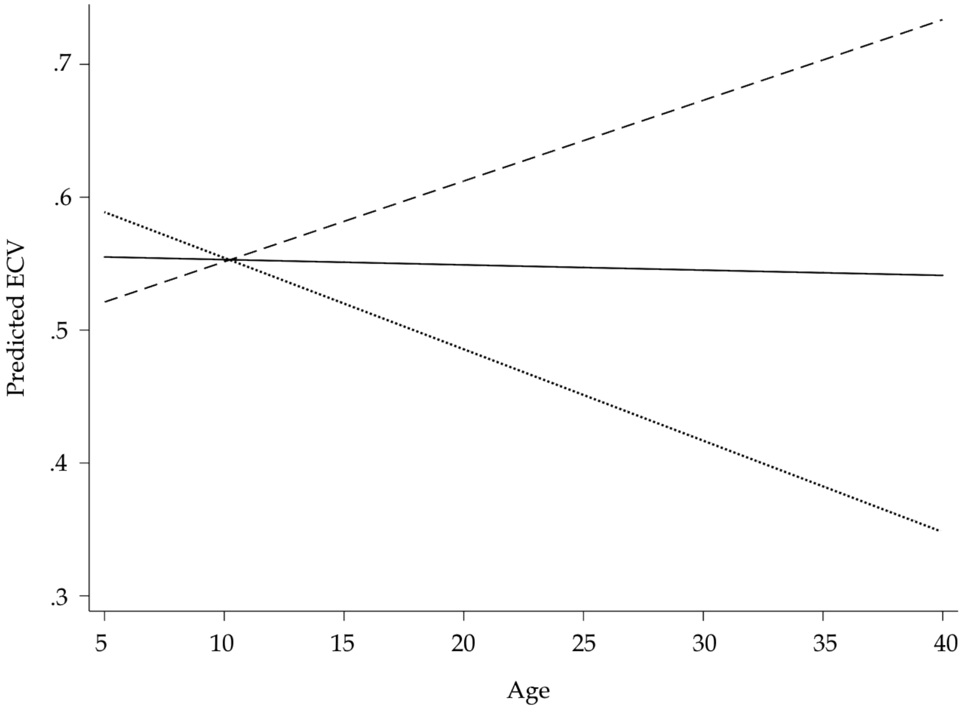

Figure 3.1. Predicted ECV values across age groups at PUC levels of .5 (dotted), .7 (solid), and .9 (dashed).

[6]The standardized regression coefficients fall outside of the expected bounds of -1, 1, suggesting that they are inflated by multicollinearity (Jöreskog, 1999). The correlation between age and PUC was *r* = .51, *p* < .001. Therefore, these estimates should be interpreted with caution.

Omega coefficients. The mean omega was 0.93 (SD = 0.06), hence 93% of the variance in raw total scores was attributable to the general and specific factors. On average, the *p* factor accounted for 74% of the raw total score variance with error (omega hierarchical) and 79% of the variance in raw total scores without error (relative omega). Therefore, most of the variance in raw total scores was explained for by a single source, despite the latent construct of psychopathology being multidimensional (see ECV above).

The mean omega-subscale was .86 (SD = .11) indicating that the majority of variance in raw subscale scores was attributable to the general and specific factors. Nonetheless, only 37% of the raw subscale variance with error (omega hierarchical subscale), and 43% of the reliable subscale variance without error (relative omega subscale), was explained by the relevant specific factor. Therefore, the *p* factor explained more than half of the variance in raw subscale scores (e.g., 57%), drawing doubt over the extent to which subscales reflect the precise measurement of specific problems beyond general psychopathology. It is important to note that the omega-subscale coefficients were somewhat variable; these estimates might not be representative of all studies reviewed.

FD and H. On average, *p* factor scores showed high reliability, regardless of whether they were observed (FD = 0.93, SD = .16) or latent (H = 0.91, *SD* = .07). No single measure was associated with higher FD and H values; instead, measures with more items were associated with higher FD and H values, which would also explain why item vs. subscale-level indicators and questionnaire vs. interviews predicted higher H values (both tend to feature more items; see partial regression coefficients in

Table 3.3). Therefore, using measures with more items (that load well onto *p*) will produce more reliable *p* factor scores.

The reliability of observed factor scores for a given specific factor fell slightly below Gorsuch's (1983) recommended threshold of ⩾ .90 (*FD* = 0.85, *SD* = .24). Furthermore, the reliability of latent factor scores for a given specific factor fell just under an acceptable value (*H* = 0.69, *SD* = .17). Therefore, specific psychopathology factors produce factor scores with near-acceptable reliability.

3.4 Discussion

The bifactor model of psychopathology often attracts researchers and clinicians alike as it emphasises the dimensional and transdiagnostic nature of mental health problems. What is often overlooked, however, is the ability to separate out the unique contributions of the general factor from the specific factors. In turn, we can evaluate the strength of the general factor relative to the specific factors in describing individual differences in psychopathology measures (see below), or examine the unique association between specific psychopathology factors and other variables, free from the confounding influence of the general factor (see Chapters 4 and 5). In this Chapter, I summarized the statistical indices for assessing the reliability of the general and specific factors and applied them to bifactor studies of psychopathology published to date. I set out to answer three main research questions which I will address in turn.

3.4.1 Is the latent structure of psychopathology measures multidimensional?

The *p* factor explained more than half (57%) of the common variance in psychopathology measures compared to specific psychopathology factors, but not enough to suggest that their underlying latent structure is 'essentially unidimensional'. Therefore, both general and specific factors are necessary to represent the latent structure of psychopathology. There was some variability in the strength of the *p* factor between the studies reviewed, 62% of which could be explained by their methodological features. This is a humbling finding: it demonstrates that the strength of the *p* factor is, in part, determined by how it is measured. We must not let theoretical wonder obscure the fact that the *p* factor is first and foremost a statistical construct, one that is subject to the peculiarities of a particular sample, measure, or assessment occasion (Coan, 1964).

Two interesting findings emerged from the analysis of ECV values. First, the strongest predictor of ECV variability was the informant: parent- and teacher- reported measures were associated with higher and lower ECV values, respectively, relative to self-reported measures. This raises a contentious issue in psychological assessment: who knows best? (Achenbach, 2006). Most researchers would agree that different informants offer a unique and complementary perspective on mental health, since these constructs are context-dependent (Achenbach, McConaughy, & Howell, 1987). The *p* factor is not exempt from this issue; the current findings suggest that there needs to be a greater emphasis on multi-informant designs in bifactor studies of psychopathology.

The other interesting finding was that the relationship between age and ECV changed with different PUC levels. Higher PUC levels (i.e. measures that are more suited to estimating a general factor) were associated with increasing ECV values across age groups (i.e. stronger *p* factors with age), while lower PUC values (i.e. measures that are less suited to estimating a general factor) were associated with decreasing ECV values across age groups (i.e. weaker *p* factors with age). In other words, the *p* factor explained more variation in psychopathology measures with age, but only if the measures were suited to estimating a *p* factor. If the measures were more suited to representing both general and specific factors, then the *p* factor explained less variation in psychopathology measures with age. By implication, it may not be appropriate to attribute age-related changes in the *p* factor's reliability to developmental changes in general psychopathology (see section 2.3.1). Instead, it may be our ability to design measures that better estimate the *p* factor *or* specific factors that increases with age.

It is, however, interesting that PUC levels influenced ECV values differently in younger and older samples. Before age 10, higher and lower PUC levels were associated with slightly lower and higher ECV values, respectively, whereas after age 10, higher and lower PUC levels were associated with higher and lower ECV values, respectively, albeit more strongly. The reversal in the association between PUC and age in older childhood might just be a product of unstable regression coefficients. However, it might also indicate that there is a qualitative shift in the expression of psychopathology between childhood and adolescence/adulthood. Psychopathology might be less differentiated in childhood; measures with fewer symptom domains might provide a better representation of individual differences in general psychopathology than measures with multiple symptom domains. However, in adolescence and adulthood, psychopathology might be expressed in more defined (i.e. disorder-specific) ways that vary in their comorbidity rates. Therefore, the effect of PUC on age-related changes in the reliability of the *p* factor and specific factors might not be completely artifactual, since the design of our measures implicitly reflect how psychopathology is expressed at different ages.

3.4.2 Do total and subscale scores reliably reflect variation in the general and specific psychopathology factors, respectively?

Most of the variance in raw total scores was attributable to the *p* factor (74%), indicating that the inter-relatedness among all items is mainly explained by a single dimension. This is particularly interesting given that the common variance was multidimensional (see ECV above). Therefore, while the underlying construct of psychopathology is multidimensional, our measurement appears to be largely explained by a single source. This might reflect a difficulty in assessing specific domains of a dimensional construct (Gignac, 2014), or the confounding effects of state (and trait) levels of general psychopathology on symptom reporting, similar to the confounding effect of motivation on test taking ability (Duckworth, Quinn, Lynam, Loeber, & Stouthhamer-Loeber, 2011).

Less than half of the variance in raw subscale scores was attributable to the specific factors; the remaining variance was accounted for by the *p* factor. Therefore, while specific factors may be important to include in a measurement model, the extent to which they can be precisely assessed beyond the *p* factor is questionable. Practically speaking, this finding casts doubt on the use of subscales to assess specific domains of psychopathology, at least when assessed among multiple problems domains (see Rodriguez et al., 2016a for a similar result with psychology measures). Many studies are therefore at risk of interpreting the unique effects of subscales which are in fact a product of the *p* factor (see Chapters 5 and 6).

Of course, the reduced ability of specific factors to explain the inter-relatedness between groups of items does not mean that subscales have no value or that specific problems cannot be reliably measured. Consider Rodriguez et al.'s (2016a) thought experiment: two subscales might show low omega hierarchical values when assessed within a broader measure, but high omega hierarchical values when assessed as stand-alone measures. What matters is the *context* of a measure: the assessment of specific problems will be less precise among multiple problems than when assessed alone. However, this raises a thorny issue: should we assess what is common among mental health problems or focus on specifics? The answer probably depends on our ontological position: realists will insist on assessing both general and specific aspects of psychopathology, as this is what is reflected in the latent structure, whereas constructivists/pragmatists will settle for measures that meet their demands in practice.

3.4.3 Are certain measures best suited to estimating factor scores?

No single measure was associated with more reliable general and specific factor scores. Instead, measures that included more items produced more reliable factor scores, both latent and observed. As with the measurement of any construct, using more items that relate well to the construct will produce more reliable factor scores (Rozeboom, 1982). This is particularly true of the *p* factor given that it is estimated using all items. By contrast, specific factors feature fewer items and hence showed near-acceptable factor score reliability. One could increase the number of items predicted by each specific factor, but this might also lower the PUC, which might not be desired if the goal is to assess a single dimension of psychopathology.

Ultimately, measures must be created or selected with their purpose in mind: If the goal is to evaluate the predictors or outcomes of the *p* factor in structural equation models, then a measure that maximizes its items and includes brief subscales will suffice. However, if the goal is to assess the predictors or outcomes of the specific factors free from the general variance, then a measure that maximises the number of items per specific factor will be important for estimating reliable specific factor scores.

3.4.4 Limitations

Several compromises were made in the study selection process to ensure an adequate coverage of bifactor studies. Consequently, the included studies differed in how they estimated the bifactor dimensions (e.g., CFA vs. Schmid-Leiman transformation) and the characteristics of their solutions (e.g., orthogonal vs. correlated specific factors, inclusion of cross-loadings). Nonetheless, the formulae for calculating reliability indices assume that the general and specific factors are estimated from a CFA with a simple structure (i.e. no cross-loadings or specific factor correlations; Rodriguez et al., 2016b). Therefore, the reliability estimates presented should be treated with caution. Further bias would have also been introduced by the fact that the reliability estimates were not adjusted for differential samples sizes or dependencies in entries that used the same sample at different ages or different members from the same family.

There is also some debate about applying model-based reliability estimates to categorical outcomes. Model-based reliability estimates are designed for continuous outcome variables (Rodriguez et al., 2016b). While they can be applied to categorical outcomes, their reliability will ultimately depend on how skewed the categorical outcome variables are (Flora & Curran, 2004). Many opt for weighted least-squares estimation of their factor models to avoid the distributional problems associated with categorical outcomes. However, calculating reliability estimates from the resultant polychoric correlation matrices requires extreme caution as we are no longer estimating the reliability of raw scores, but their hypothesised continuous distributions (Chalmers, 2017). The current analysis included factor loading matrices estimated using maximum-likelihood (raw correlation matrix) and weighted-least squares (polychoric correlation matrix), and so the estimates should be treated with caution as neither estimation approach is unbiased.

3.5 Conclusions and Further Methodological Issues

Three conclusions can be reached in this chapter. First, the latent structure of psychopathology is multidimensional, represented by a general latent dimension that reflects commonalities among all the problems assessed, as well as specific dimensions that reflect commonalities among subsets of problems. The strength with which the general *p* factor explains individual differences in psychopathology assessments is dependent on method characteristics, particularly the informant.

Second, while the latent structure of psychopathology is multidimensional, its measurement tends to be dominated by a single dimension. Subscales provide some information beyond total scores, but not enough to be considered unique from the single dimension.

Third, the reliability of both general and specific factor scores can be increased by sampling more items, following the laws of reliability theory. In all, the *p* factor (or more appropriately, a *p* factor) is sensitive to its measurement context, which raises further questions about methodological issues surrounding *p* that I will now review.

3.5.1 Is the *p* factor a statistical artifact?

The positive correlations among symptoms and disorders are thought to be underpinned by a single dimension of psychopathology, but they might instead be a product of method effects (Lahey et al., 2012; Lahey et al., 2015; Lahey et al., 2017). For example, items on a questionnaire or interview may positively co-occur due to responses biases, such as agreeing or disagreeing with items regardless of their content. Furthermore, some items encourage certain ways of responding that are unrelated to the construct measured, such as responding in a socially acceptable manner. Assessments are conducted within a certain context (e.g., a given time and place; blocks of items also serve as a context) which might influence the interpretation of items and hence variation in scores (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

It is easy to forget that the bifactor model was initially used to control for method effects. Researchers would estimate a general factor to capture the common variance associated with the use of a single measure or informant, as well as specific factors to capture specific constructs (Billiet & McClendon, 2000; Marsh, 1989; Widaman, 1985). It is therefore likely that some of the variance explained by the *p* factor is attributable to common method effects. If we take Cote & Buckley's (1987) meta-analytic estimate of the proportion of variance in personality measures explained by method effects, then 25% of the variance in *p* may be due to method effects.

While 25% of the variance is not enough to claim that the *p* factor is an artifact, it might still impact the correlations between symptoms from different domains (e.g., social anxiety and conduct problems), which are naturally weaker than the correlations between symptoms from similar domains (e.g., social anxiety and generalized anxiety). In other words, controlling for method effects might weaken or negate some of the *p* factor's loadings, meaning that symptoms do not show a true positive manifold attributable to a single dimension of psychopathology.

Few have tested the amount of variance in the *p* factor explained by common method variance, mainly because they are both estimated with the same method (e.g., a general factor). However, Lahey, Rathouz, Krueger, Waldman, and Zald (2017) controlled for common method variance using a multitrait multimethod (MTMM) matrix. With MTMM, one can estimate the sources of variance associated with traits and methods without specifying latent variables. Specifically, one estimates a correlation matrix for items that share the same content but are assessed with at least two different methods (Campbell & Fiske, 1959).

Lahey et al. (2017) scored various disorders for the frequency and severity of their symptoms using child and parent reports on the Child and Adolescent

Psychopathology Scale. The trait assessed was general psychopathology, and the methods used were two different informants (one could have also used two different assessment measures). The MTMM matrix included a within-method portion (e.g., correlations between disorders reported by a single informant) and between-method portion (e.g., correlations between disorders reported by different informants). If the associations between disorders reported by a given informant are substantive rather than a product of method effects, then they should be present when assessed across informants.

Lahey et al. (2017) found a positive manifold across disorders in the between-informant ratings (e.g., symptoms rated by the child and parent for different disorders were positively correlated). However, the median cross-disorder correlations dropped from *r* = .42 (parent) and *r* = .51 (child) within each informant, to *r* = .23 between informants. Lahey et al. argued that the between-informant correlations were unlikely to be caused by method effects alone, as 91% remained significant. Nonetheless, it is questionable whether the correlations are strong enough to be considered the result of a general trait. Moreover, parent and child reports are not fully independent: heritable biases in responding may have influenced both parents and their children (Alessandri et al., 2010; Melchers et al., 2018).

Lahey et al. (2017) also argued that the pattern of between-informant correlations was not uniform, as would be expected if they were driven by common method effects. For instance, depression correlated more strongly with generalized anxiety than with specific phobia, both within and between informants. Nonetheless, it is possible that traits and method effects interact, which would produce varied correlations patterns (Podsakoff, MacKenzie, & Podsakoff, 2012). For instance, conduct disorder symptom ratings may have shown weak between-informant correlations with other disorders because they possess lower social desirability, making people reluctant to respond in a truthful manner, rather than because conduct problems are distinct from other disorders.

3.5.2 Is the *p* factor a consequence, rather than cause, of mental health problems?

van der Maas et al. (2006) argued that the positive manifold among intelligence tests could be explained by reciprocal interactions between cognitive processes during development rather than a single underlying trait. They presented a model used in biology to describe cooperation and competition dynamics between species or parts of an ecosystem. When describing lake health, for instance, we might expect a positive manifold whereby better lakes are healthier in all aspects of the ecosystem (e.g., better water quality, more diverse flora and fauna, etc.). Rather than a single factor explaining variation in lake health, such variation could be described by cooperative interactions between components of the ecosystem. For example, cleaner water would provide a stable living environment for freshwater lake bacteria, and freshwater lake bacteria would also keep the water clean (Newton, Jones, Eiler, McMahon, & Bertilsson, 2011). No single latent factor is necessary to describe the mutual relationships between parts of the ecosystem that determine lake health.

van der Maas et al. (2006) defined a model of cognitive development with two parts: a logistic growth component that described the development of individual cognitive processes, and an interaction component that described how cognitive processes interacted during development. Through a simulated dataset, they demonstrated that cognitive processes showed a positive manifold when certain growth parameters were correlated, and this was explained by a single factor, as if an underlying dimension underpinned the resources allocated to each cognitive process.

However, van der Maas et al. (2006) also observed a positive manifold when the cognitive processes were correlated, rather than their growth parameters. In other words, the positive co-occurrences between cognitive processes were apparent even when no single underlying dimension produced their interactions. All that was needed to produce the positive manifold was the interaction between cognitive processes during development, which is known as dynamic mutualism.

The concept of dynamic mutualism has inspired some to describe psychiatric comorbidity using network models. Network models include a set of observed variables (‘nodes’) and connections between them (‘edges’; Borsboom & Cramer, 2013). Networks have several features, including centrality (e.g., how connected a node is to all the other nodes), edge weight (how strongly connected a node is to another node), and edge direction (e.g., whether two nodes are positively or negatively connected). Networks can also be estimated in different ways to yield different kinds of information. For example, a directed network involves unidirectional edges, where the direction of relationship between nodes is estimated (e.g., Node A → Node B), whereas undirected networks do not specify the direction of the relationship (e.g., Node A ←→ Node B).

A growing number of studies have applied network models to patient-reported symptom data. For example, Boschloo et al. (2015) applied an undirected network model to data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), which features symptom-based reports of 12 DSM-IV disorders by 34,653 adults. They found that the connections between symptoms strongly clustered around DSM-IV diagnoses, but each disorder was also weakly connected to at least three other disorders.

The lack of sharp boundaries between disorders mirrors the finding of a general psychopathology factor using this dataset (Greene & Eaton, 2017; Lahey et al., 2012), but the weak between-disorder correlations question the extent that this is underpinned by a single underlying trait–much like Lahey et al.'s (2017) findings after controlling for common method effects. Further questioning a single underlying trait is the finding that disorders were connected through different symptoms (i.e. 'bridge symptoms'; Borsboom, 2017). If disorders are connected by a single underlying trait, then symptoms from different disorders should be equally connected. Instead, Boschloo et al.'s findings suggest that each symptom is a unique causal entity that interacts with other symptoms for different reasons rather than a single underlying cause.

Boschloo, Schoevers, van Borkulo, Borsboom, and Oldehinkel (2016) replicated the strong and numerous connections within clusters of symptoms that mirrored DSM-IV child and adolescent disorders, as well as weak connections between clusters of disorders. Other studies have focused on comorbidity among depression and other disorders, such depression and anxiety (Beard et al., 2016; Cramer, Waldorp, van der Maas, & Borsboom, 2010), depression and OCD (McNally, Mair, Mugno, & Riemann, 2017), and depression and PTSD (Choi, Batchelder, Ehlinger, Safren, & O'Cleirigh, 2017). Bridge symptoms vary with each disorder pairing, suggesting that different mechanisms underpin their overlap. Some have also investigated comorbidity between disorders other than depression, such as OCD and autism (bridged by repetitive behaviours; Ruzzano, Borsboom, & Geurts, 2015) and eating disorders and social anxiety (bridged by nervousness of one's appearance; Levinson et al., 2018).

While network models provide compelling evidence against a substantive *p* factor, they too are susceptible to method effects. For example, connections between symptoms may be driven by similar wording patterns or expectations about how symptoms co-occur (Constantini et al., 2015). Item blocks might also explain why symptoms cluster around DSM disorders, which in itself contradicts the argument that mental disorder is characterized at the symptom-level rather than the disorder-level (Borsboom & Cramer, 2013). Moreover, just because a network model 'fits' the observed data, does not mean that mental health symptoms are structured like a network. There is an inherent divide between theory and method which network modellers, like latent variable modellers, suffer from. It is therefore necessary to directly compare network and latent variable models to make claims about the structure of psychopathology (Eaton, 2015).

Another criticism of network models is that by re-specifying the correlation matrix–rather than reducing it to a smaller set of variables–there is a risk that they produce structures that are not replicable across samples and measures (Krueger, DeYoung, & Markon, 2010). Indeed, Forbes, Wright, Markon, and Krueger (2017) found that the centrality of specific nodes and direction and weight of specific edges lacked replicability across different conditionally independent networks (e.g., networks that partial out the influence of other variables on the relationship in question). Moreover, Steinley, Hoffman, Brusco, and Sher (2017) found that several centrality statistics and edge weights did not differ significantly from what would be expected by chance when analysing binary data. They also highlighted the need to correct for multiple testing and provided a nonparametric alternative to computing 95% confidence intervals via bootstrapping (which are otherwise overestimated using standard methods). In line with Forbes et al.'s findings, Steinley et al. concluded that "psychometric network models should be employed with extreme caution and interpreted guardedly" (p. 1000).

While network models and latent variable models are typically pitted against each other, their explanations of individual variation in psychiatric comorbidity overlap. According to van der Maas et al. (2006), mutualistic interactions are weighted by the resources allocated to cognitive (or emotional) processes, which creates variation in the extent to which processes interact. Resources are not sampled from the same multivariate distribution, i.e. they are uncorrelated, meaning that the comorbidity among psychiatric disorders can arise differently for everyone. For example, someone might allocate resources to various symptoms leading to broad disorder overlap, whereas someone else might allocate resources to certain symptoms which act as a central node that attracts other symptoms.

While it may not be apparent prima facie, the mutualism explanation for psychiatric comorbidity mirrors the bifactor model in its structure. Both feature common and unique components that cause the variation in comorbidity: the mutualism model includes interactions between symptoms (common) that are weighted by uncorrelated resources (unique), while the bifactor model includes a general latent factor (common) as well as uncorrelated specific factors (unique). They differ in that the common influence in the mutualism model is described from the bottom-up, whereas the common influence in the bifactor model is described from the top-down.

Network and bifactor models produce similar expected covariance matrices, suggesting that they are statistically equivalent (van der Maas et al., 2006). Therefore, these models might be best suited to different levels of analysis rather than an endless theoretical struggle for the 'true' population model. For example, the bifactor model is useful for assessing the between-person reliability of specific factors or subscales when controlling for the general variance (Chen, West, & Sousa, 2006), while network models is useful for predicting within-person changes in network states that predict recovery or relapse (Lutz et al., 2018; van de Leemput et al., 2013).

3.5.3 Does the *p* factor have its basis in nature?

If the *p* factor reflects a general vulnerability to psychopathology that varies in the population, then we would expect it to show some degree of heritability and a plausible neural basis. Of course, it is misguided to think that neurobiological studies are 'proofs' of a construct in nature: data are still collected and analysed, removing the scientist from the natural phenomenon to the realm of empiricism (Tallis, 2016). However, carefully conducted neurobiological studies provide an indirect representation of the underlying organ, in as much as carefully conducted factor analyses provide an indirect representation of a psychological construct (Coan, 1964).

Genetics. An early twin-study by Lahey, van Hulle, Singh, Waldman, and Rathouz (2011) showed that the genetic covariation among disorders (assessed by stronger associations among monozygotic compared to dizygotic twins) was best explained by a general psychopathology factor as well as specific internalizing and externalizing factors. Non-shared environmental covariation among disorders was also best explained by a bifactor model, but the specific factors explained more variance in each disorder than the general factor. Lahey et al. (2011) thus argued for a "generalist genes, specialist environments" etiological model, where genes non-specifically or 'pleiotropically' increase risk for mental disorders, while specific environmental experiences shape the nature of the disorders present.

Pettersson, Larsson, and Lichtenstein (2016) also found that the genetic covariation among mental health disorders (assessed by the stronger incidence rates in full compared to half siblings) was best explained by a general factor as well as two specific factors, e.g., a psychosis-spectrum factor, accounting for the genetic overlap among schizoaffective disorder, schizophrenia and bipolar disorder, and an externalizing factor, with loadings from drug abuse, alcohol abuse, violent criminal convictions, ADHD, but also anxiety (depression and bipolar disorder loaded onto a non-shared environmental factor). Pettersson, Lahey, Larson, and Lichtenstein (2018) further replicated the influence of additive genetic effects on *p* in a twin-study.

Nonetheless, not all studies have shown that *p* is predominantly influenced by genetic variance. For example, Waldman, Poore, van Hulle, Rathouz, and Lahey (2016) explicitly estimated additive genetic, shared environmental, and non-shared environmental factors and used them to predict variation in *p*, internalizing, and externalizing, rather than splitting the covariance matrix into samples with different degrees of relatedness. Based on a process of removal, they found that the *p* factor was equally influenced by genetic, shared environmental and non-shared environmental factors. This study highlights a recurring theme in this chapter: changes in method can have quite important implications the bifactor dimensions estimated. Indeed, family studies tend to produce the strongest, and potentially inflated, general factor estimates compared to other genetic methods (Selzam, Coleman, Caspi, Moffitt, & Plomin, 2018).

Others have directly estimated the proportion of variance in *p* explained by genetic effects. For example, Neumann et al. (2016) reported that 38% of the variance in *p* was explained by single nucleotide polymorphisms (i.e. hereditary genetic variants), but this estimate was highly variable (95% CI: 6-69%). In a twin-study by Harden et al. (2019), 72% of the variance in a parent-reported *p* factor was accounted for by genetic similarity; however, only 49% of the variance in a child-reported *p* factor was attributable to genetic similarity. Similarly, Rosenström et al. (2018) reported that 48% of the variance in the *p* factor was explained by genetic variation in twins. It appears that roughly half of the variance in *p* is hereditary, which fits Waldman et al.'s (2016) findings of a balanced influence between genetic and environmental effects on *p*.

The biometric studies reviewed demonstrate that some aspect of *p* is heritable, but it is unclear what is inherited and whether it is has any relevance to psychopathology. For example, the genetic associations among disorders might simply reflect the transmission of general intelligence, language abilities, or response biases, all of which show some degree of heritability and would influence psychometric test scores (Plomin et al., 2013). Tackett et al. (2013) reported moderate-sized correlations between the *p* factor and negative emotionality, both of which were estimated using genetic associations in twins. Furthermore, Jones et al. (2019) found a significant, albeit weak, association between *p* and polygenic risk scores for neuroticism. Finally, Schaefer et al. (2018) found that 63% of the variance shared between *p* and experiences of victimization (e.g., physical, sexual, or emotional abuse) was explained by genetic overlap between monozygotic compared to dizygotic twins. Therefore, there is good evidence to show that the genetic basis of *p* is linked to factors that would pose a vulnerability to emotional distress and victimization, both of which contribute to psychopathology (Finkelhor, Ormrod, & Turner, 2007; Jeronimus, Kotov, Kiese, & Ormel, 2016).

Neuroimaging. There have been some recent attempts to link the general and specific psychopathology factors to brain structure and function (Dickinson, 2017; Zald & Lahey, 2017). If *p* reflects a disposition to emotional distress (see section 2.2.2), then we would expect it to be associated with key areas involved in explicit emotion regulation strategies such as reappraisal, e.g., the dorsolateral and ventrolateral prefrontal cortex, and implicit emotion regulation strategies such as fear inhibition, e.g., the ventromedial prefrontal cortex and ventral anterior cingulate (Etkin, Büchel, & Gross, 2015). There might also be associations with areas linked to emotion reactivity, e.g., the dorsal anterior cingulate, amygdala, and insula (Ochsner, Silvers, & Buhle, 2012).

Consistent with these predictions, Snyder, Hankin, Sandman, Head, and Davis (2017) found that higher *p* factor scores were associated with reduced grey matter volume in the dorsal, ventrolateral, and orbitofrontal prefrontal cortex in a community sample of 254 adolescents. Furthermore, higher specific internalizing scores were associated with reduced grey matter volume in the insula, amygdala, and medial temporal lobe (including the hippocampus). These findings support the hypothesis that *p* is related to prefrontal regions associated with implicit and explicit emotion regulation strategies. They also provide insight into the neural basis of specific internalizing, which might be associated with bottom-up emotion processing (Etkin et al., 2015). However, we must take care not to misinterpret measures of structure for function; lower grey matter volume does not necessarily reflect poorer emotion regulation abilities (Poldrack, 2010).

Romer et al. (2017) found that higher *p* factor scores were associated with reduced grey matter volume in the cerebellar and occipital cortices in a sample of 1,246 undergraduates. The cerebellum is traditionally associated with the coordination of complex movements, but its role in coordinating thought and affect has become more evident in recent years (Koziol et al., 2014). The cerebellum is heavily connected to the prefrontal cortex (Balsters et al., 2010), and further analyses indicated that *p* was associated with reduced integrity of the white matter pathways within the pons, which acts as a bridge between the cerebellum and the cortex (Middleton & Strick, 2001). The association between *p* and cerebellar grey matter volume was replicated by Moberget et al. (2019), albeit weakly and in the opposite direction (*r* = .13), and using a proxy measure of *p,* e.g., the first principal component extracted from a battery of independent scales administered to a sample of 1,401 adolescents.

One limitation of Romer et al.'s (2017) study is that the statistical correction required for a whole-brain voxel-based analysis might have reduced their ability to detect more subtle differences in morphology associated with *p*. Hinton et al. (2019) took a more detailed approach and investigated the microstructure of major white matter tracts that are critical for connecting different grey matter regions. Hinton et al. found that *p* was positively associated with fractional anisotropy (i.e. white matter integrity) in the body of the corpus callosum, the largest white matter tract that connects the two cerebral hemispheres, in a sample of 410 high-risk adults. While this finding might explain why various disorders are associated with the corpus callosum microstructure (Phillips, Hewedi, Eissa, & Moustafa, 2015), it is surprising that higher *p* scores were associated with an index of greater white matter integrity.

By contrast, Riem et al. (2019) found that an index of *p* (e.g., the first principal component from a battery of independent scales) was negatively associated with fractional anisotropy in the body of the corpus callosum in sample of 126 clinical and non-clinical adolescents. The discrepancy between Hinton et al. (2019) and Riem et al.'s findings might be explained by age-related differences in cortical connectivity associated with psychopathology. For example, general psychopathology in childhood and early adolescence might be associated with hypo-connectivity in cortical networks (Sato et al., 2016), while general psychopathology in adulthood might be associated with hyper-connectivity in networks (Elliot, Romer, Knodt, & Hariri, 2018). It should also be noted that the association between general psychopathology and irregular cortical connectivity has not always been replicated (van Hoof et al., 2019).

On the whole, emerging neuroimaging studies suggest that the *p* factor is associated with structural and functional differences in regions and networks implicated in the coordination of neural activity. This is reminiscent of Spearman's (1927) hypothesis that general intelligence reflects the amount of mental "energy" expendable rather than the integrity of a specific neural substrate. As exciting as these findings are, future studies are needed to ensure that the irregular cortical connectivity associated with *p* is not simply a by-product of distress, interpersonal hardship, or other factors that accompany psychopathology.

A recent study investigating a hypothesised biomarker of *p* (e.g., shorter telomere length) failed to show a prospective link between *p* at age 9 and telomere length at age 13 (Wade, Fox, Zeanah, Nelson, & Drury, 2019). While there was a negative association between *p* and telomere length at age 13 (i.e. higher *p* scores were associated with shorter telomere length), we cannot conclude that general psychopathology *caused* a shortening in telomeres; other variables might account for the relationship. Nonetheless, the prospective association between a marker of brain integrity at age three and *p* factor scores in adulthood reported by Caspi et al. (2014) suggests that irregular cortical connectivity might be an endophenotype of general psychopathology.

3.5.4 Is the *p* factor replicable?

The studies reviewed so far demonstrate a link between *p* and dysregulation across emotional, behavioural and cognitive domains (see section 2.2), which may, in part, be heritable and observable in the efficiency of cortical connectivity (see section 3.5.3). Therefore, the *p* factor shows good construct validity, i.e. the *p* factor likely reflects a transdiagnostic continuum of severity. However, the validity of a construct does not ensure its reliability: How can we be sure that the *p* factors estimated by different research groups broadly reflect the same construct? How can we guarantee that the same construct will be estimated by the same research group on different occasions? It is for this reason that Thomson (1939) argued that a factor's reliability ultimately determines its usefulness, almost at the expense of its theoretical meaning.

We have already seen that there is some variability in the strength of the *p* factors estimated between studies, which is explainable, in part, by methodological differences (see section 3.4.1). We have also seen that the general and specific factors show moderately high differential stability when estimated across time (see section 2.3.2-2.3.4). These stability estimates (median $r = .65$) are of a similar magnitude to those reported in personality research (median $r = .60$; Roberts & DelVecchio, 2000) and intelligence research (median $r = .66$; Schalke et al., 2013), suggesting that the *p* factor shows a trait-like reliability.

3.5.5 Does the bifactor model provide a better representation of the data than other models?

The bifactor model typically reproduces the sample variance-covariance matrix better than the single factor model and correlated factors model (and hence also the higher-order model, which is equivalent to the latter). However, psychopathology researchers can be criticized for their eagerness to demonstrate the bifactor model's superior fit, without considering why this is the case. For a start, the bifactor model is the least constrained model (i.e. the bifactor model features the most estimated parameters and fewest degrees of freedom), so it should naturally fit better than its nested counterparts (Yung et al., 1999). Furthermore, large samples are often needed to adequately estimate bifactor models. However, this often produces significant, albeit minor, differences between the bifactor model and its competitors when tested with chi-square model comparison tests, as these are sensitive to sample size (Bentler & Bonett, 1980). Given the natural bias towards the bifactor model, we should perhaps be asking 'by how much does the bifactor model fit better?' rather than 'which model fits better?'.

The bifactor model has also been criticized for being almost indistinguishable from competing models in statistical terms, despite implying a radically different interpretation of psychopathology (van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017). Indeed, many of the studies reviewed, including the seminal work by Caspi et al. (2014)[7], demonstrate a slightly better fit of the bifactor model compared to

[7] Another reason why some of these studies demonstrate a similar fit between the bifactor and correlated factor models is because thought disorder and internalizing items load preferentially

the correlated factors model (Brodbeck et al., 2014; Conway et al., 2019; Deutz et al., 2018; Lahey et al., 2012; Fernandez de la Cruz et al., 2018; Gomez et al., 2019; Haltigan et al., 2018; Laceulle et al., 2016; Liu et al., 2017; Miller et al., 2019; Patalay et al., 2015; Pettersson et al., 2018; Snyder et al., 2017; St Clair et al., 2017; Tackett et al., 2013; Weissman et al., 2019). However, the implication is that individual differences in all psychiatric problems can be summarized by a single latent trait (as well as more specific traits which account for a small proportion of the variation), which contrasts the notion that there are robust groupings of symptoms (e.g., internalizing, externalizing), which are the product of a more abstract latent trait. It is somewhat unsettling to think that small differences in sampling and measurement may lead to vastly different ways of interpreting the data (see Stochl et al., 2015, for an example).

An increasing number of studies support the claim that the bifactor model overfits the data (i.e. capitalizes on noise), which might explain its marginally superior but significant fit. For example, Murray and Johnson (2013) fitted higher-order and bifactor models to datasets simulated from a higher-order population model. They also included increasing numbers of residual correlations and cross-loadings in the population models, which were not modelled in the higher-order and bifactor test models ('unmodelled complexity'). The bifactor model showed superior fit to the higher-order model, despite the population structure matching the latter, particularly for models with small amounts of unmodelled complexity. Interestingly, fit statistics

onto *p* rather than the specific thought disorder or internalizing factor, respectively. When a specific factor is 'explained away' by the general factor, the bifactor model becomes a bifactor 'S-1' model, which is statistically equivalent to the correlated factors model, except that one of the 'common factors' incudes loadings from all items and is orthogonal to the remaining factors, though they are allowed to freely correlate (Koch et al., 2018).

that penalized for model complexity based on the number of freely estimated parameters (e.g., BIC) favoured the higher-order model for population structures with high amounts of unmodelled complexity, suggesting that overfitting can be accounted for in model comparison. But in practice, unmodelled complexity is likely to be too small to warrant action, but numerous enough to cause a problem, such as being re-expressed through inflated general factor loadings.

Model complexity is not limited to the number of parameters estimated. Bonifay and Cai (2017) argued that complexity also takes a functional form, based on the way in which parameters are specified. For instance, basic non-linear models naturally have a better fitting propensity than linear models, despite having the same number of parameters, due to there being fewer linear constraints. To determine the fitting propensity of the bifactor model, Bonifay and Cai simulated a large random data space of 1,000 datasets which varied in their population structure (e.g., bifactor models, exploratory factor models, unidimensional IRT models, and latent class models), and determined the percentage of datasets that the bifactor model as well as other models fit at different thresholds of the Y2/N fit statistic, which is akin to RMSEA for binomial IRT data.

At a moderate threshold of model fit (Y2/N = .05), the bifactor model fit almost as many datasets as an EFA model (64% vs. 79%) which is designed to find an optimal solution to the data, while the latent class and unidimensional models were more conservative (2.3-5%). Critically, all models had the same number of parameters, demonstrating that a model’s functional form is also important in determining model fit estimates (and the extent to which a model might overfit the data). The bifactor model was also better at explaining local dependencies between items compared to the EFA model (e.g., residual correlations which reflect model misspecification). Bonifay and Cai argued that the bifactor model's ability to fit local noise in the data might account for its superior fit, but further work is needed to determine a direct link.

Such work was conducted by Reise, Kim, Mansolf, and Widaman (2016), who identified participants that showed implausible response patterns (i.e. a large distance between a person's actual and estimated response pattern) and un-modelable response patterns (i.e. large residuals after fitting a person's response pattern using a given model) on the Rosenberg Self-Esteem Questionnaire. The same percentage of cases showed implausible item response patterns in the bifactor model, correlated factors model, and single factor model (31% in each), which is unsurprising given the similarity in implied variance-covariance matrices between models (van Berk et al., 2017). However, fewer participants showed unmodelable response patterns in the bifactor model (11%) compared to the correlated factors (12%) and single factor (14%) models. Most participants who showed un-modelable response patterns contributed to the superior fit of the bifactor model over the single factor model. Therefore, the bifactor model shows a superior fit, in part, due to its ability to accommodate unlikely or noisy item response patterns.

One problem with this interpretation is that it only applies to the bifactor and single factor models: it does not explain why both bifactor and correlated factor models accounted for almost the same number of item response patterns (87% and 88%, respectively). Furthermore, responses are 'implausible' relative to the unidimensional model, which does not mean that they represent random noise (although some of the example response patterns presented by Reise et al., 2016 were clearly nonsensical). To understand this point, consider Reise et al.'s (2016) observation that participants with implausible response patterns showed a more reliable specific negative-wording factor than those with plausible response patterns. A stronger negative wording factor has also been observed in people who score highly on neuroticism (Quilty et al., 2006), show lower verbal abilities (Gnambs & Schroeders 2017), and are raised in poorer socioeconomic conditions (Schmitt & Allik, 2005). In other words, these 'implausible' responses may have a substantive basis (e.g., treating negatively worded items as a separate domain related to negative affect) that is readily accommodated by specific factors in the bifactor model.

In all, the evidence reviewed suggests that the bifactor model shows a superior fit because it accommodates local noise, not because it identifies the 'true' population structure. However, 'noise' may have substantive meaning, and hence the bifactor model may just be a 'supercharged' unidimensional model that accommodates individual variation in response styles. This might also explain why the unidimensional model typically shows near-acceptable model fit when applied to psychopathology data: the structure of psychopathology is "essentially unidimensional" (Reise et al., 2010) but includes groupings of symptoms that capture specific styles of expression (compare, for instance, the omega hierarchical and explained common variance estimates presented earlier in the chapter).

It is also worth noting that the unidimensional, correlated traits, higher order, and bifactor models are restricted versions of each other and fall on a continuum (Widaman & Thompson, 2003). In fact, most of these models tend to recreate the variance-covariance matrix adequately at an absolute level (Morgan et al., 2015) which leaves the SEM field in a conundrum: how do we select between near-equivalent models? Reise et al. (2010) suggested that model selection is ultimately guided by the research context, often inadvertently. Therefore, it might be helpful to think of models as tools for summarizing a measure in a manner that is most useful for a given need (Thomson, 1939), rather than a representation of a true population structure (Murray et al., 2016). When there is a high degree of overlap between indicators, the bifactor model might be the preferred option to assess the impact of a purported general trait or the reliability of specific factors after controlling for the general trait.

3.5.6 What is the impact of shared variance beyond the *p* factor?

Confirmatory factor models, bifactor or otherwise, assume a simple structure, also known as an independent cluster structure, where each item loads onto one and only one factor (McDonald, 1999; Thurstone, 1947). An item that loads onto two or more factors is said to 'cross-load'. Cross-loadings violate the independent cluster structure and can distort model fit statistics and factor loadings (Reise, Moore, & Maydeu-Olivares, 2011). The presence of cross-loadings implies that there are unmodelled factors and hence the model is mis-specified (McDonald, 1999).

The impact of cross-loadings on the confirmatory bifactor model has yet to be thoroughly investigated. From a theoretical standpoint, if the general factor accounts for the shared variance among items, why should there be additional shared variance indicated by the presence of cross-loadings? Correlations between specific factors also reflect shared variance beyond the general factor. In the original bifactor model, specific factors are assumed to be orthogonal (Holzinger & Swineford, 1937), yet freeing the correlations among specific factors is becoming increasingly popular in psychopathology research (i.e. the 'revised' model; Afzali et al., 2017; Arrindell et al., 2017; Brodbeck et al., 2014; Carragher et al., 2016; Caspi et al., 2014; Hyland et al., 2018; Laceulle et al., 2016; McElroy et al., 2017; Neumann et al. 2016; Patalay et al., 2015; Pettersson et al., 2018; Preti et al., 2018; Urbán et al., 2016; Urbán et al., 2014). Cross-loadings and specific factor correlations ultimately suggests that there are unspecified sources of variance beyond the general factor, and hence the bifactor model can be mis-specified.

Aside from the study by Murray and Johnson (2013; see above), only two studies have assessed the impact of cross-loadings on the bifactor model. Reise, Moore, and Maydeau-Olivares (2011) simulated exploratory bifactor models with independent cluster structures that differed in the strength of their general and specific factors, and whether they included cross-loadings. Items that cross-loaded showed upward-biased general factor loadings, and downward-biased specific factor loadings. The degree of bias lessened with larger simulated sample sizes, but only if the general factor strength was weak. That is, the ameliorating effect of larger sample sizes on parameter bias was lessened if the general factor accounted for most of the variance in the simulated dataset, presumably because it was more able to absorb the unmodelled complexity (Murray & Johnson, 2013).

In a confirmatory multidimensional IRT model, Finch (2011) found that discrimination parameters (which are somewhat akin to factor loadings) were also overestimated in the presence of cross-loadings. Clearly, more studies are needed to determine the influence of cross-loadings on the bifactor model. However, it is likely that freeing cross-loadings, or even constraining small but substantial cross-loadings to zero, artificially inflates the *p* factor variance, particularly when the general variance is already strong. The *p* factors reviewed earlier on in the chapter were moderately strong on average (see section 3.3), and hence more likely to accommodate unmodelled noise. The most frequent way of managing a surplus of shared variance is to free the specific factor correlations, but as described above, this introduces a new issue of model misspecification and how to interpret the shared variance beyond *p*.

One potential solution to managing unmodelled shared variance is bifactor exploratory structural equation modelling (ESEM; Asparouhov & Muthén, 2009). To understand ESEM, it is helpful to reiterate the main difference between EFA and CFA. In EFA, all loadings (and hence cross-loadings) are specified, but errors are assumed to be random. In CFA, errors have a systematic and random component, allowing one full control over the error structure. In other words, one can constrain factor loadings to zero; an independent cluster structure is the product of freeing factor loadings on some factors and constraining loadings to zero on others. In ESEM, one specifies all loadings but with constraints. All loadings and cross-loadings can be specified and adjusted for using a factor rotation, as in typical EFA, but since it is in an SEM context, one gets CFA parameters, including standard errors of factor loadings, goodness of fit statistics, and residual correlations. With E-SEM, one is no longer tied to the independent cluster assumption of CFA; all cross-loadings can be estimated and tested with significance tests. One can also specify a bifactor model using a target rotation, providing a powerful tool to model both the hierarchical nature of items (e.g., symptoms measuring both common and specific constructs), as well as their infallibility (e.g., symptoms hardly ever measuring one and only one specific construct; Morin, Arens, & Marsh, 2016).

One study has estimated a bifactor model of psychopathology using ESEM. Lahey et al. (2017) compared a confirmatory bifactor model that included a *p* factor and orthogonal internalizing and externalizing factors to an ESEM model that included the same factors but also permitted all cross-loadings in a sample of 499 young adults (corrected for clustering due to twin-pair membership), using diagnostic scores from the Young Adult version of the Diagnostic Interview for Children. The E-SEM model fit the data excellently ($\chi^2(53)$= 70, p = .055, CFI = .969, TLI = .946, RMSEA = .026, SRMR = .034, BIC = 21,019), as did the CFA model ($\chi^2(64)$= 81, p > .05, CFI = .969, TLI = .956, RMSEA = .023, SRMR = .036, BIC = 20, 978). Furthermore, the factor loading matrices were similar between models: all items in the E-SEM model loaded significantly onto one specific factor only.

Lahey et al.'s (2017) findings suggest that unmodelled complexity might not always be a threat, since their ESEM model maintained a simple structure despite specifying all cross-loadings. However, it is likely that cross-loading strength and hence unmodelled complexity was masked by the use of disorder-level indicators. Symptom-based indicators might cross-load more strongly as some symptoms are featured in multiple disorders and overlap in their wording. Nonetheless, Lahey et al. recreated their disorder-level indicators (e.g., disorder sum scores) after removing overlapping symptoms, and reproduced the simple structure solution. However, it should also be noted that the depression and PTSD indicators no longer loaded significantly onto the internalizing factor (they loaded exclusively onto *p*), and marijuana use no longer significantly loaded onto *p* (it loaded exclusively onto specific externalizing) in the ESEM model. Therefore, changes might still occur in the loading structure as a result of freeing cross-loadings with ESEM.

3.5.7 Does the *p* factor predict all psychiatric problems equally?

In his theorem on the indifference of the indicators, Spearman (1927) argued that the type of intelligence test does not matter when assessing *g*. So long as the battery of tests are equally associated with *g*, they will provide equal measurements of general intelligence. The same logic can be applied to the assessment of psychopathology: if *p* represents a single underlying dimension of psychopathology that exists beyond the tests used to measure it, then it too should be equally predictive of multiple symptoms or disorders.

Not everyone agrees that *g* is invariant across indicators. Historically, Thurstone (1940) argued that *g* was influenced by the particular test used, and therefore varies between test batteries. Note that Thurstone did not refute the invariance of indicators theorem, but thought that it only applied to his common factors. Despite the evidence supporting the invariance of *g* factor indicators (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, Nijenhuis, & Bouchard, 2008; Warne & Burningham, 2018), there is an element of circularity in the theorem: any test, regardless of its content, supports the invariance of *g*, if it is associated with *g* in the first place. Furthermore, bifactor studies of intelligence show that fluid intelligence tasks load almost exclusively onto *g*, implying a special relationship between the two that should not occur if all tasks are invariant (Gustafsson & Åberg-Bengttson, 2010).

One can test the indifference of the indicators hypothesis by assessing whether *g* (or *p*) is strongly correlated across different measures (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, Nijenhuis, & Bouchard, 2008), or by estimating a meta-general factor over multiple samples (Warne & Burningham, 2018). As this is rather difficult to achieve in practice, one can instead use the correlated vectors method, where the general factor loadings of a set of indicators is correlated with those indicators' associations with an outcome variable that is thought to influenced by the underlying trait (Jensen, 1998; Vainik, Mõttus, Allik, Esko, & Realo, 2015). Indicators are invariant to the extent that their general factor loadings are equally correlated with their association with an external criterion. The correlated vectors method is not without issues, however, such as producing significant associations even when the underlying trait does not fully explain the association between the indicator and external variable, such as when content-unrelated indicators are used (Wicherts, 2017). A simple alternative is the item removal method, where a bifactor model is continuously re-run, each time removing a different indicator that loads on the general factor.[8] If there is no interaction between certain indicators and the general factor, then the specific indicator removed should have no effect on the factor loading matrix (Vainik et al., 2015).

[8] One could also test the indicator invariance hypothesis by comparing the fit of a model with factor loadings constrained to equality compared to one where they are freely estimated (i.e. metric invariance).

One study has investigated the indifference of indicators assumption of *p* using the item removal method. Lahey et al. (2017) estimated a bifactor model with a *p* factor and orthogonal internalizing and externalizing factors (see above). Lahey et al. repeatedly estimated the model, each time removing a diagnostic indicator and examining the impact on model fit indices. The models showed equally excellent fit. Furthermore, most *p* factor loadings were similar across models, correlating *r* = .9 on average. The only substantial change occurred when the antisocial personality disorder indicator was removed: drug and alcohol abuse indicators no longer loaded significantly onto the externalizing factor. By and large, these findings support the view that *p* is invariant to the indicators used.

However, it can be argued that Lahey et al.'s (2017) test of indicator invariance lacks sensitivity. The fact that each bifactor model converged and maintained excellent fit despite removing an indicator is not surprising, since Lahey et al. essentially re-specified a model that still included all relevant paths. A more sensitive approach would be to examine how *p* factor loadings changed in strength and direction after constraining an indicator to zero. Such a model could also be compared statistically to a model without these constraints; Lahey et al. could not compare their models statistically as they were not nested. Moreover, reliability theory suggests that removing a single indicator is unlikely to dominate model fit indices when there are multiple indicators present.

Unfortunately, Lahey et al. (2017) only reported whether indicators remained significant with each removal. However, substantial changes in loading strength can occur without disturbing an indicator's significance. This thesis' review of bifactor studies showed that both depression and thought disorder items tend to load most strongly onto the *p* factor, and most weakly onto the specific internalizing factor (see sections 2.2.1 and 2.2.3, respectively). One does wonder whether the internalizing indicators in Lahey et al.'s study showed substantial changes in loading strength after removing the depression indicator (thought disorder was not assessed).

More generally, the fact that depression and thought disorder items load most strongly onto *p* contradicts the indifference of indicators hypothesis. I have already outlined potential reasons for these preferential loadings (see section 2.2). For instance, the *p* factor might reflect a disposition to emotional distress (see section 2.2.1), a continuum of disordered thought (see section 2.2.2), or a universal response to human suffering (see section 2.2.3). Alternatively, depression indicators might load preferentially onto *p* because they overlap in content with other disorders, while thought disorder indicators might load preferentially onto *p* because they are inherently unreliable (see section 2.2.1). Regarding the role of content overlap, Lahey et al. showed that the depression indicator did not change in loaded strength after removing overlapping symptoms, making this an unlikely contributor to its preferential loading strength. Regardless of the cause, symptoms and disorders do not appear to represent *p* in a similar way, questioning traditional views that general factors reflect latent constructs distinct from the data.

Chapter 4 Fact or Artifact? Testing the Response Bias Hypothesis of the General and Specific Psychopathology factors

4.1 Introduction

I concluded Chapter 3 with a methodological review showing that the *p* factor is observed across multiple respondents, is moderately heritable and associated with cortical dysconnectivity, and is stable over time. However, it also drew attention to issues such as the *p* factor's varied rather than uniform relationship with psychiatric problems, its tendency to overfit noise in the data, and its weaker presence when assessed with network models and multi-informant designs. Overall, the studies reviewed suggest that the *p* factor has a substantive basis, but it might be overstated due to issues with our methods.

One methodological issue that has yet to be tested is whether the *p* factor is, in part, a product of response biases (i.e. consistencies in the way that people respond to content-unrelated features of self-report measures). The aim of this chapter is to test the response bias hypothesis of the general and specific psychopathology factors. I begin by defining response biases, followed by a review of methods used to control for response biases in the context of psychopathology measures. I then estimate a bifactor model of psychopathology in a large online sample of community adults (N = 1,200) and examine the amount of variance in the general (*p*) and specific psychopathology factors explained by response bias factors estimated from a set of heterogeneous items. I also test whether responses biases account for the relationship between the *p* factor and neuroticism trait, as well as the correlations between specific factors.

4.1.1 What are response biases and how might they account for the general and specific psychopathology factors?

Response biases (also known as response styles or response sets) are systematic patterns in responding that are unrelated to the construct assessed (Paulhus, 1991; Wetzel, Böhnke, & Brown, 2016). Common response biases include agreement bias (a tendency to agree with questions regardless of their content; also known as acquiescence or yeasaying), disagreement bias (a tendency to disagree with questions regardless of their content; also known as disacquiescence or naysaying), extreme responding (a tendency to use the extreme ends of the response scale regardless of an item's content), and mid-point responding (a tendency to use the middle response options regardless of the item's content; Van Vaerenbergh & Thomas, 2012).

Baumgartner and Steenkamp (2001) outlined two effects of response biases on the validity of self-report data. The first, which I call 'intra-scale effects', describes how response biases can artificially inflate or deflate scores on an item or scale (Bentler, Jackson, & Messick, 1971). Therefore, a discrepancy may exist between a respondent's 'true' score on an underlying construct, and their observed score contaminated by response biases. As a result, two respondents might score similarly on a factor despite ranking differently on the underlying trait because one respondent's scores were skewed by response biases (Böhnke & Croudace, 2015). Response biases thus are a source of systematic measurement error that need to be controlled for (Block, 1965; Savalei & Falk, 2014).

The second 'inter-scale' effect that response biases can have is by inflating or deflating the relationship between scales completed by the same respondent (Baumgartner & Steenkamp, 2001). Two different scales might be correlated, in part, because systematic response tendencies influence scores on each scale in similar ways (Danner, Aichholzer, & Rammstedt, 2015; Messick, 1991). Response biases can effectively act as 'third variables' that distort our understanding of the relationships between scales and increase the likelihood of type I and type II errors (Podsakoff, MacKenzie, & Podsakoff, 2012). Response biases are seldom controlled for in research and applied settings, despite the threats they pose to scale interpretation (Aichholzer, 2014; Forbey, Lee, Ben-Porath, Arbisi, & Gartland, 2013; Rammstedt & Farmer, 2013; Rios, Guo, Mao, & Liu, 2017; Wiggins, Wygant, Hoelzle, & Gervais, & Gevais, 2012).

Some have raised concerns that the *p* factor might be a product of response biases (Böhnke & Croudace, 2015; Caspi & Moffitt, 2018; Lahey et al., 2012). The *p* factor captures the common variance among items, part of which will be underpinned by individual differences in the experience of various problems, but other parts will be driven by systematic method effects, including response biases, item wording, item ordering, scale presentation, and the psychological and situational context (Böhnke & Croudace, 2015). The question is, how much of the variance in the *p* factor is accountable by common method effects, in particular, response biases?

One possibility is that most of the variance in the *p* factor is due to the intra-scale effects of response biases that induce positive correlations between symptom-items. For example, individual variation in the tendency to agree to items indiscriminately (e.g., agreement bias), combined with variation in the tendency to disagree to items indiscriminately, could produce a positive manifold across items, without the need for an underlying trait (see Figure 4.1). However, it is more likely that a small to moderate portion of the variance in *p* is attributable to response biases, given that 25% of the variance in personality scales is estimated to be driven by common method effects more generally (Cote & Buckley, 1987). This could still influence the weaker correlations between symptoms, since the *p* factor does not predict all problems similarly (see section 3.5.7). The intra-scale effect of responses biases could inflate or even induce the weak correlations between distant symptoms, which would challenge the notion of a positive manifold between symptoms.

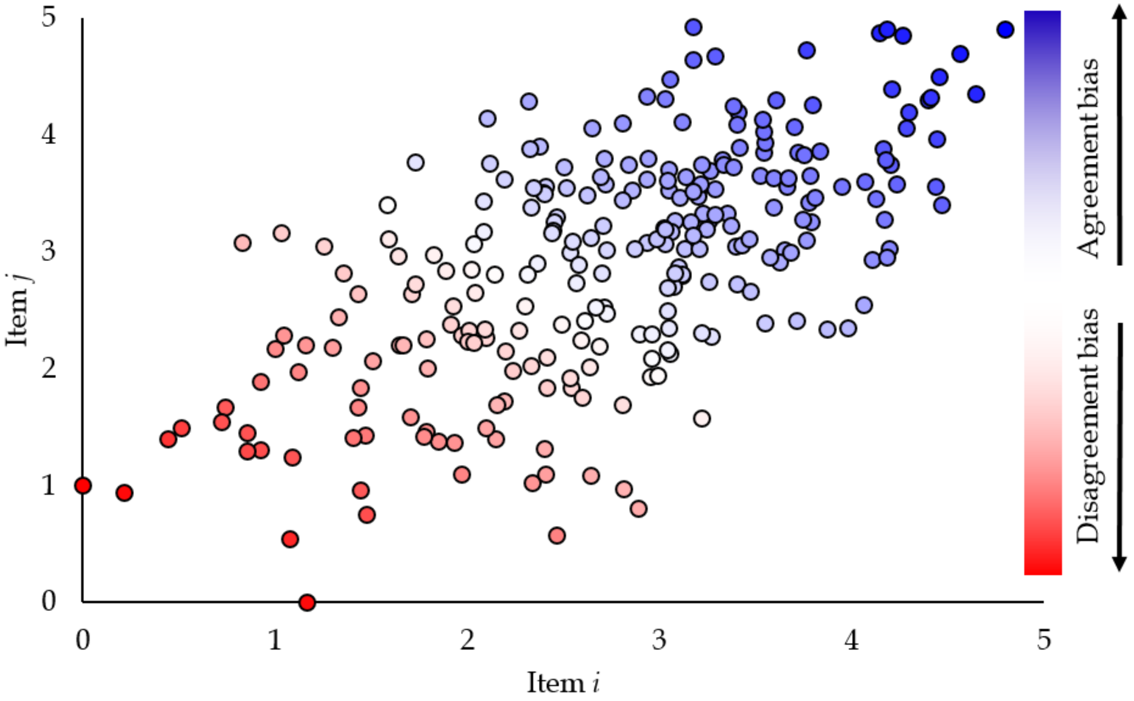

Figure 4.1. Illustration of how agreement and disagreement biases could induce positive correlations between any two items in a scale. Item scores were simulated as continuous variables for demonstraton purposes.

Furthermore, the inter-scale effects of response biases could inflate the associations between the *p* factor and theoretically relevant variables. For example, the *p* factor is consistently associated with neuroticism, which implies that the *p* factor reflects a disposition to emotional distress (see section 2.2.1). However, most studies that assess *p* and neuroticism using use single respondent. Therefore, these two variables might be positively correlated because assessments of psychopathology and neuroticism are both influenced by negative response tendencies (i.e. individual differences in the extent that respondents view themselves and the world in negative terms; Lahey et al., 2012), rather than because true scores on the underlying constructs overlap substantively.

The reader might recall that the *p* factor is associated with various external outcomes that are unaffected by response biases, including performance on neuropsychological tests (Bloeman et al., 2018; Caspi et al., 2014; Castellanos-Ryan et al., 2016; Harden et al., 2019; Martel et al., 2017; Neumann et al. 2016; White et al., 2017), academic attendance and attainment (Constantinou et al., 2019; Lahey et al., 2015; Patalay et al., 2015; Pettersson et al., 2018; Sallis et al., 2019), and biological data (Neumann et al., 2016; Romer et al., 2017; Rosenström et al. 2018; Snyder et al., 2017). While the associations cannot, strictly speaking, be driven by the inter-scale effects of response biases, they tend to be weak in magnitude, and hence provide weak evidence for the *p* factor's criterion validity.

The intra-scale and inter-scale[9] effect of response biases might also account for the variation and covariation in specific psychopathology factors, respectively. Some view specific factors in the bifactor model as method factors or nuisance variables that capture method effects such as response biases or superficial item groupings (Cho, Cohen, & Kim, 2014). Correlations between method factors would also explain why there is shared (artifactual) variance beyond the common factor (see section 3.5.6). Treating specific factors as nuisance variables might also justify their weak reliability beyond the *p* factor (see section 3.4.2). While we generally refer to specific factors with substantive names (e.g., internalizing and externalizing), we must not forget that these are factors are distinct from and residualized for the general factor. Therefore, they do not have the same interpretation as the internalizing or externalizing factors we refer to (Bonifay, Lane, & Reise, 2017).

4.1.2 How do you measure response biases?

Several methods have been developed to assess response biases (Van Vaerenbergh & Thomas, 2011). To narrow the scope of the review, I will focus on latent variables methods as they can be easily integrated into the bifactor model.

Item Reversals. Perhaps the most common method to measuring response biases is to invert the response scale for roughly half of the questionnaire items. Some items might be positively worded, where agreement indicates higher scores on the underlying trait (e.g., agreement to the item "I feel I have a number of good qualities"

[9] I recognise that the covariance between specific factors attributable to response biases would technically count as an 'intra-scale' effect, since specific factors are assessed with the same scale, but for exposition purposes I have classed them as inter-scale effects.

reflects higher self-esteem), whereas other items might be negatively worded, where disagreement indicates higher scores on the underlying trait (e.g., disagreement to the item “I certainly feel useless at times” reflects higher self-esteem). Item reversals were initially used to prevent indiscriminate agreement to items, since agreeing to a negatively worded item would run counter to previous ratings (Bentler, Jackson, & Messick, 1971). Researchers extended the approach by estimating separate method factors for positively and negatively worded items, to capture the degree of agreement and disagreement bias, respectively, after accounting for a common trait running across items (Billiet & McClendon, 2000; DiStefano & Motl, 2006; Tomás, Oliver, Galiana, Sancho, & Lila, 2013).

There is, however, some debate about whether item wording factors reflect response biases or introduce variations of the construct in question. Most research surrounding this debate comes from studies of the Rosenberg Self-Esteem Scale (RSES), an early measure of self-esteem that includes five positively worded items (e.g., ‘I feel that I have a good number of qualities’) and five negatively worded items (e.g., ‘I feel I do not have much to be proud of’; Rosenberg, 1965). For example, Horan, Distefano, and Motl (2003) assessed the longitudinal stability of a negative wording factor assessed from the RSES across three bi-annual waves in 14,374 secondary school students. If the negative wording factor has substantive qualities, then adolescents’ rank-ordering should remain relatively stable over time (see differential stability, section 2.3.2-2.3.4).

Horan et al. (2003) found that the negative wording factor showed moderate stability over time: the autoregressive coefficient (*B*) was .44 between waves one and two, and also .44 between waves two and three. Horan et al. argued that the 'trait-like' stability observed in the negative wording factor, and hence the associated response biases, would not be observed unless it had a substantive quality. However, these stability estimates conflate consistencies in general disagreement tendencies with consistencies in self-esteem. Horan et al. argued that their negative wording factor was 'content-free' as it correlated with negative wording factors estimated from the Locus of Control Scale (*r* = .32) and Attitude to School Scale (*r* = .31) at wave one. I would, however, argue that these correlations are too weak to reflect an overarching construct, and the poor fit of a general negative wording factor estimated across these scales weakens their case. In all, Horan et al.'s (2003) study, while innovative, does not reliably address the question of whether item wording factors introduce substantive traits.

Stronger evidence for the substantive nature of item wording factors is by Marsh, Scalas, and Nagengast (2010). Also taking the assumption that nuisance variables should be transient over time, while substantive response biases should be stable, Marsh et al. assessed the stability of a bifactor model with a general self-esteem factor and specific positive and negative wording factors estimated from the RSES. Marsh et al. first demonstrated scalar invariance of the bifactor model (i.e. consistency in the factor loadings, intercepts, and factor variances), which is critical to ensuring that any change observed in factor means is due to changes in the factors themselves and not changes in item meaning. They found that the positive and negative wording factor means did not change over time and showed moderate autoregressive correlations over four waves (*r* = .40-.65). Therefore, the trait-like nature of Marsh et al.'s positive and negative wording factors suggest that the (implied) response biases have substantive qualities.

While Marsh et al.'s (2010) study was well-controlled, it lacked measures of criterion validity. Therefore, it is uncertain whether their wording factors represented response biases as they argued. To test this, Arias and Arias (2017) estimated a negative wording factor using the Core Self-Evaluations Scale, which overlapped with self-reported negative affect over and above a general self-evaluations factor. These findings suggest that their negative wording factor reflected a response bias associated with negative responding (e.g., viewing oneself, others and the world in negative terms; Podsakoff et al., 2003). However, Arias and Arias face the same problem as Horan et al. (2003): the negative wording factor could simply reflect a variation of self-esteem associated with negative items (e.g., depressive characteristics). Others have also demonstrated that negatively worded items on the self-esteem measures reflect low mood or negative affect, but the direction of association between the negative wording factor and low mood is mixed (DiStefano & Motl, 2006; Lindwall et al., 2012; Michaelides, Koutsogiorgi, & Panayiotou, 2016; Quilty, Oakman, & Risko, 2006).

There are several lessons to be learned about estimating response biases in psychopathology measures using item reversals. The studies by Horan et al. (2003) and Arias and Arias (2017) suggest that specific wording factors reflect more than just responses to item wording; they also carry meaning associated with the content of items.[10] Therefore, unless they are estimated across different measures that are

[10] This also implies that specific internalizing and externalizing factors might reflect problems in each domain rather than nuisance variables, since they, like wording factors, are estimated with a bifactor model.

unrelated in content, wording factors will conflate response biases with specific domains of a construct. This is not to say that response biases are not underpinned by substantive characteristics separate from the content of the items used to measure them. For example, agreement bias has been linked to lower cognitive ability and flexibility (Knowles & Nathan, 1997; Lechner & Rammstedt, 2015) and collectivistic cultural norms (Rammstedt, Danner, & Bosnjak, 2017), while disagreement bias has been linked to oppositional characteristics (Knowles & Nathan, 1997) and individualistic cultural norms (Baumgartner & Weijters, 2015). The substantive nature of response biases might also interact with the specific content of the questionnaires (Podsakoff et al., 2003), adding a further level of complexity that is difficult to control for when both response biases and the construct of interest are assessed with the same scale.

Another obstacle to using item reversals to assess response biases in psychopathology measures is that most, if not all, measures are dominated by positively worded symptom-items.[11] Rarely do assessment measures ask whether people do not experience a certain symptom (unless a diagnosis has exclusion criteria). Some bifactor studies of psychopathology have included well-being items (Murray et al., 2016; St Clair et al., 2017), which could be regarded as negatively worded items. However, estimating a specific wellbeing factor and negative wording factor from the same set of items is not possible (due to a singularity in the factor matrix), and would

[11] Some mental health questionnaires, such as the General Health Questionnaire (GHQ-12) feature positively and negatively worded questions (e.g., "felt constantly under strain" vs. "able to face problems"), but they tend to have a broad focus compared to symptom-items that describe circumscribed problematic thoughts, feelings, and behaviours.

otherwise conflate content (e.g., well-being) and style (e.g., responses to negatively worded items).

One study to my knowledge has controlled for a wording effects when estimating a general factor from mental health scales. Böhnke & Croudace (2016) estimated a general well-being factor (which is not quite general psychopathology) from items on the General Health Questionnaire (GHQ-12), Warwick Edinburgh Mental Well-being Scale, and EuroQol Health Status measure (EQ-5D), as well as specific factors for each scale. They also estimated a negative wording factor from the negatively worded items on the GHQ-12. Böhnke & Croudace were able to estimate a healthy general well-being factor while controlling for a negative wording factor, which suggests that the response biases associated with negative wording (e.g., disagreement bias) do not account for general factors in mental health scales. However, the reliability of their negative wording factor is questionable, since only 2/6 items showed practical significant loadings (e.g., λ = > |.3|; Hair et al., 1998), and it was not validated against substantive qualities.

Response Bias Indicators. An alternative method to estimating response biases is to use an independent measure of heterogeneous items. If the items are heterogeneous in content, then any consistencies in responding should be due to response biases (or other method effects) rather than content (Greenleaf, 1992). Unlike positive and negative wording factors, a heterogenous measure of items does not conflate content with style, making it ideal for measuring response biases alongside psychopathology.

Early studies analyzed response biases by counting the frequency of certain responses to heterogeneous items. For example, agreement bias scores for each participant would be calculated by summing the number of times that the highest agreement response category was used (e.g., 'Strongly Agree'; Bachman & O'Malley, 1984), or by producing a weighted average of the use of each agreement response category (e.g., 'Strongly Agree' * 3 + 'Agree' * 2 + 'Somewhat Agree'*1/total number of items; Baumgartner & Steenkamp, 2001). In either case, each participant would be scored on the extent to which they agreed to items regardless of their content.

More recently, Weijters, Schillewaert, and Geuens (2008) extended this method by estimating factors for different response biases rather than simply adding or averaging the response frequencies. In the 'representative indicators response style means and covariance structure' method, or the response bias indicators (RBI) method, the items in a heterogeneous measure are split into an equal number of parcels. The frequency of certain response options is then averaged for each parcel of items, producing indicators representative of a given response bias. Finally, the indicators from each parcel are used to identify response bias factors that can be estimated alongside substantive factors (e.g., psychopathology factors) in a structural equation model (the RBI method is described more fully in section 4.2.5).

Prior studies using the RBI method have shown that response bias factors influence responses consistently throughout the completion of a questionnaire (Weijters, Geuens, & Schillewaert, 2010a), are stable over different measurement occasions (Weijters, Geuens, & Schillewaert, 2010b), and show similar characteristics across different modes of data collection (e.g., pen-and-pencil, telephone, and online questionnaires; Weijters, Schillewaert, & Geuens, 2008). Furthermore, by separating out the substantive variance from the error variance in response bias measures with latent variables, one increases the predictive power of responses biases, making the RBI method suitable for testing the contribution of response biases to substantive factors, such as psychopathology factors (Weijters et al., 2008).

The RBI method was developed by marketing researchers and hence has not been widely adopted in psychopathology research. However, Stone, Schneider, Junghaenel, and Broderick (2019) used the RBI approach to investigate the impact of response biases on global health. They found that disagreement bias, but not agreement bias, was weakly but significantly correlated with perceived physical and mental health (r = .23) and life-satisfaction (r = .33). Furthermore, disagreement bias weakened the positive prediction of age on perceived health and life satisfaction, particularly in younger adults. It is uncertain why a tendency to disagree indiscriminately to questionnaire items would predict higher, rather than lower, ratings of perceived health and life satisfaction. However, these findings illustrate how self-reported health outcomes are not unbiased estimates of people's 'true' health status, and partly reflect content-unrelated response tendencies. Early studies also highlighted the biasing effect of both agreement and disagreement biases on disorder-specific reports using classical measures of response bias (Phillips & Clancy, 1970; Roberts, Forthofer, & Fabrega Jr., 1976; Williams, Tarnopolsky, & Hand, 1980).

4.1.3 Study Aims

This study investigates the contribution of response biases to the general (*p*) and specific psychopathology factors. It also investigates the influence of response biases on the associations between specific psychopathology factors, and association between the *p* factor and neuroticism. A bifactor model was estimated from responses on the Achenbach Adult Self Report completed by a large online sample of community adults (N = 1,200). Response biases were assessed with an independent measure of heterogeneous items from which response bias indicators were derived and used to model latent response bias factors. The general and specific psychopathology factors were then regressed onto the agreement and disagreement response bias factors to examine the latter's influence.

If the general or specific psychopathology factors are partly attributable to response biases, then agreement and disagreement biases should positively and negatively predict a moderate amount of variance in each set of factors, respectively (e.g., up to 25% of the variance attributable to method effects; Cote & Buckley, 1987). Furthermore, if the correlations between specific factors are inflated or deflated by response biases, then there should be significant differences between the correlation coefficients before and after correcting them for response biases. Similarly, the association between the *p* factor and neuroticism should be significantly different after controlling for response biases.

4.2 Methods

4.2.1 Participants

Respondents were recruited via the online crowdsourcing site Prolific Academic (www.prolific.ac). They were pre-screened for age (18+), nationality and country of residence (United Kingdom), language fluency (English), and approval rate (>=90%). A total of 1,200 respondents were recruited from a pool of 11,800 who were eligible to participate. The sample mainly consisted of Caucasian adults (94% White/White British; 66% female), with a mean age of 37 (*SD* = 13, range = 18-81). Most respondents reported being in full- or part-time work (68%) and completed some form of higher education (e.g., A-levels or an undergraduate degree; 69%). Almost one third of the sample (30%) reported a household income of ≤£5,199-£20,799, 22% reported a household income of £20,800-£31,199, and 48% reported an annual household income of £31,200-£52,000+. Most participants (69%) reported being in a relationship, but many reported never being married or registered in a same-sex civil partnership (59%). A total of 28% of respondents reported having a current mental health diagnosis, of which 58% reported they were taking medication for. Furthermore, 36% and 37% reported a personal or family history of mental health problems, respectively. A full demographics breakdown can be found in Table 4.1.

Table 4.1

Demographic Breakdown for the Community Sample of Online Respondents (N = 1200)

Demographic	*N*/*M*	%/*SD*
Age (*N* = 1200)	37	13
Study completion time (min; *N* = 1200)	27	14

Sex (*N* = 1200)		
Male	409	34%
Female	791	66%
Ethnicity (*N* = 1200)		
White/White British	1,127	94%
Mixed/multiple ethnic groups	25	2%
Asian/Asian British	25	2%
Black/Black British	19	2%
Other	4	0.3%
Employment Status (*N* = 1200)		
Full-time	547	46%
Part-time	269	22%
Unemployed	68	6%
Student	119	10%
Not in paid work	197	16%
Highest Level of Education (*N* = 1200)		
No formal qualifications	16	1%
Secondary school/GCSEs or equivalent	180	15%
College/A-levels or equivalent	400	33%
Undergraduate degree (BA/BSc/other)	428	36%
Graduate degree (MA/MSc/other)	134	11%
Professional degree (PhD/Other)	42	4%
Annual Household Income (*N* = 1190)		
Up to £15,599	227	19%
£15,600-£-£31,199	407	34%
£31,200-££52,000+	556	48%
Marital Status (*N* = 1198)		
Never married/in civil partnership	701	59%
Married/in civil partnership	413	34%
Separated/Divorced/Widowed	84	7%
Currently in a relationship (*N* = 1199)		
Yes	831	69%
No	368	31%
Current MH diagnosis (*N* = 1200)		
Present	340	28%
Absent	860	72%
History of MH problems (*N* = 1200)		
Present	435	36%
Absent	765	64%
Family history of MH problems (*N* = 1200)		
Present	446	37%
Absent	754	63%
Medication for MH problems (*N* = 340)		
Present	197	58%
Absent	143	42%

Note. MH = Mental Health.

4.2.2 Measures

Adult Self-Report (ASR; Achenbach & Rescorla, 2003). The ASR is a 126-item self-report measure of common mental health problems in adulthood, including anxiety, depression, somatic complaints, thought problems, attention problems, aggressive behaviour, rule-breaking behaviour, intrusive behaviour, and substance misuse. Items are rated on a 3-point scale (0-Not True, 1-Somewhat or Sometimes True, 2-Very True or Often True). The ASR was born from a tradition of assessing hierarchically organized dimensions of psychopathology with the Child Behavior Checklist (Achenbach, 2009). As such, the ASR is one of the few, if not only, adult measures that can be scored on narrowband syndrome scales, broadband internalizing and externalizing scales, and a general psychopathology scale (Achenbach & Rescorla, 2003). The hierarchical structure of psychopathology as measured by the Achenbach System of Empirically Based Assessments has been replicated in clinical and normative samples across the life-span (e.g., ages 1½ -90+) in over 57 societies (Achenbach et al., 2017; Achenbach, Ivanova, Rescorla, Turner, & Althoff, 2016; Ivanova et al., 2015). The ASR's design and extensive prior use make it ideal for assessing the bifactor structure of psychopathology in adults. The 99 items validated for internalizing, externalizing, and cognitive subscales were used in the CFA models (see below).

Response Bias Scale. A scale of heterogeneous items was developed following Weijters, Schillewaert, and Geuens' (2008) guidelines. Specifically, items were randomly selected from 25 random subscales from Bruner's (2013) Marketing Scales Handbook, which includes open-source marketing, personality, and social attitude questionnaires. The items covered a range of domains, from consumerism to collectivism, and had a mean inter-item correlation of .05 (range = -.12–.27; Cronbach's α = .44). The heterogeneity among items was important to ensure that response patterns reflected response biases rather than content (Baumgartner & Steenkamp, 2001). Items were removed if they overlapped substantially in content with another item, or with specific symptoms or broad vulnerabilities to psychopathology (e.g., low self-esteem, neuroticism). All items were adapted to a 7-point response scale (1-Strongly Disagree to 7-Strongly Agree) as this was the most common scale used. Eleven items were excluded post data collection due to their heavily skewed distributions (*k* = 9) or potential overlap with symptoms (*k* = 2). Table 4.2 lists the final list of 14 items.

Table 4.2

Items Randomly Sampled From Bruner's (2013) Marketing Scales Handbook and Used in the Response Bias Scale

Item	Author(s) and Scale/Subscale
1. Individuals should sacrifice self-interest for the group.	Chan, Yim, and Lam (2010) *Individualism/Collectivism*
2. Fitness is a virtue.	Hung and Labroo (2011) *Health Motivation*
3. I would feel secure sending sensitive information over the web	John, Acquisti, and Loewenstein (2011) *Attitude Toward the Website (Security)*
4. I am attracted to rare objects.	Lynn and Harris (1997) *Need for Unique Products*
5. I prefer specific instructions to broad guidelines.	Sharma (2009) *Personal Cultural Orientation (Ambiguity Intolerance)*
6. Traditional values are important for me.	Sharma (2009) *Personal Cultural Orientation (Tradition)*
7. My shopping decisions are influenced by reviews that I read online.	Khare, Labrecque, and Asare (2011) *Attitude Toward Word-of-Mouth (Online)*
8. I often look at my life in philosophical ways	Trapnell and Campbell (1999) *Reflection*
9. People can do things differently, but the important parts of who they are can't really be changed.	Levy, Stroessner, and Dweck (1998) *Implicit Person Theory*

10. I consider myself a creative person.	Hoffman, Kopalle, and Novak (2010) *Creativity*
11. I find I have fewer problems than other people in making technology work for me[a]	Parasuraman (2000) *Technology Readiness Index 2.0: Innovativeness*
12. Products don't seem to hold much value when they are purchased regularly by everyone.	Tian, Bearden, and Hunter (2001) *Need for Uniqueness (Consumer's)*
13. I am dissatisfied with how frequently (or infrequently) my friends want to spend money.	Rick, Small, and Finkel (2011) *Financial Harmony*
14. The salaries of executives should be cut if their firms lay off U.K. workers in order to send jobs overseas.	Thelen, Yoo, and Magnini (2010) *Attitude Toward Offshore Services (Free-Trade Resentment)*

Note. Items were randomly ordered for each participant. Items are not intended to form a reproducible scale but demonstrate the process of randomly sampling heterogeneous items to assess response biases (Weijters et al., 2008).

[a] This item is from the Technology Readiness Index 2.0 which is copyrighted by A. Parasuraman and Rockbridge Associates, Inc., 2014. This scale may be duplicated only with written permission from the authors, which was granted by A. Parasuraman on 10/05/19.

Items were randomly grouped into three parcels and scored for agreement bias, disagreement bias, extreme responding, and mid-point responding (the latter two biases were included for completeness). The first two parcels contained five items while the third parcel contained four items. The frequency of item responses across all items within a parcel was scored using the following formulae (from Weijters et al., 2008):

$$Agreement = [f_5 \times 1 + f_6 \times 2 + f_7 \times 3]/k,$$

$$Disagreement = [f_1 \times 3 + f_2 \times 2 + f_3 \times 1]/k,$$

$$Extreme\ responding = [f_1 + f_7]/k,$$

$$Midpoint\ responding = [f_4]/k,$$

where f_j is the frequency of response category j (j = 1, ..., 7) over k items within a parcel. Each parcel was scored for agreement bias, disagreement bias, extreme responding, and mid-point responding, producing three indicators for each response bias that were used to estimate the response bias factors.

Big Five Inventory (BFI-44; John, Donahue, & Kentle, 1991). The BFI is a 44-item measure of five core personality traits: extraversion (vs. introversion), neuroticism (vs. emotional stability), conscientiousness (vs. lack of direction), agreeableness (vs. antagonism), and openness to experiences (vs. closedness). Items are rated on a 5-point scale (1-Disagree Strongly to 5-Agree strongly). The BFI-44 shows good internal consistency, test-retest reliability, and convergent and discriminant validity (John, Naumann, & Soto, 2008; Rammstedt & John, 2007; Soto & John, 2009). Items from the neuroticism subscale formed a neuroticism factor that was used to test the inter-scale effects of response biases.

4.2.3 Procedure

Participants completed the ASR, BFI-44, and Response Bias Scale online via Qualtrics survey builder (Qualtrics, Provo, UT). They were encouraged to complete the study in one sitting (with breaks if necessary) and took 27 minutes on average (SD = 14, range = 8-293). Participants were informed beforehand about the potential concern or unease caused by reflecting on troubling experiences and provided sample questions to help guide their decision to participate. They could skip questions, entire questionnaires, or withdraw from the study completely without giving a reason and without affecting their monetary compensation. An 'opt-out' button and links to mental health resources were placed on every webpage. Those who reported feeling concerned about their mental health after completing the study (6%) were followed up to ensure their safety and awareness of professional help. Published scales were presented in the original item order and response format; items on the custom response bias scale were presented in a random order. The response bias scale was completed first to determine the influence of responses biases on the psychopathology ratings, followed by the ASR and BFI-44. Ethical approval was obtained from the University College London Research Ethics Committee.

4.2.4 Data Quality Checks

Missing Data. There were a small proportion of missing responses on the ASR and BFI-44 (.01%) and covariates (.10%). The missing data patterns appeared non-systematic[12]: no psychopathology or personality item showed more than two

[12] Parametric tests of missing completely at random, such as Little's (1988) MCAR test, are not available for ordinal or nominal outcomes.

missing responses (.2%). As for covariates, the income variable showed the largest amount of missing responses (10 responses or .8%), which is still minor in absolute terms. Furthermore, missing responses were not disproportionately associated with any given participant. Of the 25 participants who showed missing responses, no participant showed more than two missing responses (0.9%). Missing data were therefore handled using item-based single mean imputation (i.e. using the mean response on a given item to replace missing values), which is as effective as multiple imputation when missing data rates are roughly 5% or less (Shrive, Stuart, Quan, & Ghali, 2006).

Response Distributions. On average, the first response option on the ASR ('Not True') was used 64% of the time (*SD* = .20, range = .15-.98), the second response option ('Somewhat or Sometimes True') was used 27% of the time (*SD* = .27, range = .02-.54), and the third response option ('Very True or Often True') was used 9% of the time (*SD* = .08, range = .003-.37). Response options with infrequent use (e.g., <5%; Achenbach & Rescorla, 2003) are often collapsed with adjacent categories to reduce standard errors, semantic redundancy, and 'disordered' categories, where response options no longer share a linear relationship with the underlying trait (Adams, Wu, & Wilson, 2012). Nonetheless, collapsing adjacent categories is not without criticism and can result in a loss of information. Furthermore, disordered categories can still differentiate participants on the underlying trait (García-Pérez, 2017; Manor, Matthews, & Power 2000; Wetzel & Carstensen, 2014). Therefore, response options were not collapsed, and the robust maximum likelihood estimator was used to adjust the standard errors for skewness. There were few differences between the estimated and observed response distributions (*M* = .004, *SD* = .002, range = -.01–.01).

Residual Correlation Matrix. The residual correlation matrix included 4,851 unique polychoric correlations between ASR items. No model substantially under-estimated (i.e. positive residual) or over-estimated (i.e. negative residual) the item correlations. On average, positive and negative residuals fell below the standard cut-off of .20 (Christensen, Makransky, & Horton, 2017) and stricter cut-off of .10 (Goodboy & Kline, 2017; see Table 4.3).

Less than 1% of residuals were 'potentially problematic', i.e. falling above or below an absolute residual value of .27 (the average residual +/- .2), and less than 0.1% of residuals were 'problematic', i.e. falling above or below an absolute residual value of .37 (the average residual +/- .3; Pallant & Tennant, 2007). Residuals in the single factor and correlated factors model were clustered among externalizing items, while residuals in the bifactor models were diffuse.

Table 4.3

Summary of Residual Correlations for Each Within-Person CFA Model

Model	Positive Res	Negative Res	*M* +/- 0.20	*M* +/- 0.30
Single factor	.07 (.07)	-.07 (.05)	66 (1.2%)	21 (0.4%)
Correlated factors	.07 (.07)	-.07 (.05)	36 (0.7%)	12 (0.2%)
Bifactor (uncorrelated)	.06 (.05)	-.06 (.05)	8 (0.2%)	1 (.02%)
Bifactor (x-loadings)	.05 (.05)	-.05 (.04)	8 (0.2%)	1 (.02%)

Note. *M* = Mean; Res = Residual. Mean and standard deviations (parenthesis) are provided for the average positive and negative residual correlations. Counts and percentages (parenthesis) are provided for the number of residuals falling above or below the mean residual +/- .20 or .30.

Completion speed. There is, to my knowledge, no formal guidance on how to identify and treat respondents that complete online studies exceptionally quickly or slowly. One can follow the conventions of experimental studies and identify outliers based on study completion times that are either 2, 2.5, or 3 standard deviations above or below the mean completion time. Alternatively, one can use the median absolute deviation (MAD) as it is less sensitive to outliers (Leys, Ley, Klein, Bernard, & Licata, 2013).

The median completion time in the current study was 25 minutes, with a MAD of 8 minutes (range = 8-293 mins; note, the next slowest completion time was 112 mins). 'Fast' completers could not be identified using cut-offs derived from the MAD because only one respondent fell 2 MADs below the median (100 participants [8%] fell 2 MADs above). Therefore, the sample was stratified into fast study completers (completion times ≤ 25 mins) and slow completers (completion times > 25 mins) using a median split.

Fast and slow study completers were compared for their factor loadings and item thresholds using multi-group measurement invariance testing. There was no major improvement in fit between the configural and metric models ($\chi^2(198) = 13.68$, $p > .05$), or between the metric and scalar models ($\chi^2(198) = 171.15$, $p > .05$). Therefore, the fast and slow completers showed scalar invariance in terms of their factor loadings and item thresholds. Fast and slow study completers also showed similar model-based reliability estimates for the general and specific factors (see Appendix B), which has also been reported in personality studies (Harms, Jackel, & Montag, 2017; Montag & Reuter, 2008).

4.2.5 Statistical Analysis

The analysis was performed in three parts: 1) clarifying the optimal measurement model for the ASR and response bias scale; 2) identifying the contribution of response biases to the general and specific psychopathology factors, and 3) identifying the impact of response bias factors on the relationships between specific factors, and relationship between the *p* factor and neuroticism. I will outline the analytic strategy for each part in turn.

Part 1: Factor Structure of the ASR and Response Bias Scale

ASR. A bifactor model, with a general *p* factor and specific internalizing, externalizing, and cognitive factors, was compared to a single-factor model that included a general factor with loadings from all symptoms, and a correlated factors model with three correlated broadband factors: internalizing, externalizing, and cognitive problems. In the bifactor model, the associations between *p* and the specific factors were constrained to zero, as well as the associations between specific factors. A bifactor model with correlated specific factors was also estimated. It should be noted that subscale-level factors (e.g., anxious-depressed, withdrawn, aggressive, rule-breaking behaviour, etc.) generally fit better than broadband factors (e.g., internalizing and externalizing), but the more theory-driven and parsimonious solution was preferred. All model solutions were standardized (e.g., the first indicator of each factor was freely estimated and factor variances were set to 1).

Models were estimated using the robust maximum likelihood estimator (MLR) which adjusts for the biases in standard errors associated with non-normal indicators (Yang-Wallentin, Jöreskog, & Luo, 2010). While there are estimators specifically designed for non-normal indicators (e.g., weighted-least squares), MLR was used as the model would later include continuous estimators (e.g., response bias indicators), and the relationship between factors tends to be over-estimated using weight-least squares (Beauducel & Herzberg, 2006; Li, 2016).

Models were compared using the Akaike information criteria (AIC), Bayesian Information criteria (BIC), and sample-size adjusted Bayesian information criteria (aBIC). A difference of 2 (AIC/BIC/aBIC) between models was considered negligible; a difference of 2-7 (AIC) or 2-6 (BIC/aBIC) suggested some evidence favouring the competing model; a difference of 7-10 (AIC) or 6-10 (BIC/aBIC) suggested strong evidence favouring the competing model, and a difference greater than 10 (AIC/BIC/aBIC) suggested very strong evidence favouring the competing model (Fabozzi, Focardi, Rachev, & Arshanapalli, 2014).

The Vuong closeness test for nested and non-nested models was used to formally compare the difference in BIC values between models. The Vuong test includes a likelihood ratio adjusted for the number of parameters estimated in each model (Vuong, 1989). The adjusted likelihood ratio is equivalent to the difference between BIC values (Merkle, You, & Preacher, 2016). A test statistic with a standard normal distribution can be derived from the following equation (simplified from Merkle et al., 2016):

$$z = n^{-0.5}\left[((k-q)\log n) - 2\log\frac{L_A}{L_B}\right],$$

where k is the number of parameters in model A, q is the number of parameters in model B, and L_A and L_B are the likelihood values for model A and B, respectively. The resultant statistic tests the null hypothesis that the BIC values for model A and B are equal (i.e. $\mathrm{H_o} = \mathrm{BIC}_A = \mathrm{BIC}_B$; $\mathrm{H_1} = \mathrm{BIC}_A \neq \mathrm{BIC}_B$).

Models were re-estimated using the means and variances-adjusted weighted least squares estimator (WLSMV) to evaluate their goodness of fit. Acceptable and excellent fit, respectively, were defined by Comparative Fit Index (CFI) values ≥ .90 and ≥ .95, Tucker-Lewis Index (TLI) values ≥ .90 and ≥ .95, and Root Mean Square Error of Approximation (RMSEA) values ≤ .08 and ≤ .06 (Hu & Bentler, 1999).

Bifactor models might fit better than competing models because they better accommodate noise in the data (i.e. overfitting; Greene et al., 2019). Models were assessed for overfitting with double cross-validation, whereby model parameters in a sample are tested in an independent sample and vice versa (Cudeck & Browne, 1983). It is often impractical to recruit another sample, so the current sample was randomly split into a calibration group and a test group. Parameters for the bifactor model (both standard and revised), correlated factors model, and single factor model were freely estimated in the calibration group and used to fix the parameters in the test group. Substantial differences between the calibration and test models, determined by the difference in information criteria (see above for cut-offs), suggest that the model parameters are sensitive to noise within each sub-sample. The process is then repeated, with participants who previously served as the calibration group now used as the test group and vice versa.

Model-based reliability estimates, including omega (ω), omega hierarchical (ω_H), omega hierarchical-subscale (ω_{Hs}), explained common variance (ECV), explained common variance-subscale (ECV_s), construct reliability (*H*), and factor determinacy (*FD*) were calculated using the MLR factor loadings and Bifactor Indices Calculator (Dueber, 2017). Furthermore, the mean parameter change (MPC) for a given broadband dimension (e.g., internalizing, externalizing, thought problems) was calculated by subtracting the standardized factor loading of the relevant specific factor from the standardized factor loading of the respective factor in the correlated factors model, and averaging the values. Positive mean parameter change values suggest that factor loadings decreased on average for a broadband domain from the correlated factors model to the bifactor models, and hence included common variance accounted for by the general factor, whereas negative values suggest that factor loadings increased on average and were less affected by the common variance. The standard deviation of the parameter change (SDPC) was calculated by taking the standard deviation of the difference scores between factor loadings in the correlated factor and bifactor models.

Response Bias Scale. A standardized two-factor model was estimated with the three agreement bias (AB) indicators loading on an AB factor, and the three disagreement bias (DB) indicators loading on a DB factor. Factors were free to covary, as well as indicators scored using the same parcel of items. The two-factor model was estimated using MLR and compared to a single-factor model upon which all response bias indicators loaded using the information criteria described above. Models were also estimated with the standard maximum likelihood estimator to evaluate their goodness of fit with the indices described above (with the caveat that the DB indicators showed positive skews). The process was repeated with the extreme responding (ER) and mid-point responding (MR) indicators (e.g., a two-factor model with ER and MR factors was compared to a single factor with loadings from ER and MR indicators).

Part 2. Intra-scale Effect of Response Biases

Structural equation models were used to estimate the contribution of response biases to the general (*p*) and specific psychopathology factors. One could simply regress the *p* factor and specific factors onto the response bias factors, but this would create a path between the general and specific factors which violates the bifactor model's orthogonality constraint and can bias parameter estimates or result in model non-convergence (Koch, Holtmann, Bohn, & Eid, 2018). Therefore, response biases were first residualized for the influence of the psychopathology factors that were not of interest. For example, to estimate the contribution of response biases to the *p* factor, the response biases factors were first regressed onto the specific psychopathology factors. The residual component of the response bias factors (i.e. the component that that is free from the influence of specific factors) was then saved as a new latent variable and used to predict variation in the *p* factor. If the influence of response biases on specific psychopathology factors was of interest, then the response bias factors were first residualized for the *p* factor, and the latent residuals were used to predict the specific factors.

In all models, the residuals of the response bias factors residualized for the general or specific psychopathology factors were correlated, as well as the residuals of the response bias indicators created from the same item parcels. The contribution of residualized response bias factors to the *p* factor and specific factors was assessed by the strength of the standardized regression coefficients and proportion of variance explained (R^2).

Part 3. Inter-scale Effect of Response Biases

A structural equation model was used to estimate the relationship between the *p* factor and neuroticism after controlling for response biases. Specifically, the *p* factor and a neuroticism factor (estimated using BFI-44 neuroticism items) were estimated within the same model and regressed onto response bias factors, which were first residualized for the specific factors. The strength of the correlation between the *p* factor and neuroticism factor before correcting for response biases was compared to their correlation after correcting for response biases using a Fisher transformation, which first involves converting the *r* coefficients to *z* scores and then comparing them using the normal distribution (Steiger, 1980). Changes in the relationship between specific factors and neuroticism before and after correcting for response biases was also compared using the same method for completeness.

The contribution of response bias factors to the specific factor correlations was also assessed by comparing the strength of the correlation coefficients corrected and uncorrected for response biases using a Fisher transformation. All analyses were ran in Mplus 8.0 (Muthén & Muthén, 2017).

4.3 Results

4.3.1 Factor Structure of the ASR and Response Bias Scale

ASR. Both the single factor model and correlated factor model (with internalizing, externalizing, and cognitive factors) showed an acceptable absolute fit (RMSEA), but near-acceptable comparative fit (CFI, TLI; see Table 4.4). The correlated factors model explained substantially more information than the single factor model (ΔAIC = 2079; ΔBIC = 2064; ΔaBIC = 2073; z = 42.78, p < .001). Both single factor and correlated factor models showed healthy loadings, and the internalizing, externalizing, and cognitive factors were strongly and positively correlated, indicating the presence of a general factor (see Table 4.5).

Estimating a general p factor as well as orthogonal specific internalizing, externalizing, and cognitive factors fit the data well (see Table 4.4) and improved substantially on the correlated factors model (ΔAIC = 3032; ΔBIC = 2108; ΔaBIC = 2414; z = 59.80, p < .001) and single-factor model (ΔAIC = 5111; ΔBIC = 4172; ΔaBIC = 4486; z = 41.40, p < .001; z = 102.58, p < .001). Modification indices from the WLSMV model suggested that the correlations among specific factors could be freed (expected parameter changes > .4; Stevens, 1992). Furthermore, somatic items loaded negatively on the internalizing factor (see Table 4.5). Therefore, the bifactor model was revised by freeing the correlations among specific factors and estimating a separate factor for somatic problems. The revised bifactor model showed a near-excellent fit (see Table 4.4) and made a substantial improvement on the original bifactor model (ΔAIC = 814; ΔBIC = 1219; ΔaBIC = 1238; $z = 16.61$, $p < .001$). The specific factors showed moderate positive associations (see Table 4.5). Cross-validation tests demonstrated that all models differed substantially between the calibration and test groups and were therefore sensitive to sample-specific characteristics (see Table 4.6).

The common variance in the revised bifactor model was multidimensional but favoured the *p* factor over the specific factors (62% vs. 38%). By contrast, the variance in raw total scores was largely explained by the *p* factor (ω_H = 83%). Most items showed moderate-to-strong loadings on the *p* factor except for rule-breaking and intrusive problems, which were weak and not significant (see Table 4.5). The specific internalizing factor showed low reliability because several items that loaded onto the internalizing factor in the correlated factors model now loaded almost exclusively onto the *p* factor (mean parameter change [MPC] = .47, standard deviation of the parameter change [SDPC] = 0.23).

Somatic items (MPC = .25, SDPC = .15) and cognitive items (MPC = .26, SDPC = 0.18) also decreased in specific factor loading strength compared to the correlated factors model, albeit to a lesser extent than internalizing items. Externalizing items loadings showed the smallest change between correlated factor and bifactor models (MPC = .06; SDPC = 0.26), but changes in different externalizing domains cancelled each other out. For example, aggressive problems showed a small bias towards the *p* factor (MPC = 0.23, SDPC = 0.24), while intrusive items were biased towards the specific externalizing factor (MPC = -0.23, SDPC = 0.07). Rule-breaking items loaded strongly on both *p* and specific externalizing factors (MPC = 0.01, SDPC = 0.21; see Table 4.5).

Table 4.4

Model Fit Values for the Confirmatory Factor Analysis Models of the Adult Self-Report (1-5) and Response Bias Scale (6-7)

	Fit Estimate							
Model	χ^2	*df*	CFI	TLI	RMSEA	AIC	BIC	aBIC
1. Single Factor	16,074	4,752	.88	.86	.05	152,213	153,725	152,781
2. Correlated Factors	14,066	4,749	.89	.89	.04	150,134	151,661	150,708
3. Bifactor	11,105	4,653	.92	.92	.03	147,102	149,553	148,295
4. Bifactor (revised)	9,781	4,647	.94	.94	.03	146,288	148,334	147,057
5. Bifactor (attention on externalizing)	9,907	4,647	.94	.94	.03	146,445	148,491	147,214
6. Response bias single factor	199	6	.85	.62	.16	7,947	8,054	7,987
7. Response bias two factor	42	5	.97	.91	.08	7,797	7,909	7,840

Note. χ^2 = chi-square statistic; aBIC = sample size adjusted Bayesian information criterion; CFI = Comparative Fit Index; *df* = degrees of freedom; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation.

Table 4.5

Standardized Factor Loadings for the Single Factor, Correlated Factor, and Bifactor Models of the Adult Self-Report

		Correlated Factors			Bifactor (Standard)				Bifactor (Revised)				
ASR Item	1-Fac	INT	EXT	COG	*p*	INT	EXT	COG	*p*	INT	SOM	EXT	COG
Anxious/Depressed													
12. Lonely	.64	.67			.67	.20			.66	.24			
13. Confused	.76	.75			.75	-.11			.72	.13			
14. Cries	.66	.66			.68	-.15			.71	-.11			
22. Worries about future	.67	.70			.69	.10			.72	.05[a]			
31. Fears thinking/doing something bad	.70	.68			.68	.09			.65	.21			
33. Unloved	.71	.74			.73	.23			.69	.40			
34. Others out to get him/her	.69	.69			.69	.05[a]			.65	.29			
35. Worthless	.83	.86			.85	.19			.84	.22			
45. Nervous	.77	.80			.80	-.01[a]			.85	-.08[a]			
47. Lacks self-confidence	.73	.77			.77	.17			.80	.07[a]			
50. Fearful	.79	.82			.82	.02[a]			.88	-.08[a]			
52. Too guilty	.71	.71			.72	-.05[a]			.72	.03[a]			
71. Self-conscious	.67	.71			.71	.14			.75	.03[a]			
91. Thinks about suicide	.74	.73			.72	.16			.69	.28			
103. Sad	.85	.87			.87	.05[a]			.86	.16			

107. Can't succeed	.79	.81	.81	.17	.80	.19	
112. Worries	.73	.76	.77	-.03[a]	.86	-.20	
113. Worries about relations with opposite sex	.54	.54	.54	.27	.51	.30	
Withdrawn							
25. Doesn't get along with others	.57	.55	.54	.33	.47	.58	
30. Poor relations with opposite sex	.53	.55	.54	.32	.51	.40	
42. Would rather be alone	.47	.48	.48	.20	.45	.33	
48. Not liked	.68	.68	.67	.30	.62	.52	
60. Enjoys little	.76	.77	.77	.20	.74	.28	
65. Won't talk	.64	.64	.63	.11	.58	.34	
67. No friends	.65	.65	.65	.35	.60	.51	
69. Secretive	.51	.50	.49	.19	.45	.37	
111. Withdrawn	.57	.58	.58	.24	.54	.40	
Somatic Complaints							
51. Dizzy	.64	.65	.68	-.36	.64		.42
54. Feels tired	.70	.69	.70	-.16	.68		.26
56a. Aches	.53	.53	.55	-.37	.50		.51
56b. Headaches	.46	.48	.51	-.46	.47		.51
56c. Nausea	.62	.64	.67	-.51	.62		.57
56d. Eye problems	.45	.44	.45	-.32	.39		.47
56e. Skin problems	.37	.36	.36	-.12	.32		.25
56f. Stomach aches	.56	.57	.59	-.40	.54		.52
56g. Vomits	.62	.61	.61	-.36	.55		.46

56h. Heart pounds	.67	.67		.69	-.30		.66	.39	
56i. Numbness	.61	.60		.62	-.41		.56	.57	
100. Sleep problems	.50	.50		.51	-.22		.48	.31	
Aggressive Behavior									
3. Argues	.38		.55	.33		.44	.31		.45
5. Blames others	.52		.54	.49		.25	.48		.28
16. Mean	.40		.57	.31		.56	.27		.60
28. Gets along badly with family	.47		.48	.44		.21	.41		.28
37. Fights	.60		.70	.48		.53	.39		.62
55. Elation-depression	.76		.77	.74		.25	.72		.33
57. Attacks	.58		.68	.41		.56	.30		**.65**
68. Screams	.52		.65	.46		.43	.42		.47
81. Behavior changes	.64		.75	.60		.40	.58		.46
86. Stubborn	.62		.68	.59		.32	.56		.39
87. Mood changes	.76		.81	.73		.31	.73		.37
95. Temper	.48		.63	.44		.43	.42		.46
97. Threatens	.59		.70	.44		.56	.35		**.65**
116. Upset	.71		.61	.73		.04[a]	.77		.08
118. Impatient	.53		.63	.50		.38	.49		.41
Rule-Breaking Behavior									
6. Uses drugs	.31		.38	.22		.36	.19		.42
20. Damages own things	.62		.68	.52		.45	.47		.52
23. Breaks rules	.28		.44	.19		.55	.13		.60
26. Lacks guilt	.09		.26	.01[a]		.46	-.05[a]		.47
39. Bad companions	.51		.65	.35		.64	.28		**.69**

41. Impulsive	.44	.62		.35	.56		.30	.63	
43. Lies, cheats	.45	.58		.37	.51		.31	.57	
76. Irresponsible	.62	.72		.52	.54		.45	.63	
82. Steals	.46	.62		.32	.61		.26	**.66**	
90. Gets drunk	.23	.34		.16	.41		.13	.43	
92. Trouble with the law	.47	.60		.32	.58		.24	**.65**	
114. Fails to pay debts	.48	.51		.43	.26		.41	.34	
117. Can't manage money	.45	.52		.40	.32		.37	.39	
122. Can't keep a job	.65	.55		.62	.09[a]		.58	.22	
Intrusive									
7. Brags	.14	.35		.05[a]	.60		.03[a]	.56	
19. Demands attention	.33	.52		.22	.61		.22	.59	
74. Shows off	.24	.43		.13	.63		.09	**.64**	
93. Talks too much	.18	.32		.11	.45		.10	.44	
94. Teases	.20	.36		.11	.54		.07[a]	.54	
104. Loud	.23	.46		.13	.64		.10	.63	
Thought Problems									
9. Can't get mind off thoughts	.64		.63	.64		.07[a]	.65		.14
18. Harms self	.71		.66	.69		.05[a]	.67		.17
36. Gets hurt	.52		.55	.48		.30	.44		.37
40. Hears things	.62		.64	.54		.36	.46		.50
46. Twitches	.61		.62	.59		.18	.55		.29
63. Prefers older people	.34		.33	.34		.02	.33		.10
66. Repeats acts	.44		.47	.40		.20[a]	.38		.28
70. Sees things	.65		.66	.57		.29	.48		.48

84. Strange behaviour	.50			.55	.43			.31	.37				.45
85. Strange ideas	.58			.60	.53			.22	.50				.35
Attention Problems													
1. Forgetful	.45			.52	.41			.44	.37				.45
8. Can't concentrate	.62			.68	.57			.39	.54				.43
11. Dependent	.55			.57	.53			.18	.54				.19
17. Daydreams	.42			.45	.38			.26	.37				.29
53. Can't plan	.71			.72	.70			.17	.69				.24
59. Fails to finish	.65			.71	.59			.42	.56				.49
61. Poor work performance	.73			.76	.69			.28	.64				.39
64. Can't prioritize	.66			.74	.60			.47	.56				.51
78. Trouble with decisions	.68			.71	.67			.24	.67				.25
101. Avoids work	.54			.56	.48			.23	.44				.35
102. Lacks energy	.72			.71	.73			.11	.71				.20
105. Disorganized	.54			.64	.45			.69	.40				.66
108. Loses things	.52			.60	.46			.53	.43				.52
119. Poor at details	.53			.59	.47			.42	.43				.48
121. Tends to be late	.40			.47	.33			.41	.30				.41
Mean	.56	.65	.56	.60	.53	-.01	.44	.29	.50	.22	.44	.49	.36
Standard Deviation	.16	.12	.14	.10	.19	.25	.16	.16	.21	.20	.11	.15	.14
ECV/ECV$_s$	–	–	–	–	.71	.06	.18	.06	.62	.05	.05	.20	.08
ω/ω_s	–	–	–	–	.98	.97	.95	.94	.98	.97	.92	.95	.94
ω_H/ω_{Hs}	–	–	–	–	.89	.00	.55	.22	.83	.09	.37	.64	.32
Relative Omega	–	–	–	–	.90	.00	.58	.23	.85	.09	.40	.67	.34

H	–	–	–	–	.98	.74	.92	.78	.98	.74	.76	.93	.83
FD	–	–	–	–	.99	.92	.96	.91	.99	.90	.91	.97	.92
Inter-factor Correlations		INT	EXT	COG						INT	SOM	EXT	COG
	INT	–			–	–	–	–	INT	–			
	EXT	.75	–		–	–	–	–	SOM	.05[a]	–		
	COG	.89	.81	–	–	–	–	–	EXT	.39	.26	–	
					–	–	–	–	COG	.44	.33	.65	–

Note. 1-Fac = Single factor model; COG = Cognitive; ECV/ECV$_s$ = Explained Common Variance/Explained Common Variance-Subsale; Ext = Externalizing; FD = Factor Determinacy; Int = Internalizing; ω/ω_s = Omega/Omega-subsale; Som = Somatic Problems. ω_H/ω_{Hs} = Omega hierarchical/Omega hierarchical-subscale.

[a] Estimates that were not significant (p > .05).

Table 4.6

Differences in Information Criteria Between the Calibration and Test Groups for Each CFA Model of the ASR

	Order A			Order B		
Model	ΔAIC	ΔBIC	ΔaBIC	ΔAIC	ΔBIC	ΔaBIC
Single Factor	1873	581	1515	1852	3154	2214
Correlated Factors	1955	649	1592	1998	3313	2364
Bifactor	1614	-114	1134	1745	3482	2228
Bifactor (revised)	1611	-143	1124	1437	3200	1927

Note. aBIC = sample size adjusted Bayesian information criterion. Order A and B reflect the sequence that each half of the sample was allocated as the calibration or test group. Negative values indicate that the calibration sample showed a lower (better) fit compared to the test sample.

In a sensitivity analysis, the fit of the revised bifactor model was compared to a bifactor model with attention items loading on the specific externalizing factor rather than the specific cognitive factor. While attention items loaded well on the specific externalizing factor, the model fitted slightly worse than when attention items loaded on the cognitive factor (ΔAIC = -157; ΔBIC = -157; ΔaBIC = -157; see Table 4.4 for model fit and Appendix B for loadings).

Response Bias Scale. A single factor with loadings from agreement bias (AB) indicators and disagreement bias (DB) indicators showed a poor fit (see Table 4.4 for model fit and Table 4.7 for factor loadings). By contrast, a two-factor model with separate AB and DB factors fitted well and better than the single factor model (ΔAIC = 150; ΔBIC = 145; ΔaBIC = 147). Items loaded well on each factor (see Table 2.5). The AB and DB factors were weakly correlated (*r* = -.07, *p* = .321, 95% CI [-.21, .07]), despite moderate and negative residual correlations among the AB and DB indicators scored from the same item parcels (average *r* = -.47, 95% CI [-.53, -.41]).

The single-factor model for extreme responding (ER) and mid-point responding (MR) indicators also showed a poor fit (see Table 2.3 and Table 2.5), while a two-factor model with separate ER and MR factors showed a superior fit (ΔAIC = 378; ΔBIC = 374; ΔaBIC = 378). Items loaded well on each factor (see Table 2.5). The ER and MR factors were negatively correlated (*r* = -.32, *p* < .001, 95% CI [-.39, -.25]), as were the residuals among the ER and MR indicators scored from the same item parcels (average *r* = -.16, average 95% CI [-.22, -.10]).

Table 4.7

Standardized Factor Loadings for the Single Factor and Two-Factor CFA Models of the Response Bias Scale

		Two-factor	
Response Bias Indicator	Single Factor	AB	DB
Agreement and Disagreement Biases			
AB Indicator 1 (Items 9, 2, 4, 5, 11)	.60	.68	
AB Indicator 2 (Items 13, 6, 10, 8, 1)	.42	.53	
AB Indicator 3 (Items 14, 7, 3, 12)	.47	.57	
DB Indicator 1	.23		.63
DB Indicator 2	.33		.45
DB Indicator 3	.22		.44
Extreme and Mid-point Responding		ER	MR
ER Indicator 1	0.68	0.70	
ER Indicator 2	0.59	0.63	
ER Indicator 3	0.59	0.63	
MR Indicator 1	-0.38		0.65
MR Indicator 2	-0.34		0.61
MR Indicator 3	-0.37		0.57

Note. AB = Agreement Bias; DB = Disagreement Bias. Factor loadings were significant unless marked otherwise. Items comprising each response bias indicator were the same across response biases.

4.3.2 Intra-scale Effect of Response Biases on the General (*p*) and Specific Psychopathology Factors

***p* factor.** In a model with AB and DB factors predicting the *p* factor after being residualized for the specific psychopathology factors, the AB factor weakly but significantly predicted *p* (*B* = 0.15, *p* = .001, 95% CI [0.06, 0.23]), as did the DB factor (*B* = 0.16, *p* = .002, 95% CI [0.06, 0.26]). That is, a tendency to indiscriminately agree and disagree with items was slightly related to higher *p* factor scores. The AB and DB factors explained 4% of the variance in *p*, and were weakly correlated (*r* = -0.07, *p* = .343, 95% CI [-0.22, 0.08]).

A sensitivity analysis was run treating the DB indicators as over-dispersed count variables to adjust for their positive skew. The DB factor's negative binomial regression weight increased slightly (B = 0.21, p = .01, 95% CI [0.05, 0.37]; R^2 = 4%), as did the AB factor's linear regression weight (B = 0.18, p < .001, 95% CI [0.10, 0.26]; R^2 = 3%).

In an exploratory analysis with ER and MR factors predicting the p factor after being residualized for the specific psychopathology factors, the ER factor weakly but significantly predicted p (B = 0.15, p < .001, 95% CI [0.06, 0.23]), as did the MR factor (B = -0.09, p = .023, 95% CI [-0.16, -0.01]). In other words, a tendency to use the highest and lowest response categories was weakly related to higher p scores, while a tendency to use the middle response option was weakly associated with lower p scores. The ER and MR factors explained 4% of the variance in p, and were moderately correlated (r = -0.32, p < .001, 95% CI [-0.39, -0.25]).

In a sensitivity analysis treating the ER and MR indicators as over-dispersed count variables, the ER factor's negative binomial regression weight increased slightly (B = 0.20, p = .02, 95% CI [0.04, 0.36], R^2 = 4%), while the MR factor's negative binomial regression weight was longer reached significance (B = -0.07, p = .424, 95% CI [-0.23, 0.10], R^2 = .5%).

When the AB, DB, ER and MR response bias factors were included as predictors in the same model, the AB factor (B = 0.18, p = .292, 95% CI [-0.16, 0.53]) and DB factor (B = -0.18, p = .132, 95% CI [-0.42, 0.06]) alone predicted the p factor. However, as is evident, the regression coefficients and standard errors were inflated, mostly likely because the AB and DB factors were collinear with the ER factor ($r_{AB,ER}$ = -.61; $r_{DB,ER}$ = .52). Therefore, further models included complementary pairs of response bias factors only (e.g., AB and DB or ER and MR).

Specific Psychopathology Factors. In a model with AB and DB factors predicting the specific factors after being residualized for the *p* factor, the AB factor weakly but significantly predicted the specific externalizing factor (*B* = 0.09, *p* = .049, 95% CI [0.00, 0.18]). Hence, a tendency to indiscriminately agree with items was associated with slightly higher externalizing scores. The DB factor did not significantly predict any specific factor. The AB and DB factor accounted for 0.2%-1% of the variance in the specific factors (see Table 4.8).

In a model with ER and MR factors predicting the *p* factor after being residualized for the specific psychopathology factors, the ER factor weakly but significantly predicted the specific externalizing factor (*B* = 0.09, *p* = .029, 95% CI [.01, .18]). Therefore, a tendency to use the highest and lowest response categories was weakly associated with higher externalizing scores. The MR factor did not significantly predict any specific factor. The ER and MR factors accounted for 0.4%-0.9% of the variance in the specific factors (see Table 4.8).

Table 4.8

Standardized Regression Coefficients for the Response Bias Factors (Residualized for the p Factor) Predicting the Specific Psychopathology Factors

Factor	*B*	*z*	*p*	95% CI	R^2
Agreement Bias					
Internalizing	.05	0.82	.412	-.07, .16	0.4%
Externalizing	**.09**	**1.96**	**.049**	**0, .18**	1%
Somatic	.04	0.72	.471	-.06, .13	0.2%
Cognitive	.08	1.53	.126	-.02, .17	0.7%
Disagreement Bias					
Internalizing	-.03	-0.65	.513	-.13, .07	“
Externalizing	.07	1.42	.155	-.03, .16	“
Somatic	-.02	-0.33	.745	-.13, .09	“
Cognitive	-.03	-0.64	.522	.14, .07	“
Extreme responding					

Internalizing	.06	1.16	.247	-.04, .15	0.4%
Externalizing	**.09**	**2.19**	**.029**	**.01, .18**	0.9%
Somatic	.08	1.59	.111	-.02, .17	0.8%
Cognitive	.07	1.37	.171	-.03, .17	0.4%
Midpoint responding					
Internalizing	.05	0.91	.366	-.05, .15	"
Externalizing	-.01	-0.26	.797	-.10, .08	"
Somatic	.08	1.66	.097	-.01, .17	"
Cognitive	.03	0.68	.497	-.06, .12	"

Note. Significant results are in bold.

4.3.3 Inter-scale Effect of Response Biases on the Relationship Between Specific Factors and Between the *p* Factor and Neuroticism

Table 4.9 shows the correlations between specific factors before and after controlling for response biases. *z* tests comparing the magnitude of the uncorrected coefficients with either the coefficients corrected for AB and DB biases, or ER and MR biases, were not significant (*z*s < |1.36|, *p*s > .05). The average decrease in magnitude from the uncorrected to the corrected coefficients was .02 (*SD* = .02, range = 0-.05).

Table 4.9

Correlation Matrix of the Specific Psychopathology Factors Without Controlling for Response Biases (First Estimate), After Controlling for Agreement and Disagreement Biases (Second Estimate), and After Controlling for Extreme Responding and Midpoint Responding (Third Estimate).

	INT	SOM	EXT	COG
INT	–			
SOM	.05/.03/.02	–		
EXT	.39/.39/.38	.26/.23/.22	–	
COG	.44/.47/.46	.33/.29/.28	.65/.65/.64	–

Note. Cog = Cognitive; Ext = Externalizing; Int = Internalizing; Som = Somatic.

As for the inter-scale effect of response biases on neuroticism, the BFI-44 items loaded strongly onto a neuroticism factor when estimated alongside the bifactor dimensions (average λ = .76, *SD* = .10, range = .67-87). The uncorrected correlation between the neuroticism factor and *p* factor was near-perfect (*r* = .97, *p* < .001, 95% CI [.97, .98]), and remained strong but was significantly smaller when items that overlapped with internalizing symptoms were removed (*r* = .87, *p* < .001, 95% CI [.85, .90]; *z* = 18.76, *p* < .001).

When the AB and DB factors were added to the model and residualized for the specific psychopathology factors, the DB factor weakly but significantly predicted the neuroticism factor (*B* = .15, *p* < .001, 95% CI [.06, .23]), while the AB factor marginally predicted the neuroticism factor (*B* = .07, *p* = .069, 95% CI [-.004, .14]). Hence, a tendency to indiscriminately disagree with items was associated with a slight increase in neuroticism scores. The AB and DB factors explained 2% of the variance in neuroticism. The relationship between the neuroticism factor and *p* factor was similar after controlling for the AB and DB factors (*r* = .98, *p* < .001, 95% CI [.97, .98]; *z* = -.99, *p* = .32).

When the ER and MR factors were added to the model and residualized for the specific psychopathology factors, the ER factor weakly but significantly predicted the neuroticism factor (*B* = .12, *p* < .001, 95% CI [.02, .19]), while the MR factor did not significantly predict the neuroticism factor (*B* = -.04, *p* = .190, 95% CI [-.11, .02]). Therefore, a tendency to use the highest and lowest response options was associated with a slight increase in neuroticism scores. The ER and MR factors explained 2% of the variance in neuroticism. Again, the corrected relationship between the neuroticism factor and p factor was similar overall (r = .98, p < .001, 95% CI [.97, .98]; z = -.49, p = .62).

For completeness, the associations between neuroticism and the specific psychopathology factors were also compared before and after controlling for response biases. The neuroticism factor was strongly correlated with the specific factors, particularly internalizing and externalizing problems (see Table 4.10). As with the p factor, there was little change in the correlation coefficients after controlling for AB and DB factors (zs < |1.61|, ps > .05; see Table 4.10). The model controlling for ER and MR biases did not converge, nor did a model controlling the neuroticism factor the p factor (most likely due to the indirect path with the specific factors).

Table 4.10

Associations Between Neuroticism and Specific Psychopathology Factors Before and After Correcting for Response Biases

Specific Factor	Uncorrected r	Corrected r (AB DB)	Corrected r (ER MR)
Internalizing	.90	.92	No convergence
Externalizing	.92	.92	"
Cognitive	.87	.88	"
Somatic	.70	.73	"

Note. AB = Agreement bias; DB = Disagreement bias; ER = Extreme responding; MR = Mid-point responding.

In a sensitivity analysis, the associations between an extraversion factor (λ = .72, SD = .15) and the specific psychopathology factors were compared before and after controlling for response biases. The extraversion factor was moderately and negatively correlated with the internalizing factor before (r = -.59, p < .001, 95% CI [-.68, -.50]) and after (r = -.59, p < .001, 95% CI [-.68, -.51]) controlling for response biases. The extraversion factor was also weakly but significantly associated with the externalizing factor (uncorrected r = .36, p < .001, 95% CI [.19, .56]; corrected r = .36, p < .001, 95% CI [.16, .51]) and somatic factor (uncorrected r = .16, p = .02, 95% CI [.03, .29]; corrected r = .15, p = .04, 95% CI [.01, .29]). The extraversion factor did not correlate with the cognitive factor before (r = .03, p = .742, 95% CI [-.15, .21]) or after (r = .01, p = .936, 95% CI [-.18, .19]) controlling for response biases.

4.4 Discussion

The bifactor model of psychopathology has attracted much excitement from clinical researchers seeking to reform the current diagnostic system. While there is a growing body of evidence supporting the substantive basis of the *p* factor (see Chapter 3), there are also untested methodological issues that could challenge our understanding of the bifactor dimensions. One such issue is whether the variance explained by the *p* factor and/or specific factors, and the associations between the bifactor dimensions and theoretically-related variables, is driven by response biases, such as indiscriminately agreeing (agreement bias) or disagreeing (disagreement bias) with questionnaire items, or exclusively using the extreme ends (extreme responding) or middle option (mid-point responding) of a response scale.

The current study showed that agreement bias, disagreement bias, extreme responding, and mid-point responding explain a negligible proportion of the variance in the *p* factor and specific internalizing, externalizing, cognitive, and somatic factors. Furthermore, the moderate associations among specific factors, and the strong association between the *p* factor and neuroticism, showed little change after controlling for response biases. I will discuss these findings in turn, starting with the latent structure of psychopathology, followed by the minimal impact of response biases.

4.4.1 Are responses on the Adult Self Report (ASR) best described by a bifactor structure?

The relationships among ASR items were best captured by a revised bifactor model with a *p* factor and correlated specific factors compared to a bifactor model with orthogonal specific factors, a correlated factors model, and a single-factor model. The revised bifactor model showed a multidimensional data structure, with 62% of common (i.e. modellable) variance explained by *p*, and the remaining 38% explained by specific factors. Nonetheless, the majority of reliable variance in raw total scores was explained by *p* (83%). Therefore, both general and specific factors are necessary to account for the latent structure of psychopathology, but the observed data is 'essentially unidimensional', as is commonly found in bifactor studies of psychopathology (see Chapter 3) and psychological traits more generally (Rodriguez, Reise, & Haviland, 2016a).

Freeing the correlations among specific factors substantially improved the information explained by the bifactor model. Others have also reported an improvement in fit after freeing the specific factor correlations (Afzali et al., 2017; Arrindell et al., 2017; Brodbeck et al., 2014; Carragher et al., 2016; Caspi et al., 2014; Hyland et al., 2018; Laceulle et al., 2016; McElroy et al., 2017; Neumann et al. 2016; Patalay et al., 2015; Pettersson et al., 2018; Preti et al., 2018; Urbán et al., 2016; Urbán et al., 2014), but this raises concerns about how to interpret the shared variance beyond the general factor (Markon, 2019; see section 3.5.6). There is no clear answer to this: Not freeing the specific factor correlations may lead to model misspecification and inflate general factor loadings (Reise, 2012). Yet freeing the specific factor correlations implies the presence of unmodelled factors (McDonald, 1999) and could lead to an unidentified model (Markon, 2019).

There is one form of the bifactor model that permits correlated specific factors. In the bifactor S-1 model, one of the specific factors is not estimated, and the items of the omitted specific factor load exclusively onto the general factor (Eid, Krumm, Koch, & Schulze, 2018). Correlations among specific factors reflect shared variance beyond the general factor as identified by the items that exclusively load onto it. Take, for example, the finding that psychosis items sometimes load exclusively onto the *p* factor, rendering a specific thought disorder factor obsolete (see section 2.2.1). In a model where psychosis items load exclusively onto a *p* factor, and a specific thought disorder factor is omitted (e.g., Caspi et al., 2014), the correlation between the remaining specific factors (e.g., internalizing and externalizing) would represent the relationship between internalizing- and externalizing-specific variance whilst controlling for the common variance identified by psychosis items.

The revised bifactor model in the current study could be considered a 'quasi' bifactor S-1 model, since most of the anxiety and depression items loaded almost exclusively onto the *p* factor rather than the internalizing factor. Therefore, the specific factor correlations are admissible and reflect the shared variance between internalizing (free from anxiety and depression), externalizing, cognitive problems, and somatic complaints, independent from the common variance identified by anxiety and depression. It should be noted that the bifactor S-1 model is a re-expression of the correlated factors model (Eid et al., 2018). Therefore, while the bifactor S-1 model showed a better fit than the correlated factors model, this does not mean that the former reflects the correct or 'true' population variance-covariance matrix and the latter does not. It simply implies that reproducing the correlations among symptoms with a broad 'common factor', so to speak, as well as more specific common factors provides the most efficient solution. A model might appear structurally superior but could have the same underlying constraints that are elusive to model comparison tests (Markon, 2019).

One point of difference is that the specific internalizing and externalizing factors were positively correlated in the current study, while prior studies report a negative correlation (Caspi et al., 2014; Haltigan et al., 2018; Laceulle et al., 2016; McElroy et al., 2017; Neumann et al. 2016; Patalay et al., 2015). This might be due to content overlap: Both internalizing and externalizing factors included items related to interpersonal problems, such as 'Doesn't get along with others' and 'Gets along badly with family', respectively. Moreover, some externalizing items overlapped with emotional problems (which are characteristic of internalizing), such as 'mood changes', 'elation-depression', and 'temper'. Alternatively, prior studies may have included a larger proportion of respondents falling within the clinical range. The nature and structure of factors can vary between clinical and non-clinical samples due to the differential responses to symptom items (Ferentinos et al., 2019; Krueger, 1999). More broadly, this finding raises the debate about whether the general and specific factors are invariant across the instrument used, which is unlikely to be the case given differences in the content, wording, sample, and overall measurement context (Gustafsson & Åberg-Bengttson, 2010; see Chapter 3).

Broadly speaking, the current study supports the bifactor model as a viable alternative to current models (e.g., correlated factors model), with some differences with past studies that are likely specific to the ASR.

4.4.2 Are the general and specific psychopathology factors meaningful?

Even if the bifactor model fits the data best, items might not load onto the general and specific factors in plausible ways. In other words, the model might be identified but the factors are not meaningful or could even imply an alternative model (Markon, 2019). It is thus important to evaluate the factor loading matrix to determine whether the nature of the factors aligns with prior studies and expectations about the bifactor structure of psychopathology (which may or may not be synonymous with past findings).

***p* factor.** The *p* factor showed healthy and positive loadings from most items (except for some externalizing items, see below) which partially supports the positive manifold among psychiatric symptoms. Anxious-depressed items loaded most strongly onto *p,* sometimes at the expense of their internalizing factor loadings. Items with the strongest standardized loadings (λ > .8) included 'fearful', 'worries', 'sad', 'nervous', 'worthless', and 'can't succeed'. Prior studies in adolescents and adults have also found that depression and anxiety symptoms or disorders load most strongly onto *p* (Arrindell et al., 2017; Black, Panayiotou, & Humphrey, 2019; Caspi et al., 2014; Lahey et al., 2012; Lahey et al., 2015; Liu et al., 2017; Martel et al., 2017; Miller et al., 2019; Olino et al., 2014; Pezzoli et al., 2017; Preti et al., 2018; Romer et al., 2017; St Clair et al., 2017; Tackett et al., 2013; Urbán et al., 2016; Urbán et al., 2014). The reasons for this are unclear, but it could be that the involuntary and affectively laden nature of cognition associated with depression and anxiety, such as rumination and worry, poses a risk to, or is a common feature of, many forms of psychopathology (see section 2.2). This would also explain why many psychotic and attentional items loaded strongly onto *p*, while rule-breaking and intrusive items loaded weakly onto *p*; the latter describe impulsive and callous behaviors more so than disordered cognition.

In all, the *p* factor showed healthy loadings from most items in a pattern anticipated by past research, but there are clear differences in problems that are most (e.g., depression and anxiety) and least (e.g., antisocial problems) reflective of *p*, which contradicts the claim that the *p* factor reflects a vulnerability factor to any and all forms of psychopathology (Caspi et al., 2014; see section 3.5.7).

Internalizing. In contrast to anxious-depressed items that loaded almost exclusively onto *p*, anxious-withdrawn items (e.g., 'Doesn't get along with others') continued to load healthily onto the specific internalizing factor after accounting for *p*. Therefore, the specific internalizing factor likely reflects social withdrawal, but its reliability is too low for psychometric interpretation in terms of the proportion of variance it explained in raw and weighted internalizing subscale scores (e.g., omega hierarchical-subscale and ECV-subscale, respectively). Even if the internalizing factor was reliable, it is uncertain how 'social withdrawal' should be interpreted after being residualized for general psychopathology. It could represent social dysfunction beyond the *p* factor or an introverted trait free from pathological variance (Bonifay, Lane, and Reise, 2017; Caspi et al., 2014). Supporting the latter view, specific internalizing was most strongly and negatively correlated with extraversion (the inverse of which is believed to reflect introversion; Eysenck & Eysenck, 1963), but external criteria of personality and social functioning are needed to test this question more rigorously.

Externalizing. The externalizing factor showed the highest proportion of variance explained in raw and weighted externalizing subscale scores out of all the specific factors. Items that loaded most strongly onto specific externalizing at the expense of *p* were associated with rule-breaking (e.g., 'Has bad companions', 'Steals', 'Trouble with the law') and intrusiveness (e.g., 'Loud', 'Shows off'). Other studies have also reported preferential loadings of externalizing items, particularly those associated with drug and alcohol abuse and callous-unemotional behavior, onto the specific externalizing factor rather than *p* (Black et al., 2019; Carragher et al., 2016; Caspi et al., 2014; Conway et al., 2019; Gomez et al., 2019; Haltigan et al. 2018; Hyland et al., 2018; Lahey et al., 2012; Lahey et al., 2015; Lahey et al., 2017; Martel et al., 2017; Olino et al., 2014; Pezzoli et al., 2017; Romer et al., 2017; St Clair et al., 2016; Urbán et al. 2016).

The preferential loadings of antisocial-type externalizing items onto the specific externalizing factor raises questions about their relevance to general psychopathology. Some argue that antisocial tendencies, particularly callous-unemotional traits, have a distinct aetiology from internalizing problems (Benning, Patrick, Salekin, & Leistico, 2005; Pardini & Fite, 2010; but see Barker & Selekin, 2012; Euler et al., 2015), which would limit the extent that they converge on a general factor. Interestingly, Lahey et al. (2012) found that in addition to antisocial and drug-related problems showing balanced loadings on specific externalizing and *p* (despite internalizing items showing preferential loadings onto *p*), the specific externalizing factor, but not the specific internalizing factors, predicted aetiological variables beyond the *p* factor.

The bifactor model might separate out the unique variance associated with externalizing (e.g., callousness and proactive aggression), which is relatively strong, from the variance in externalizing shared with other disorders (e.g., reactive aggression), which is relatively weak. Indeed, there was a general trend for items associated with antagonism and callousness to load more strongly onto the specific externalizing factor (e.g., 'mean', 'attacks', 'threatens'), while items associated with reactive aggression and mood dysregulation to load more favourably onto the *p* factor (e.g., blames others, gets along badly with family, elation-depression, upset, mood changes).

Cognitive Problems. Psychotic and attention problems co-loaded onto a 'cognitive' factor that captured disturbances in attention, perception, memory and reality monitoring. Previous symptom-level bifactor analyses in adults have not included attention and psychosis items within the same model; therefore, it is uncertain whether a cognitive factor should be considered part of the bifactor structure of psychopathology or simply a product of the ASR. Most bifactor studies in adults have shown that psychosis items load exclusively onto the *p* factor (Caspi et al., 2014; Lahey et al., 2017; Martel et al., 2017; Romer et al., 2017; Rosenström et al., 2018; Urbán et al., 2016; Urbán et al., 2014; but see Arrindell et al., 2017; Conway et al., 2019; Pettersson et al., 2019). Some argue this occurs because psychosis items define the *p* factor as a continuum of severity (Caspi et al., 2014), while others propose that the inherent unreliability in psychosis items force them to load directly onto the *p* factor (Carragher et al., 2016; see section 2.2.1). The current findings support the latter argument: with an adequate number of psychosis items, a thought disorder-related specific factor can be identified (note that the cognitive factor was still identified even when attention items were moved to the externalizing factor).

Attention problems typically fall under the externalizing spectrum in child and adolescent studies (Burt, Krueger, McGue, & Iacono, 2003), but a sensitivity analysis showed that they fit slightly better on the cognitive factor. Attention problems may thus show heterotypic development, starting out as impulse control and behavioral problems in childhood, but shifting into disordered thought problems in adulthood. Nonetheless, the difference between models with attention items loading on the cognitive factor vs. externalizing factor was relatively small, and attention problems still loaded well on the specific externalizing factor. Therefore, there is no clear divide between the specific cognitive and externalizing factors, which might explain why they showed the strongest specific factor correlation (*r* = .65).

Somatic problems. Somatic problems typically show positive loadings on the internalizing factor in adults (Carragher, Krueger, Eaton, & Slade, 2015). However, somatic items loaded negatively onto the specific internalizing factor in the current study and required their own factor. This 'somatic' factor was weakly associated with the other specific factors. The independence of somatic problems in the current sample may be an effect of accounting for the common variance with *p*. That is, the association between somatic and internalizing problems may might be driven by a common underlying factor, controlling for which would remove their dependence.

The current study is not the first to demonstrate a dedicated somatic factor. For example, Pezzoli, Antfolk, and Santtila (2017) found that a 'body' factor, comprised of eating attitudes and body image concerns, ran alongside internalizing and externalizing factors in a bifactor model (although there are differences between somatic complaints and body image/eating attitudes). Moreover, Kotov et al. (2011) estimated a somatoform factor distinct from internalizing in a correlated factors model. Nonetheless, the somatic factor might be a product of subscale effects, since somatic items are answered consecutively in the ASR (all other items are randomized). Further work is necessary to determine the psychometric validity and utility of a somatic factor.

4.4.3 Are the general and specific psychopathology factors a product of response biases?

***p* factor.** Individual differences in the tendency to agree and disagree with questions regardless of their content weakly predicted the *p* factor, explaining just 4% of its variance. In an exploratory analysis, extreme responding and mid-point responding accounted for 4% of the variance in *p* (which is likely to overlap with the variance explained by agreement and disagreement biases). These findings suggest that response biases–measured with an independent measure of heterogeneous items–contribute very little to the *p* factor.

We cannot conclude that the positive manifold underlying the *p* factor is free from response biases because their influence was examined at the factor level (see section 4.4.5). However, we can infer that response biases have a minimal contribution to the *p* factor; only 4% of the systematic variance in *p* is attributable to response biases, suggesting that the majority of variance is substantive. There may, of course, be other systematic method effects that account for a larger proportion of the variance, given that a single measure and informant was used.

As predicted, the agreement bias factor positively but weakly predicted the *p* factor. Therefore, the more participants agreed indiscriminately to a heterogeneous set of items, the higher they scored on most symptom-items on the ASR, albeit slightly. Put differently, a small but significant proportion of the variance in people’s tendency to report multiple symptoms is explained by their general tendency to agree to items on questionnaires.

Contrary to predictions, the disagreement bias factor positively but weakly predicted the *p* factor. If a general disagreement to items artifactually contributed to the *p* factor, then it should have predicted lower *p* factor scores, while a general agreement to items artificially inflated the *p* factor. Instead, the more participants disagreed indiscriminately to a heterogenous set of items, the higher they scored on most symptom-items and hence the *p* factor.

One reason for the positive association between the disagreement bias factor and *p* factor is that a negative attitude towards benign content might overlap with a general disposition towards mental health problems. Lahey et al. (2012) proposed that a negative view of the self and world might be expressed as response biases that encourage agreement to any and all symptoms. A 'pessimistic' response bias still provides a substantive explanation of *p*: a negative attributional style might be a psychological feature or trait that contributes to negative internal working models or schemata that are characteristic of psychopathology (Carver & Scheier, 2017).

The current measure of disagreement bias might tap a broader oppositional tendency that positively predicts *p*. Indeed, disagreement bias tends to be stronger in people who show oppositional characteristics (Knowles & Nathan, 1997) and in countries that promote individualism (Baumgartner & Weijters, 2015). This argument is, of course, highly speculative, and it would be improper to make such inferences from a weak association. The main take-home is that disagreement bias contributes very little to the *p* factor, but its small contribution might be due to substantive reasons.

Specific factors. Externalizing was the only specific psychopathology factor predicted by response biases, including agreement bias and extreme responding (which likely predicted overlapping variance). Therefore, a preference to agree with a heterogeneous set of questions predicted a slightly higher tendency to report aggressive, rule-breaking and intrusive behavior, as would be expected from the inflating effects of an agreement bias. Like the *p* factor, the amount of variance in the specific factors explained by each response bias was minimal (<1%). Therefore, response biases assessed with a heterogeneous set of questions are unlikely to explain the systematic variance in externalizing. As for the other specific factors, absence of evidence is not evidence of absence; there may still be response bias measures or other systematic method factors that explain variation in the specific factors, including the externalizing factor.

Of all the specific factors, why was the specific externalizing factor predicted by response biases? Externalizing showed the highest specific factor reliability (omega hierarchical = .64) which neared the suggested cut-off for psychometric interpretability (≥ .7; Rodriguez et al., 2016a). Therefore, it is likely that there was more reliable variance in the externalizing factor that could be explained by response biases, rather than a special relationship between externalizing and agreement bias.

The correlations between specific factors were also minimally affected by response biases, at least when assessed with a heterogeneous set of questions. This is hardly surprising given what little variance was predicted by response biases. However, it remains uncertain what the specific factor correlations represent. We have already seen how the positive association between specific internalizing and externalizing factors is probably be a consequence of content overlap, and the positive association between cognitive problems and externalizing might be due to comorbidity in attention problems. Furthermore, the somatic factor might have shown the least overlap with other specific factors because it is a product of a subscale effect which lacks substantive variance. In all, there appears to be logical reasons for the overlap between specific factors–at least in the context of the ASR–that would explain why response biases contributed little.

4.4.4 Is the relationship between the *p* factor and neuroticism driven by response biases?

The relationship between the *p* factor and neuroticism factor was strong and positive; higher ratings of emotional instability were associated with higher symptom reports, almost perfectly. Some have hypothesised that neuroticism provides the substantive basis of the *p* factor (Lahey et al., 2012; Tackett et al., 2013; see section 2.2.2). While our results support this hypothesis, they do so with little error, which should be treated with caution. Removing the neuroticism items that overlapped with depression and anxiety symptoms significantly reduced the strength of the correlation, but it was still large in absolute terms. It is likely that other common method effects inflated the relationship between the neuroticism and *p* factors, since the underlying measures were assessed using the same medium in a short space of time by the same respondent (Podsakoff et al., 2012).

Controlling for agreement and disagreement biases, or extreme responding and mid-point responding, did not significantly alter the strength of the relationship between the neuroticism and *p* factors. It is, however, interesting that the disagreement bias factor positively but weakly predicted the neuroticism factor (as did extreme responding, but the variance explained is likely overlaps with disagreement bias). That is, a tendency to disagree with a heterogeneous set of items was associated with slightly higher ratings of emotional instability. While this association is too weak to infer anything conclusive, it does support the argument that the disagreement bias factor partially tapped a pessimistic tendency that tends to be more pronounced in people who score highly neuroticism (Baumgartner, Schneider, & Capiola, 2018).

In an exploratory analysis, the neuroticism factor was strongly and positively associated with the specific factors, even after controlling for agreement and disagreement bias, or extreme responding and midpoint responding. Therefore, it is unlikely that response biases (as measured with a heterogeneous set of items) contribute to the relationship between the specific psychopathology factors and neuroticism

4.4.5 Strengths and Limitations

The current findings demonstrate that response biases contribute little to the bifactor dimensions of psychopathology and their relationships with other variables. However, the strength with which response biases predicted the bifactor dimensions might have been limited by the way in which response biases were assessed and analysed. For example, the current study examined the influence of response biases on ASR ratings using a scale other than the ASR. Therefore, the current measure of response biases is indirect. Nonetheless, response biases are thought to influence all measures within a session (Podsakoff et al., 2012; Weijters et al., 2010a). Furthermore, the advantage of using a separate scale to assess response biases is that content and style were not conflated like in studies that use a single measure to assess both the construct of interest and response biases (see section 4.1.2).

Another reason for the weakened relationships between psychopathology factors and response biases is the limited number of response bias indicators. While each response bias indicator was created from 4-5 items, there were only three indicators per response bias factor. Therefore, the amount of reliable variance that could overlap between the response bias factors and psychopathology factors may have been limited (Epstein, 1983). Furthermore, there may have been a limited amount of variance in the psychopathology factors accountable by responses biases because the ASR's brief response scale limits biased responding (Achenbach & Rescorla, 2003). The fact that a *p* factor was estimated despite minimizing response biases is a testament to its substantive nature. However, brief response scales introduce other method effects that could underpin the positive manifold among symptoms, such as limited distributions that inflate inter-item correlations (Sellbom & Tellegen, 2019).

In the current study, response biases were controlled for at the factor level. Therefore, we can say something about the extent to which the psychopathology factors are predicted by response biases, but not the extent to which the positive manifold in symptoms ratings is a product of response biases. An alternative method proposed by Williams and Anderson (1994) is to control for response biases at the item-level by loading the items of interest (e.g., ASR items) onto a response bias factor (e.g., a factor that includes response bias indicators for a given response style), in addition to substantive factors (e.g., a *p* factor). In turn, the items that load onto the substantive factors are residualized for response biases (provided that the response bias factors reflect response biases).

A problem with the item-level approach to controlling for response biases is that it is incompatible with the bifactor model because it is, itself, a bifactor model. One tests for the presence of response biases by comparing a model that freely estimates the response bias indicators loading onto the response bias factor to one where the response bias indicator loadings are constrained to zero (Williams &

Anderson, 1994). However, constraining the response bias indicator loadings to zero in the context of a bifactor model produces two general factors (e.g., a *p* factor and a response bias factor that includes all the ASR items but no response bias indicators; essentially two *p* factors) and hence a singularity in the factor matrix. Estimating the overlap between the bifactor dimensions and response biases at the factor-level was a natural compromise, but one that lacks the power to explicitly control for response biases at the level of responding.

The current sample was recruited through crowdsourcing (i.e. engaging groups in a common task, typically online; Estellés-Arolas & González-Ladrón-de-Guevara, 2012). Crowdsourcing is ideal for conducting large-scale psychometric studies with relative ease, speed, and cost-effectiveness (Hewson, Vogel, & Laurent, 2016). Comprehensive psychiatric assessments run in sufficiently large samples are needed to estimate the bifactor model of psychopathology, which limits much of the work to secondary analyses of large, multi-wave clinical trials and epidemiological studies. This also limits what can be investigated, since the data has already been collected. Crowdsourcing made it possible to recruit a large sample of respondents and test the contribution of response biases to the bifactor dimensions.

The drawback of efficiently collecting large sums of online data is that there is greater scope for breaching ethical and data protection guidelines, particularly when collecting and managing sensitive data about mental health (Miller, Crowe, Weiss, Maples-Keller, & Lynam, 2017). For instance, asking respondents to reflect on troubling experiences can cause them concern and evoke emotions such as shame and embarrassment (Tourangeau & Yan, 2007). Respondents were therefore given the option to skip single questions or whole questionnaires, provided information about mental health support on each webpage, and immediately contacted following a set protocol if they reported feeling concerned or triggered by questionnaire material. Furthermore, efforts were made to ensure that responses were not identifiable from their data by pseudo-anonymizing personal details, limiting the collection of personal information such as IP addresses, and storing data in secure environment (e.g., in an encrypted drive on a protected server).

Another issue of crowdsourcing is that we cannot verify the identity of our respondents–how do we know who is in the crowd? (Stewart, Chandler, & Paolacci, 2017). Crowdsourcing samples are self-selected and thus not representative of the general population (Chandler & Shapiro, 2016). However, crowdsourcing samples may still be a useful alternative to student samples. For example, Peer, Brandimarte, Samat, and Acquisti (2017) found that a sample recruited from Prolific.ac showed similar education levels, English fluency, and naivety to the measures taken compared to a student sample, but lower ethnic diversity. While Peer et al.'s participants were mainly recruited from the U.S, we see a similar trend in the current U.K sample, who were an educated group of mainly Caucasian adults. Given that we rely on student samples for personality research despite them lacking characteristics of the general population, crowdsourcing samples might offer a viable alternative that, if anything, are more representative of the general population in terms of age and income, with the obvious caveat that direct generalization in both samples is limited.[13]

There are also concerns about the data quality of crowdsourced samples (Chandler & Shapiro, 2016). Researchers have little control over how questionnaires

[13] Prolific.ac piloted a tool for stratifying samples to characteristics in the general population, but data for the current study had already been collected. This will prove to be an exciting tool that addresses generalizability issues associated with crowdsourced samples.

are completed online; respondents might engage in behaviours that threaten data quality, such as multi-tasking (Chandler et al., 2014; Necka et al., 2016). Nonetheless, Peer et al. (2017) found that the quality of the data collected from a Prolific.ac sample matched or surpassed the data quality of a student sample in terms of the internal consistency of questionnaire responses, replication of existing experimental effects, performance on attentiveness checks, and response rates.

Crowdsourcing studies of psychopathology have also shown that the five-factor structure of personality and personality disorder traits is replicable in both crowd-sourced and student samples (Feitosa, Joseph, & Newman, 2015; Miller et al., Lynam, 2017). Furthermore, prevalence rates and test-retest reliability of depression reported via crowdsourcing is similar to the general population (Shapiro, Chandler, & Mueller, 2013). Lastly, common effects in psychopathology research are replicated in crowdsourced samples, including the mediating effect of emotion regulation on various symptoms (Fergus & Bardeen, 2014; Raines, Boffa, Allan, Short, & Schmidt, 2015; Rose & Segrist 2012). The studies reviewed, as well as the current study[14], demonstrate that crowdsourcing offers a viable and efficient alternative for collecting psychopathology data that is of a good standard.

4.4.6 Implications and Future Directions

The current study demonstrates that basic response biases have little influence on the general and specific psychopathology factors, but there may other response biases that are more sensitive to psychopathology measures. For example, a tendency to portray oneself and the world in negative terms (pessimistic

[14] Consider the minimal amount of missing responses and measurement invariance across completion times.

responding) or positive terms (optimistic responding) could encourage people to agree and disagree to any and all symptoms, respectively (Lahey et al., 2012). Furthermore, respondents (particularly patients) might under-report or over-report their symptoms to appear more or less impaired, respectively (Ben-Porath, 2013), which could artificially inflate and deflate the relationships among symptoms.

In future, researchers could investigate the extent to which optimistic and pessimistic response tendencies inflate or deflate scores on the bifactor dimensions of psychopathology. Standard self-report measures of optimism and pessimism are unlikely to be appropriate, however, as the overlap in content between optimism/pessimism questions and well-being/depression items would conflate the associations between response biases and psychopathology factors. 'Pseudo-behavioural' measures could be developed that capture optimistic and pessimistic response patterns, rather than people's beliefs about how optimistic or pessimistic they are. Perhaps the most creative example is the work of Paul Meehl, who developed scales for detecting patients who would "fake good" based on the extent to which they agreed to questions reflecting socially desirable behaviour, e.g. "I have very few quarrels with members of my family", and patients who would "fake bad" based on their responses to questions such as "I hate my whole family"(Meehl & Hathaway, 1946).

Future research could also broaden the methods used to assess response biases. For example, response biases can be modelled with Item Response Theory, which separates out the influence of latent traits from item characteristics on response patterns (Lord & Novick, 1968). Using the Multidimensional Nominal Response Model, one can model the probability of choosing certain response categories over others, such as the highest response category (agreement bias) or lowest response category (e.g., disagreement bias), whilst also modelling variation in responses explained by latent traits (Bolt & Johnson, 2009; Johnson & Bolt, 2010). The multidimensional nominal response model has been used to estimate response biases in the context of a bifactor model (von Davier & Khorramdel, 2013).

An alternative approach to measuring response biases is through anchoring vignettes: written descriptions of hypothetical people or scenarios that are rated with the same response scale as the substantive measure (King, Henry, Salomon, & Tandon, 2004). The vignettes are designed with a common response option in mind; deviations from the "correct" response option index response biases. Variability in responses can be analysed in various ways, including the proportion of certain selecting response categories (Mõttus et al., 2012) or multidimensional nominal response models (Bolt, Lu, & Kim, 2014). Anchoring vignettes were recently used to control for agreement bias, disagreement bias, extreme responding, and mid-point responding in personality disorder ratings (Jonas & Markon, 2019). Controlling for response biases resulted in a small but significant improvement in personality disorder trait estimates, quite like the current results.

Finally, studies could capitalize on the power of crowdsourcing to test novel questions about the structure of psychopathology. For instance, most of studies on the bifactor model have been conducted in high-income western countries. The invariance of bifactor dimensions could be studied across countries of different wealth and cultural norms made accessible online (Wüsten et al., 2018). We live in a digital age where mental health is both influenced by and expressed through digital communication (Hayes, Maughan, & Grant-Peterkin, 2016). It would also be interesting to investigate how online behaviour, such as social media use, is associated with the general and specific psychopathology factors. It may even be possible to predict the onset of mental disorder using online behaviours that correlate with *p*, like how the spread of infectious diseases can be predicted using smartphone and search engine data (Bengtsson et al., 2015; Ginsberg et al., 2009).

In all, the current study shows that the latent structure of symptom reports on the Adult Self Report is best explained by a bifactor model, with a general *p* factor that describes the commonalities among symptoms, and specific internalizing, externalizing, somatic, and cognitive factors that describe commonalities among specific problem domains. The tendency to indiscriminately agree or disagree with questionnaire items regardless of their content makes a small but significant contribution to the general and specific psychopathology factors. Furthermore, the relationships between the general factor and theoretically relevant variables (e.g., neuroticism), and relationships between specific factors, are minimally explained by these response tendencies.

Chapter 5 Changes in the General and Specific Psychopathology Factors Over a Psychosocial Intervention

The prior chapters have addressed methodological issues associated with the bifactor model of psychopathology. For instance, the latent structure of psychopathology measures was found to be multidimensional, but raw total scores and subscale scores mainly reflected the *p* factor (Chapter 3). Furthermore, response biases contributed little to the general and specific psychopathology factors, demonstrating their substantive validity (Chapter 4). The next two chapters move away from methodological issues and address the bifactor model's clinical utility when applied to clinical outcomes data (which introduces methodological issues of its own, of course). The current chapter investigates changes in the general and specific psychopathology factors over a psychosocial intervention for antisocial adolescents and what this suggests about the processes of therapeutic change (e.g., disorder-wide vs. disorder-specific change).

I begin by outlining the problems associated with disorder-specific assessment, particularly the inter-relatedness of disorder-specific measures and the confounding of shared variance on disorder-specific prediction. I then describe how the *p* factor and disorder-specific psychopathology factors can be estimated for a given individual over time, rather than across multiple respondents at a single time-point, and examine changes in these factors over a psychosocial intervention. The results are compared to a correlated factors growth model that does not control for the shared variance among disorders, with the aim of demonstrating the clinical value of the *p* factor in addressing questions such as ‘what changes?’ over an intervention.

5.1 Introduction

5.1.1 The Problem with Disorder-Specific Assessment

Psychiatric assessment in research and practice is governed by disorder-specific measures. For example, complaints of low mood and loss of pleasure are typically formalized with a self-report measure of major depressive disorder. Furthermore, researchers might investigate the neurobiological correlates of schizophrenia assessed with an interview. Disorder-specific scales have made mental disorders more accessible for research and clinical monitoring but imply that people’s problems are distinct entities that occur in isolation (Nesse & Stein, 2012).

Mental health disorders frequently co-occur: roughly 50% of people who meet the criteria for one disorder also meet the criteria for another, and roughly 50% of those people also meet the criteria for three or more disorders (Andrews, Slade, & Issakidis, 2002; Bijl, Ravelli, & Zessen, 1998; Kessler et al., 1994; Robins & Regier, 1991; Wittchen, Nelson, & Lachner, 1998). Disorders within the same diagnostic group show the highest rates of comorbidity (e.g., various anxiety disorders co-occur within the same individual), but there is also broad overlap between diagnostic groups (e.g., anxiety disorders co-occur with depression and alcohol dependence; Andrews et al., 2002). The broad overlap between disorders is not random, with depressive disorders showing the highest comorbidity rates and substance-use disorders showing the lowest (but this is influenced by absolute prevalence and sex differences; Andrews et al., 2002; Kessler et al., 1994; Wittchen et al., 1998). Systematic patterns in overlap suggest that comorbidity is not simply an artifact of the system (e.g., overlapping symptoms), but a result of a latent hierarchical structure with disorders reflecting manifestations of a higher-order dimensions (Kotov et al., 2017).

Comorbidity threatens the integrity of disorder-specific outcome monitoring. Imagine some researchers wish to investigate the efficacy of a psychological intervention for reducing depressive symptoms. They randomize two groups of adults who meet the threshold for major depressive disorder (PHQ-9 scores ≥ 10; Kroenke at al., 2010) to receive either an evidence-based therapy or a wait-list condition, and exclude participants with current psychotic problems and/or suicidal intent and a history of neurodevelopmental or personality disorders. They assess depression at three time-points (baseline, mid-treatment, post-treatment) with the self-reported Beck Depression Inventory-II (Beck et al., 1996) and clinician reported Hamilton's Rating Scale for Depression (Hamilton, 1960), and find that both self- and clinician-reported depression scores decline over time, but at a significantly steeper rate for the psychological intervention. The authors conclude that their intervention is more effective than chance at reducing depressive symptoms.

The reader is probably familiar with hundreds of trials like this, but beneath this familiarity lies a profound and as of yet unanswered question: did the psychological intervention influence a distinct clinical entity (e.g., 'depressive disorder') or a broader underlying entity from which depressive problems manifest (e.g., general psychopathology). Participants would probably meet the criteria for other disorders (only the most severe presentations tend to be excluded) that may have also been influenced by the intervention. However, limiting the assessment to a single disorder, and even designing the intervention around it, might give a false sense of treatment specificity. A recent study by Schawo, Carlier, Hemert, and Beurs (2019) showed that for most anxiety disorders, broad diagnostic instruments (e.g., Mood and Anxiety Symptoms Questionnaire and Brief Symptom Inventory) showed similar sensitivity to treatment-related change compared to disorder-specific measures.

The question of what changes over an intervention is tightly linked to the common vs. specific processes[15] debate in psychotherapy research. An intervention that works through specific processes is thought to include certain “ingredients” that remediate disorder-specific mechanisms, while an intervention that works through common processes is thought to emphasise processes common to all effective psychotherapies, such as a robust therapeutic alliance, regardless of the disorder treated (Mulder, Murray, & Rucklidge, 2017), although certain disorders might interact with different common processes (Hofmann & Barlow, 2014). A more comprehensive assessment of what changes over an intervention is therefore important for informing the debate between common vs. specific therapy processes, and ultimately how we should design our interventions.

5.1.2 Principle of Intwined Generality

Psychopathology assessment measures feature both general and specific components (see Chapter 3). The extent to which subscale scores provide precise information about specific domains of a construct will be influenced by the strength of the general construct as their variances are ‘intwined’. This is known as the principle of intwined generality (i.e. different measurement levels are conflated

[15] Common and specific processes are more commonly referred to as common and specific factors. I used the former phrase to avoid confusion between the term ‘factor’ (something that influences an outcome of some kind) and ‘factor’ (a latent variable).

within a measure), which implies that one must control for the general variance in order to assess the unique influence of the specific variance (Gustafsson, 2002). Failing to do so will confound the effects of the specific variables with their commonalities. Using the bifactor model, one can study the associations between predictors (e.g., a psychological intervention) and specific factors (e.g., disorder-specific symptom reports) free from the influence of the general factor (e.g., general psychopathology).

Several studies have demonstrated the importance of separating out the general and specific variance when investigating the relationships between psychopathology dimensions and external variables. For example, the *p* factor is most strongly associated with risk factors for psychopathology at the expense of specific internalizing and externalizing factors, despite their counterparts showing moderate-to-strong associations in the correlated factors model (Caspi et al., 2014; Lahey et al., 2012; Schaefer et al., 2018). These findings demonstrate that the associations between internalizing and externalizing factors and risk factors were underpinned by their shared variance, which is captured by the *p* factor in the bifactor model (the correlated factors model conflates the general and specific variance). However, specific factors and correlated factors are not isomorphic and hence not readily comparable (Bonifay, Lane, & Reise, 2017).

The *p* factor also shows the strongest prediction of academic outcomes and simultaneously 'explains away' the variance predicted by internalizing and/or externalizing factors in the correlated factors model (Lahey et al., 2015; Patalay et al., 2015; Sallis et al., 2019). The effect is observed with clinical outcomes too: diagnostic indicators (Thomas, 2012) and polygenic risk scores (Jones et al., 2019) show widespread associations with correlated factors which are mainly explained by the *p* factor and a domain-relevant specific factor in the bifactor model. Therefore, much of what drives the associations with psychopathology factors might be due to commonalities among the disorders assessed rather than their unique features. This does not mean that disorder-specific effects are clinically or theoretically unnecessary. Rather, the reliability of specific psychopathology domains is overshadowed by a single dimension of mental ill-health.

5.1.3 Assessing Clinical Outcomes with the Bifactor Model

The bifactor model provides a means of addressing the question: ‘what changes over a psychosocial intervention?’ Following the principle of intwined generality, disorder-specific changes can be assessed with specific psychopathology factors, free from the influence of a general underlying dimension as represented by the *p* factor–something that cannot be adequately achieved with a correlated factors model or subscale scores as they both conflate the general and specific variance.

Changes in the general and specific psychopathology factors might also capture common and specific therapy processes. For example, changes in the *p* factor over an intervention would reflect changes common to all symptoms, which is, by definition, a transdiagnostic or common therapeutic effect (McEvoy, Nathon, & Norton, 2009). Because specific factors are residualized for the general variance, they might be useful in identifying the specific effects of a treatment on the problems it was designed to engage with. The bifactor model allows us to estimate both common and specific processes based on change in the general and specific psychopathology factors, respectively, which is consistent with current theories proposing that both common and specific processes contribute to therapeutic change rather than either one alone (Fonagy & Allison, 2014; Wampold, 2015).

Only one study to date has analysed clinical outcomes with the bifactor model. Wade, Fox, Zeanah, and Nelson (2018) reanalysed data from the Bucharest Early Intervention Project, a randomized controlled trial where children were assigned to receive either foster care or institutional care in Romania when they were almost two years old. In the original analysis, the foster care group showed a lower prevalence of anxiety disorders[16] and externalizing disorders at age 4.5 compared to the institutional care group, as assessed with caregiver reports on the Preschool Age Psychiatric Assessment (PAPA) interview (Zeanah et al., 2009). Therefore, the widespread benefits of foster care might have been driven by a broader underlying factor. At age 12, the foster care and institutionalized care groups showed higher internalizing and externalizing symptom counts on the Diagnostic Interview Schedule for Children-IV than an age- and sex-matched group without a history of abandonment (Humphreys et al., 2015). However, the foster care group showed significantly fewer externalizing symptoms than the institutionalized care group. These findings further suggest that a broader underlying factor might be elevated in both groups in care, but the quality of care might have had a specific effect on externalizing problems.

Wade et al. (2018) analysed changes in caregiver- and teacher-reported disorder ratings on the MacArthur Health and Behavior Questionnaire between ages 8, 12, and 16 with a bifactor model. They estimated a *p* factor and uncorrelated specific internalizing and externalizing factors at each age and compared changes in standardized factor scores across ages between foster care, institutional care, and non-institutionalized groups using a latent growth-curve model. Wade et al. found

[16] Depression was also assessed but showed a low prevalence rate in both groups of children in care, matching the prevalence rate of age- and sex-matched children without a history of institutionalization.

that the foster care group showed a marginally significant decline in *p* factor scores across age, while the institutional care group and non-institutionalized groups showed consistently high or low scores, respectively. Moreover, the foster care group showed a significant decline in specific externalizing problems, while the institutional care and non-institutionalized groups showed small declines that did not reach significance. Finally, the foster care group showed a subtle increase in specific internalizing scores (from a lower than average starting point), while the institutional care group and non-institutionalized groups showed subtle declines, none of which were significant.

Wade et al.'s (2018) findings suggest that psychosocial interventions influence both disorder-wide and disorder-specific mechanisms. Decline in the *p* factor might reflect the common effects of foster care on symptoms, which was expected from early changes in both internalizing (anxiety) and externalizing problems in the original analysis. Furthermore, decline in the specific externalizing factor might reflect the specific effect of foster care on behavioural problems, which was also expected from the later changes in externalizing in the original analysis. The subtle increase in the specific internalizing factor in the foster care group might have masked group differences at a later age in the original analysis.

We must take care not to overinterpret Wade et al.'s (2018) findings because the nature of specific factors remains uncertain, especially without external validation (Bonifay, Lane, & Reise, 2017). Furthermore, it is hard to attribute changes in the bifactor dimensions directly to the quality of care received because the measures analysed were taken some years after the randomization to care groups. Direct comparison between observed scores and bifactor scores is also limited by the fact that the measures and timescales were not the same. Finally, there is some concern that splitting the analysis by treatment group would reduce power in a sample that is already relatively small (*N* = 220).

Wade et al. (2018) demonstrated longitudinal differences in the mean level with which symptom counts positively co-occurred *between* individuals, which does not tell us whether symptom counts positively co-occurred *within* each individual and how this changed over time. In fact, there might be a positive correlation between two disorders across a group of individuals (e.g., people who score higher on disorder X also score higher on disorder Y, whereas people who score lower on disorder X also score lower on disorder Y), but repeated measurements for each individual show that the disorders negatively correlate over time (e.g., when participants report high scores on disorder X at a given time-point, they tend to report lower scores on disorder Y; Reise, Ventura, Nuechterlein, & Kim, 2005). It is therefore important to analyse within-person vs. between-person changes in outcome data because we typically treat a single individual in practice rather than change trends in large groups of people.

A study by Wright, Hopwood, Skodol, and Morey (2016) examined changes in within-person levels of the general and specific factors estimated from personality disorder (PD) ratings in in 733 outpatients over a ten-year period. The bifactor model adequately fit the within-person data; therefore, the positive co-occurrences between PDs for each individual over time could be summarized by a general PD factor, as well as specific PD factors summarizing positive co-occurrences between overlapping PDs. Furthermore, the general PD factor significantly declined over the ten years, while the specific factors remained relatively stable apart from the compulsivity factor, which showed a marginal decline. Wright et al. also demonstrated the validity of their general and specific factors by predicting self-reported social, occupational, and vocational functioning.

Wright et al.'s (2016) findings show that the positive manifold among symptoms or disorders is observable at the within-person level. They also support the *pd* factor's sensitivity to various forms of intervention, since patients received at least one form of treatment throughout the study period. However, it is hard to infer causality when the type, duration, and intensity of treatment were unstandardized. Moreover, we cannot tease apart the effect of treatment from the natural passage of time without a control group. What is needed is a study that analyzes the bifactor dimensions over the course of a controlled intervention with external criteria.

5.1.4 Study Aims

The bifactor model offers an opportunity to investigate disorder-wide and disorder-specific changes in clinical outcomes to address the question of 'what changes?' over an intervention. Studies that have applied the bifactor model to longitudinal outcomes are limited by indirect or uncontrolled interventions, between-level vs. within-level analyses, and a lack of external validation of the bifactor dimensions. The current study evaluates changes in a general *p* factor and disorder-specific factors over an 18-month psychosocial intervention. The data are from the Systemic Therapy for At-Risk Teens (START) trial, a pragmatic randomized controlled trial that compared the effects of multisystemic therapy (MST) with those of management as usual (MAU) in decreasing antisocial behaviour in adolescents (Fonagy et al., 2018). In the original analysis of subscale scores, there were widespread reductions in self-reported conduct problems, attention problems, mood problems, and anxiety across both treatment arms, with few differences between arms.

The first part of the current analysis involves a comparison of the bifactor model against a correlated factors model (which resembles the use of subscale scores) and a single-factor model. Models were compared for how well they summarized the within-person associations among emotional and behavioural symptoms throughout the study period, collapsed across treatment arms. The second part involves an analysis of the general and specific factors' reliability using model-based reliability estimates, and concurrent validity using external records of criminal activity and academic attendance. The third and final part involves an assessment of within-person change in the general and specific psychopathology factors and between-person differences in within-person change. Between-person variation was also predicted by clinical and demographic covariates.

The following hypotheses were tested:

1) Model fit indices will favour the bifactor model in summarizing the within-person covariances in symptoms compared to the correlated factors model and single-factor model.
2) The *p* factor will explain the majority of variance in raw total and subscale scores, even if the latent data structure is multidimensional. The *p* factor will also show the strongest prediction of external criteria compared to the specific factors.
3) The *p* factor will show large reductions in mean levels over the treatment and follow-up period, since widespread reductions were reported in the primary analysis of observed subscales. There might also be a reduction in

mean levels of the specific antisocial factor, reflecting the target of intervention. When predicting between-person differences, the treatment arms might show comparable declines in the *p* factor, as there were no differences between treatment arms in the primary analysis. However, the intervention group (MST) might show steeper declines in the specific antisocial factor compared to the active control group (MAU), reflecting specific treatment differences that were masked in the primary analysis by common processes.

5.2 Method

5.2.1 Design and Participants

The START trial was a pragmatic individually randomized multicenter superiority trial that compared the effects of MST followed by MAU with MAU alone in decreasing out-of-home placements and criminal activity in adolescents with moderate to severe conduct problems (Fonagy et al., 2018). Assessment occurred at baseline, after treatment (6 months), at follow-up 1 (12 months), and at follow-up 2 (18 months).

Eligible adolescents met at least one of the following criteria: persistent (weekly) and enduring (≥ 6 months) violent and aggressive interpersonal behaviour; at least one conviction plus three additional warnings, reprimands, or convictions; a current DSM-IV diagnosis of CD that had not responded to treatment; a permanent school exclusion for antisocial behaviour; or a significant risk of harm to others or self. Eligible adolescents also met at least three severity criteria indicative of past difficulties across several settings (e.g., school non-attendance or exclusion, offending, child protection, high risk of coming into care). Because this was a pragmatic trial, participants were not excluded for comorbid disorders except for psychosis, acute suicidality, and generalized learning difficulties.

Baseline demographics for the final 684 adolescents are shown in Table 5.1. Adolescents were referred from social services, youth justice, schools, child and adolescent mental health services, and voluntary services. The final sample consisted of 684 adolescents (11–18 at baseline), most of whom were Caucasian boys of a low to moderate socioeconomic background. Most adolescents had a diagnosis of conduct disorder (78%) or any conduct disorder (81%). Other frequent disorders included attention-deficit/hyperactivity disorder combined (30%) and any emotional disorder (24%). Written consent was obtained from all participants, and the study protocol was approved by the London South-East Research Ethics Committee (09/H1102/55).

Table 5.1

Baseline Demographics of the 684 Adolescents Recruited to the START Trial

Measure	Mean or *n*	SD or %
Demographic		
Age (years)	13.8	1.4
Sex (male)	559	82%
Ethnicity		
White British/European	535	78%
Black African/Afro-Caribbean	71	10%
Asian	16	2%
Mixed/Other	51	7%
Socio-economic Status (1-6)	3	1.4
Family Income		
% on state benefits or <£20k pa	525	77%
Offences in Year Prior to Referral		
Non-offender on referral	235	34%
Total number of offences	1.2	2.5
Violent	0.4	1.0
Non-violent	0.6	1.3
Number with custodial sentences	10	1%
Psychiatric Diagnosis[a]		
Conduct disorder	532	78%
Oppositional defiant disorder	28	4%
Any conduct disorder	554	81%
Social phobia	21	3%
Obsessive-compulsive disorder	3	0.4%
Posttraumatic stress disorder	51	7%
Separation anxiety disorder	22	3%
Specific phobia	19	3%
Generalised anxiety disorder	15	2%
Panic disorder	8	1%
ADHD Combined	204	30%
ADHD Hyperactive–Impulsive	11	2%
ADHD Inattentive	25	4%
PDD/autism	7	1%
Eating disorders	4	1%
Tic disorder	11	2%
Major depression	72	11%
Any emotional disorder	163	24%
Mixed anxiety/conduct disorder	102	15%
No diagnosis	100	15%

Note. SD = standard deviation; PDD = pervasive developmental disorder; ADHD = Attention-deficit hyperactivity disorder.

[a]Assessed using the Development and Well-Being Assessment (Goodman, Ford, Richards, Gatward, & Meltzer, 2000).

5.2.2 Intervention and Randomization

MST is a family-based intervention that targets the multiple systems influencing chronic and pervasive antisocial behaviour in adolescents, including the home environment, school environment, and peer groups (Henggeler, 2012; Henggeler & Schaeffer, 2016). This is primarily achieved through caregivers who are taught how to enhance family relationships via communication skills and parenting techniques, as well as how to encourage school attendance and achievement rather than delinquent peer activity. Techniques from cognitive-behavioural therapy, behavioural parent training and pragmatic family therapy are integrated and tailored to the needs of each family. There were nine MST pilot sites across three U.K regions with at least 12 months experience of running the programme.

MAU was designed to mimic best-practice in managing the complex needs of antisocial youth in community settings. Interventions based on treatment guidelines were administered on an ad-hoc basis (e.g., support to re-engage with education, anger management, victim awareness programmes). MAU was multi-component and no less resource-intensive than MST; the main differences were that MAU lacked standardization, an overarching formulation of the problem, and weekly expert supervision. As our goal was to investigate the bifactor dimensions over a multi-component intervention, we collapsed the analysis over the treatment conditions.

Adolescents were randomized to MST or MAU by an equal allocation ratio using stochastic minimization, balancing for treatment center, sex, current age (< 15 or ≥ 15 years), and age at onset of antisocial behavior (≤ 11 or > 11 years).

5.2.3 Measures

All measures were taken at baseline, post-treatment, follow-up 1, and follow-up 2. Emotional and behavioural problems were assessed using the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), a widely used measure of young people's mental health with good validity (Lundh, Wangby-Lundh & Bjarehed, 2008; Muris, Meesters & van den Berg, 2003) and reliability (Goodman, 2001; Yao et al., 2009). Child-reported items from the emotional problems, conduct problems, and attention-hyperactivity subscales were used in the measurement models, each of which has five items rated on a 3-point scale (not true, somewhat true, or certainly true). Items from the peer problems and prosocial subscales were not included because the analysis was limited to psychiatric symptoms, which naturally excludes prosocial items, and general difficulties engaging with peers are not disorder-specific. That is, interpersonal problems reflect a broader level of analysis (e.g., children can be rejected because they appear nervous and withdrawn or because they are bold and irritable), which have not yet been thoroughly validated in the bifactor model of psychopathology.

The Mood and Feelings Questionnaire-Short Form (MFQ-SF; Angold et al., 1995) was used to increase the internalizing item pool. The MFQ-SF is a 13-item measure shown to reliably assess depression in young people (Daviss et al., 2006; Wood et al., 1995). Like the SDQ, items are scored on a three-point scale (not true, somewhat true, true). Past research suggests that the MFQ-SF captures a single depression factor (Kuo, Stoep, & Stewart, 2005; Lundervold, Breivik, Posserud, Stomark, & Hysing 2013; Sharp, Goodyer, & Croudace, 2006), but our exploratory multi-level factor analysis revealed two clear factors at the within-person level (see Table C1). The first factor reflected problems with self-attitudes and the second captured problems in mood. The five items loading on the mood factor were included in the measurement models to balance the internalizing and externalizing content of the SDQ and to ensure that the *p* factor was not biased to any one symptom domain. Five items were used to ensure that equal numbers of items loaded on each factor. The mood factor was used because the self-attitudes factor reflects a transdiagnostic construct that has not yet been validated in the bifactor model of psychopathology.

Official records of violent and nonviolent offenses committed during the study period were collected from the Police National Computer and Young Offender Information System. Violent and violent offenses were collapsed into a single count variable. Records of the number of school exclusions were collected from the National Pupil Database.

5.2.4 Data Quality Checks

Missing Data. In the START dataset, 57% (*n* = 389) of cases completed all observations. There were four main patterns of dropout:

1. Cases who withdrew before baseline data collection began (*n* = 1 or 0.1%)
2. Cases with baseline observations only (*n* = 83 or 12%).
3. Cases with observations from baseline to post-treatment (*n* = 82 or 12%)
4. Cases with observations from baseline to the first follow-up (*n* = 129 or 19%)

To investigate the assumption that data were missing completely at random (MCAR, i.e. the likelihood that a data point is missing is totally random) vs. missing at random (MAR, i.e. the likelihood that a data point is missing is dependent on some observed or unobserved data other than the missing value itself), a growth model that assumed the data were MCAR (e.g., a complete case analysis with 532 observations missing over the four time-points) was compared with a model that assumed the data were MAR (e.g., missing data were handled with full-information maximum likelihood; Little et al. 2012; see supplement C1 for background). The size of the difference between coefficients fell well below the cut-off of 10% (*M* = 0.88%, *SD* = 0.96%), suggesting that no major bias was introduced when estimating the missing data under MAR.

To test the assumption that the unobserved data were MAR vs. missing not at random (MNAR, i.e. the likelihood that the data point is missing is dependent on the missing value itself), a Diggle-Kenward selection model was adapted for multi-level models (Enders, 2011; Falkenström, Granström, & Holmqvist, 2013; see supplement C1 for background). A binary survival variable coding for the time until dropout was regressed onto the within-level *p* factor and specific antisocial, attention, anxiety, and mood factors (see 'Multilevel Factor Analysis' below). The antisocial factor positively predicted dropout onset (β = .30, *p* < .001, 95% CI [.28, .32]); therefore, adolescents who reported higher antisocial tendencies may have been more likely to dropout. Higher *p* factor scores (β = -.06, *p* < .001, 95% CI [-.12, -.03]), mood scores (β = -.02, *p* < .01, 95% CI [-.04, -.01]), and attention scores (β = -.08, *p* < .001, 95% CI [-.11, -.06]) weakly predicted a lower likelihood of dropout. Therefore, adolescents who showed less severe presentations might have dropped out slightly sooner. Findings from both sensitivity analyses suggest that the unobserved data were probably dependent on the outcome variable, particularly externalizing problems, but the extent of bias might be minimal.

Missing data were handled with full-information maximum likelihood (FIML; see supplement C1 for background) because the statistical software Mplus requires maximum-likelihood estimation to estimate random slopes. The alternative approach to handling missing data, multiple imputation (MI), requires weighted-least squares estimation which is not compatible with random slopes (see supplement C1 for comparison). Furthermore, estimating latent factors for both psychopathology dimensions and their growth curves required several levels of integration which is computationally taxing. Integrating the model over multiple replica datasets using MI would have drastically lengthened computation time compared to a single run with FIML.

Response Distributions. On average, the first response option ('Not True') was used 40% of the time (*SD* = .19, range = .14-.69), the second response option ('Sometimes True') was used 35% of the time (*SD* = .09, range = .21-.57), and the third response option ('True'/'Certainly True') was used 25% of the time (*SD* = .13, range = .08-.51). Therefore, the use of different response options was roughly equal, with fewer endorsements for the highest response (as expected) but enough to prevent large skews in the response distributions.

There were few differences between the estimated and observed response distributions in the standard bifactor model (*M* = .003, *SD* = .002, range = -.007–.008), revised bifactor model (*M* = .003, *SD* = .002, range = -.008–.009), correlated factors model (*M* = .002, *SD* = .002, range = -.006–.006), and single factor model (*M* =

.002, *SD* = .002, range = -.004–.005; see ‘Multilevel Factor Analysis’ for model specification).

Residual Correlation Matrix. The residual correlation matrix included 210 unique polychoric correlations between SDQ and MFQ items. No model substantially under-estimated (i.e. positive residual) or over-estimated (i.e. negative residual) the item correlations. On average, positive and negative residuals fell below the standard cut-off of .20 (Christensen, Makransky, & Horton, 2017) or within the stricter cut-off of .10 (Goodboy & Kline, 2017; see Table 5.2). Less than 1% of residuals were ‘potentially problematic’, i.e. falling above or below an absolute residual value of .25 (the average residual +/- .2), and no residuals were ‘problematic’, i.e. falling above or below an absolute residual value of .35 (the average residual +/- .3; Pallant & Tennant, 2007).

Table 5.2

Summary of Residual Correlations for Each Within-Person CFA Model

Model	Positive Res	Negative Res	*M* +/- 0.20	*M* +/- 0.30
Single factor	11 (.08)	-.08 (.06)	0 (0%)	0 (0%)
Correlated factors	.06 (.04)	-.05 (.04)	1 (0.4%)	0 (0%)
Bifactor (uncorrelated)	.06 (.06)	-.06 (.04)	2 (0.9%)	0 (0%)
Bifactor (x-loadings)	.05 (.04)	-.05 (.03)	0 (0%)	0 (0%)

Note. *M* = Mean; Res = Residual; x-loadings = cross-loadings. Mean and standard deviations (parenthesis) are provided for the average positive and negative residual correlations. Counts and percentages (parenthesis) are provided for the number of residuals falling above or below the mean residual +/- .20 or .30.

Measurement Invariance. Fewer parameters are needed to estimate within-person factors because the model is collapsed over time (‘multilevel approach’) rather than estimated repeatedly at each time-point (‘single-level approach’; Wright et al., 2016). However, it is not possible to test for conventional measurement invariance with the multilevel approach, i.e. the extent to which within-person change is driven by changes in measurement properties (e.g. differential item functioning or response biases) rather than the factors. Instead, parameters are assumed to be invariant and are modelled as such (see supplement C1 for background). Conventional measurement invariance tests were still ran despite not being directly translatable to the multilevel approach. In brief, all but the specific mood factor showed metric invariance (i.e. equal factor loadings between adjacent measurement waves), and all item thresholds showed scalar invariance apart from threshold B between waves one and two (see supplement C1).

5.2.5 Statistical Analysis

Multilevel Factor Analysis. Within-person psychopathology factors were estimated with multilevel confirmatory factor analysis (Muthén, 1994; 1991). Data were arranged with repeated observations in long-format (e.g., vertically) and multiple items in the wide format (e.g., horizontally; see Table 5.3). Each item was specified at the within-person level. Variances were not freed at the between-level but standard errors were corrected for nesting of observations within subjects using a subject ID cluster variable (see supplement C2 for full model details). Between-person variances were subsequently estimated in a multi-level growth model to investigate between-person differences in within-person change (see 'Multilevel Growth Model').

Table 5.3

Item Responses at Each Time-point Structured in 'Long' Format

ID	Time	Item 1	Item 2	…	Item 20

1	1	y_{11}	y_{11}	y_{11}
1	2	y_{12}	y_{12}	y_{12}
1	3	y_{13}	y_{13}	y_{13}
1	4	y_{14}	y_{14}	y_{14}
2	1	y_{21}	y_{21}	y_{21}
2	2	y_{22}	y_{22}	y_{22}
2	3	y_{23}	y_{22}	y_{23}
2	4	y_{24}	y_{24}	y_{24}
⋮	⋮	⋮	⋮	⋮
684	4	$y_{684\ 4}$	$y_{684\ 4}$	$y_{684\ 4}$

Note. The first subscript reflects the participant ID (1-684); the second subscript reflects 'time' or the assessment occasion (1-4). Therefore, y_{24} indicates participant two's response on a given item at the last time-point.

A within-person factor loading reflects the way in which responses on an item are predicted to increase or decrease over time by a factor estimated at the individual-level (e.g., an individual's expression of the *p* factor rather than relative differences in *p* factor scores). Higher positive factor loadings mean that as the expression of a factor increases in a given individual over time, their ratings on an item will also increase. This is derived from the within-person covariances between items, e.g., how item responses on a given item co-occur with other items over time for each individual. This contrasts the interpretation of a factor loading in standard single-level between-person analysis, which reflects how an item is predicted to co-occur with other items between individuals at a given time.

Model Comparison. Three within-level confirmatory factor models were estimated from the SDQ and MFQ item-level data: a single-factor model, correlated factors model, and bifactor model. The single-factor model included a single general factor upon which all items loaded. The correlated factors model included four factors reflecting antisociality (with loadings from SDQ conduct problem items), attention problems (with loadings from SDQ hyperactivity-inattention items), anxiety (with loadings from SDQ emotional symptom items), and mood (from MFQ mood problem items). These factors were identified in an exploratory within-level factor analysis and offered the best balance between model saturation and model fit ($\chi^2(116) = 1028$, $p < .001$, CFI = .95, TLI = .91, RMSEA = .06, SRMR = .04). Finally, a bifactor model was estimated with a general factor upon which all items loaded, and four uncorrelated specific factors, including antisocial, attention, anxiety, and mood factors. Correlations between the general and specific factors were constrained to zero. The bifactor model was also revised by examining theoretically plausible and substantial factor loadings (≥ .32; Tabachnick, Fidell, & Ullman, 2007) in a multilevel bifactor exploratory factor analysis, as Mplus does not provide modification indices for multilevel factor analysis. In all models, item overlap was accounted for by correlating the residuals for SDQ item 13 ('I am often unhappy') with MFQ item 1 ('I felt miserable/unhappy'), and SDQ item 2 ('I am restless') with MFQ item 4 ('I was very restless').

Models were estimated using the robust maximum-likelihood estimator and compared using the Akaike Information Criteria (AIC) and sample-size adjusted Bayesian Information Criteria (BIC). Bifactor models tend to overfit noise in the data, so it is important to use fit statistics that penalize for model complexity (Murray & Johnson, 2013). A difference of 2 (AIC/BIC) between models was considered negligible; a difference of 2-7 (AIC) or 2-6 (BIC) suggested some evidence favouring the competing model; a difference of 7-10 (AIC) or 6-10 (BIC) suggested strong evidence favouring the competing model, and a difference greater than 10 (AIC/BIC) suggested very strong evidence favouring the competing model (Raftery, 1995). The difference in BIC values was also formally tested with the Vuong test, a likelihood-based test statistic corrected for the number or freely estimated parameters in each model (Vuong, 1986; see section 4.2.5 for equation).

Models were also ran using the weighted least squares means and variances adjusted estimator (WLSMV) to assess global fit using the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). Acceptable and excellent fit, respectively, were defined by CFI values ≥ .90 and ≥ .95, TLI values ≥ .90 and ≥ .95, RMSEA values ≤ .08 and ≤ .06, and SRMR values ≤ .08 and ≤ .06 (Hu & Bentler, 1999). Models could not be compared using chi-square values because they were not nested. Therefore, we (cautiously) adopted the guidelines for comparing nested models: increases in CFI > ~.01 (and by generalization, TLI > ~.01), and decreases in RMSEA > ~.015 (and by generalization, SRMR > ~.015) between the more and less restricted models indicated a meaningful improvement in fit (Cheung & Rensvold, 2002).

Finally, models were assessed for their tendency to overfit the data with double cross-validation (Cudeck & Browne, 1983). The sample was randomly split into a calibration group and a test group. Parameters for the bifactor, correlated factor, or single-factor model were freely estimated in the calibration group and used to fix the parameters in the test group. Substantial differences between the calibration and test models, determined by critical differences in the information criteria described above, suggest that the model parameters are sensitive to peculiarities of the group used to estimate them, which is a symptom of overfitting (i.e. capitalizing on noise in the dataset). The process is then repeated, with participants who served as the calibration group now used as the test group and vice versa.

Factor Score Estimation. Within-level Bayesian Plausible Values (BPVs) were estimated for each adolescent at each time-point for the *p* factor and specific anxiety, mood, antisocial, and attention factors and used in a growth model (see 'Multilevel Growth Model'). BPVs were used instead of the latent variables themselves due to the computational issues in estimating both within-person bifactor dimensions and between-person growth curves. Since there are no established methods of evaluating BPVs, factor determinacy (FD) and information functions were estimated using the MLR and WLSMV models, respectively (see supplement C2 for background).

FD values (which reflect the reliability of estimated factor scores and range from 0-1; see section 3.2.3) for the *p* factor (.88) and specific anxiety (.77), antisocial (.75), attention (.83), and mood (.79) factors did not reach Gorsuch's (1983) recommended threshold of $FD \geqslant .90$, but were not far off, particularly for the *p* factor. Figure 5.1 displays the information curves (e.g., information summed across relevant items) for the *p* factor and specific antisocial, attention, anxiety, and mood factors (information reflects measurement precision at different levels of the latent variable and is the inverse of the standard error). The *p* factor showed the highest information, which mirrors the FD values but is more pronounced. Specific factors had relatively low information, which is typically the case when the general factor is 'essentially unidimensional' (Reise et al., 2010). In sum, factor scores for the *p* factor accurately represented individual differences on the latent variable, while factor scores for the specific factor scores lacked precision. It was therefore important to incorporate specific factor unreliability in the growth estimates using BPVs.

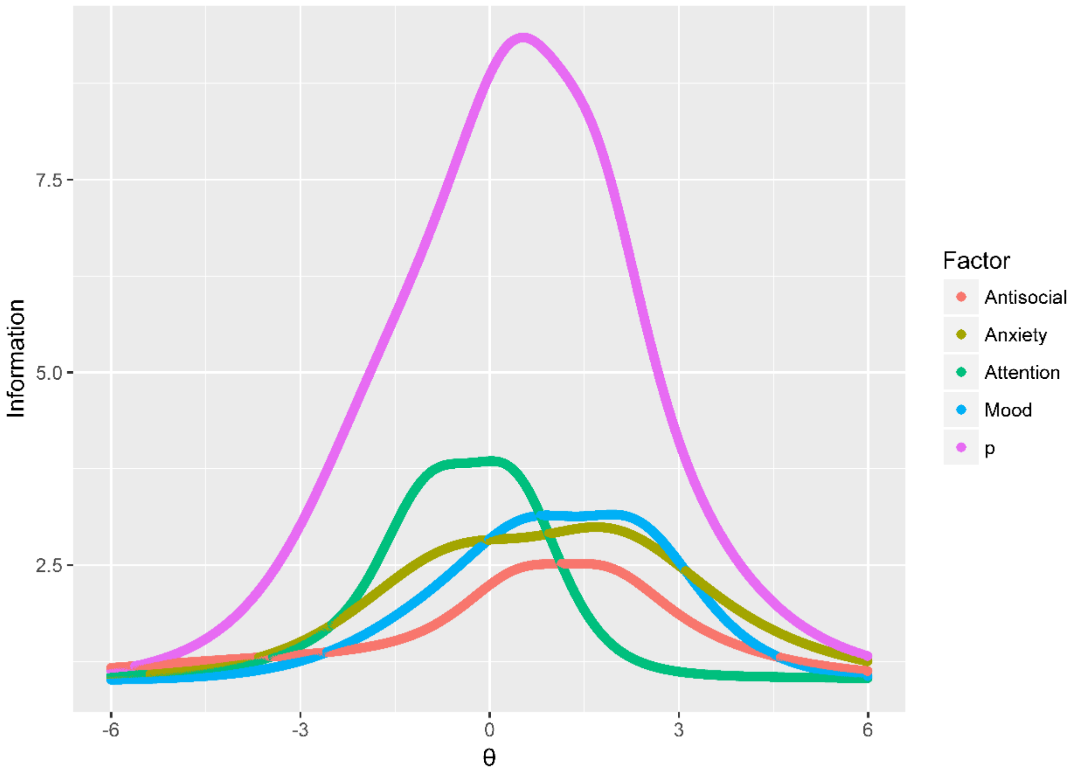

Figure 5.1. Information curves for the general p factor and specific antisocial, anxiety, attention, and mood factors. Higher information (y-axis) reflects lower standard errors and, hence, greater reliability at different levels of the latent trait (θ; x-axis). The 0 point reflects mean factor levels.

Multilevel Growth Model. Within-person changes in BPVs and between-person differences in within-person change were analysed using a parallel process multilevel growth model. Data were formatted with repeated estimates for each factor in long format (e.g., vertically) and each factor in wide format (e.g., horizontally; see Table 5.4).

Table 5.4

Bayesian Plausible Values (BPVs) Responses at Each Time-point Structured in 'Long' Format

ID	Time	θ_p	$\theta_{antisocial}$	$\theta_{anxiety}$	$\theta_{attention}$	θ_{mood}
1	0	y_{10}	y_{10}	y_{10}	y_{10}	y_{10}

1	1	y_{11}	y_{11}	y_{11}	y_{11}	y_{11}
1	2	y_{12}	y_{12}	y_{12}	y_{12}	y_{12}
1	3	y_{13}	y_{13}	y_{13}	y_{13}	y_{13}
2	0	y_{20}	y_{20}	y_{20}	y_{20}	y_{20}
2	1	y_{21}	y_{21}	y_{21}	y_{21}	y_{21}
2	2	y_{22}	y_{22}	y_{22}	y_{22}	y_{22}
2	3	y_{23}	y_{23}	y_{23}	y_{23}	y_{23}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
684	3	$y_{684\ 3}$	$y_{684\ 3}$	$y_{684\ 3}$	$y_{684\ 3}$	$y_{684\ 3}$

Note. Only one set of BPVs are depicted for demonstration purposes, but the analysis would be run and integrated over 100 datasets structured like this. The first subscript reflects the participant ID (1-684); the second subscript reflects 'time' or the assessment occasion (1-4 recoded as 0-3). Therefore, y_{23} indicates participant two's BPV on a given factor (θ) at the last time-point.

Within-level BPVs for each factor were regressed in parallel onto linear and quadratic time variables. Random intercepts, random linear slopes, and random quadratic slopes were estimated for each factor at the between level and co-varied within and between factors. A separate model with covariates predicting the random intercepts and slopes was run, including clinical covariates (e.g., treatment arm [MST vs. MAU], early vs. late onset conduct disorder, MST centre region [region 2 vs. 1, region 3 vs. 1]) and demographic covariates (e.g., baseline age, sex [boys vs. girls], composite scores on socio-economic indicators, full-scale IQ assessed with the Wechsler Abbreviated Scale of Intelligence, and ethnicity [White British vs. Other]). See supplement C2 for full model details.

Partially standardized regression coefficients were analyzed with two-tailed Wald tests (BPVs are standardized, i.e. they have a mean of 0 and variance of 1; hence, regression coefficients are partially standardized on the y-axis). Mean changes in BPVs should be interpreted as the change in standardized units of the factors with an increase in time, assuming equal item thresholds over time when factors are held constant and equal loadings.

A sensitivity analysis was run comparing the bifactor growth model to a correlated factors growth model. Initially, BPVs were derived for within-person correlated factors and regressed on linear and quadratic time variables in a multilevel growth model, but the model failed to converge as the between-person variances were collinear. Therefore, linear and quadratic growth curves were estimated at the within-level alone, providing an indication of intra-individual change but not inter-individual differences in intra-individual change (via random effects). Nonetheless, random intercepts were still estimated (i.e. variation in baseline time across adolescents) but were not predicted by covariates at the between-person level. Since within-person growth is less computationally taxing to estimate than within- and between-person growth, latent variables were used for the correlated factors rather than BPVs.

Reliability and Concurrent Validity of the Bifactor Dimensions. Model-based reliability estimates, including omega hierarchical (ω_H), omega hierarchical subscale (ω_{Hs}), explained common variance (ECV), and explained common variance subscale (ECVs), were calculated from the MLR within-level factor loading matrix using Dueber's (2017) bifactor indices calculator (see section 3.2.3 for definitions). Each reliability index ranges from 0-1, with ECV/ECV_s values ⩾.70 reflecting that the most of the common (i.e. modelled) variance is explained by the general/specific factor, and ω_H/ω_{Hs} ⩾. 80 reflecting that most variance raw total/subscale scores is explained by the general/specific factor (Rodriguez et al., 2016a). Furthermore, the mean parameter change (MPC) and standard deviation of the parameter change were computed to determine the extent that factor loadings for a given disorder decreased (positive MPC values) or increased (negative MPC values) from the correlated factors model to the bifactor models, and hence included more or less common variance.

The validity of within-person variation in the general and specific factor BPVs was assessed by regressing official records of the number of offences and school exclusions over time onto BPVs at the within-level in the multilevel growth model. The validity of within-person changes in general and specific BPVs was assessed by regressing offences and school exclusions onto each factor, a linear time variable (which captures change in the outcome variable over time), and time*factor interactions (which captures how changes in the factors over time predicted changes in the external outcomes) at the within-person level using Poisson multilevel growth models. Between-person random slopes for offences and school exclusions required too many levels of integration when estimated with a Poisson estimator, so analyses were limited to the within-person level.

5.3 Results

5.3.1 Model Comparison

Model fit indices are presented in Table 5.5 and factor loadings are presented in Table 5.6. The single-factor model showed healthy loadings from all items but fit the data poorly. The correlated factors model, with anxiety, mood, antisocial, and attention factors, approached an acceptable fit that explained more information than the single-factor model (ΔAIC = 2047, ΔBIC = 2013, ΔaBIC = 2032, z = 42.08, p < .001; ΔCFI = .20, ΔTLI = .22, ΔRMSEA = -.05, ΔSRMR = -.04). Factors were positively and moderately-to-strongly correlated with each other suggesting the presence of an overarching factor.

The bifactor model included a general (‘*p*’) factor upon which all items loaded, as well as four uncorrelated specific factors (anxiety, mood, antisocial, and attention). Like the correlated factors model, fit indices were almost acceptable but the former fit slightly better (ΔAIC = 14, ΔBIC = 94, ΔaBIC = 49, *z* = -0.68, *p* > .05; ΔCFI = -.03, ΔTLI = -.05, ΔRMSEA = .02, ΔSRMR = .00). The bifactor model still fit better than the single factor model (ΔAIC = 2033, ΔBIC = 1919, ΔaBIC = 1983, *z* = 41.40, *p* < .001; ΔCFI = -.17, ΔTLI = -.17, ΔRMSEA = .03, ΔSRMR = .04).

Four items in a multilevel bifactor EFA showed substantial and theoretically plausible cross-loadings (see Table C2 for factor loadings). Two items from the attention factor and one from the antisocial factor cross-loaded onto the anxiety factor: SDQ item 7 (‘I [do not] usually do as I am told’, λ = -.46), item 21 (‘I [do not] think before I do things’, λ = -.33) and item 25 (‘I [do not] finish the work I am doing’, λ = -.33). These items reflect behavioural control which positively co-occurs with internalizing problems after controlling for general psychopathology (Hankin et al., 2017; Neumann et al., 2016). Moreover, SDQ item 16 from the anxiety factor (‘I am nervous in new situations’) negatively cross-loaded onto the antisocial factor (λ = -.32), which is expected if the specific antisocial factor overlaps with fearlessness (Lahey et al., 2017).

A revised bifactor model in which these cross-loadings were freed fit the data well and better than the standard bifactor model (ΔAIC = 429, ΔBIC = 406, ΔaBIC = 419, *z* = 8.73, *p* < .001; ΔCFI = .07, ΔTLI = .09, ΔRMSEA = -.03, ΔSRMR = -.02), correlated factors model (ΔAIC = 415, ΔBIC = 312, ΔaBIC = 370, *z* = 8.06, *p* < .001; ΔCFI = .04, ΔTLI = .04, ΔRMSEA = -.01, ΔSRMR = -.02), and single factor model (ΔAIC = 2462, ΔBIC = 2325, ΔaBIC = 2402, *z* = 8.06, *p* < .001; ΔCFI = -.24, ΔTLI = -.26, ΔRMSEA = -.06, ΔSRMR = -.06). Therefore, BPVs were estimated for the revised bifactor model. Cross-validation tests demonstrated that all models differed substantially between the calibration and test groups and were therefore sensitive to sample-specific characteristics (see Table 5.7).

Table 5.5

Model Fit Values for the Within-Person Single-Factor, Correlated Factor, and Bifactor Models (Standard and Revised)

	Fit Statistic								
Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR	AIC	BIC	aBIC
Single Factor	5,414	168	.69	.65	.12	.11	78,975	79,316	79,126
Correlated Factors	2,013	162	.89	.87	.07	.07	76,928	77,303	77,094
Bifactor	2,558	148	.86	.82	.09	.07	76,942	77,397	77,143
Bifactor (revised)	1,466	144	.93	.91	.06	.05	76,513	76,991	76,724

Note. χ^2 = chi-square statistic; aBIC = sample size adjusted Bayesian information criterion; CFI = Comparative Fit Index; *df* = degrees of freedom; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Error.

Table 5.6

Standardized Within-Person Factor Loadings for the Single Factor, Correlated Factor, and Bifactor Models (Standard and Revised)

		Correlated Factors				Bifactor (Standard)		Bifactor (Revised)				
ASR Item	1-Fac	ANX	ANTI	ATT	MD	*p*	Specific[a]	*p*	ANX	ANTI	ATT	MD
SDQ												
3. Get a lot of headaches	0.52	0.63				0.49	0.34	0.53	0.33			
8. Worry a lot	0.55	0.70				0.46	0.63	0.56	0.48			
13. Often unhappy	0.65	0.80				0.62	0.40	0.68	0.29			
16. Nervous in new situat.	0.46	0.57				0.42	0.38	0.53	0.24	-0.32		
24. Many fears	0.45	0.57				0.34	0.59	0.44	0.45			
5. Get very angry	0.61		0.78			0.67	0.22	0.65		0.25		
7. Do not do as told	0.34		0.46			0.35	0.29	0.38	-0.48	0.26		
12. Fight a lot	0.40		0.54			0.37	0.57	0.37		0.61		
18. Lying or cheating	0.48		0.60			0.48	0.35	0.48		0.33		
22. Take things	0.31		0.42			0.27	0.55	0.28		0.50		
2. Restless	0.61			0.74		0.47	0.64	0.45			0.68	
10. Constantly fidgeting	0.64			0.78		0.51	0.63	0.49			0.66	
15. Easily distracted	0.62			0.76		0.55	0.48	0.54			0.47	
21. Do not think before do	0.41			0.54		0.39	0.27	0.45	-0.53		0.16	
25. Do not finish work	0.31			0.42		0.28	0.28	0.33	-0.40		0.16	
MFQ												

1. Felt miserable/unhappy	0.63				0.74	0.54	0.47	0.52				0.49
2. Didn’t enjoy anything	0.57				0.67	0.45	0.60	0.42				0.61
3. So tired I did nothing	0.49				0.58	0.41	0.47	0.38				0.49
4. Very restless	0.58				0.64	0.52	0.35	0.48				0.40
5. Felt no good anymore	0.73				0.86	0.66	0.50	0.63				0.53
Mean	0.52	0.65	0.70	0.56	0.65	0.46	See	0.48	0.40	0.38	0.43	0.50
Standard Deviation	0.12	0.10	0.15	0.14	0.16	0.11	Notes[b]	0.11	0.10	0.14	0.26	0.08
ECV/ECV$_s$	–	–	–	–	–	0.51		0.50	0.14	0.10	0.12	0.14
ω/ω_s	–	–	–	–	–	0.91		0.92	0.85	0.79	0.82	0.83
ω_H/ω_{Hs}	–	–	–	–	–	0.73		0.71	0.34	0.31	0.35	0.43
Relative Omega	–	–	–	–	–	0.81		0.78	0.40	0.40	0.43	0.52
H	–	–	–	–	–	0.87		0.87	0.63	0.57	0.66	0.64
FD	–	–	–	–	–	0.88		0.89	0.77	0.75	0.83	0.79
Factor Correlations		ANX	ANTI	ATT	MD							
	ANX	–										
	ANTI	0.43	–									
	ATT	0.43	0.72	–								
	MD	0.69	0.52	0.39	–							

Note. 1-Fac = Single factor model; ANTI = Antisocial; ANX = Anxiety; ATT = Attention; ECV/ECV$_s$ = Explained Common Variance/ Explained Common Variance-Subscale; FD = Factor Determinacy; MD = Mood; ω/ω_s = Omega/Omega-subsale; ω_H/ω_{Hs} = Omega hierarchical/Omega hierarchical-subscale. All loadings are significant at $p < .001$.

[a]Factor loadings for each factor are listed sequentially to save space. Factor loadings for items 3-24 represent the specific anxiety factor, loadings for items 5-22 represent the specific antisocial factor, loadings for items 2-25 represent the specific attention factor, and items 1-5 represent the mood factor.

[b]Mean factor loadings for the specific factors were as follows: anxiety = 0.47 (*SD* = 0.13), antisocial = 0.40 (*SD* = 0.16), attention = 0.46 (*SD* = 0.18), and mood = 0.48 (*SD* = 0.09). Model-based reliability indices for the specific factors were as follows: anxiety (ECV_s = 0.13, ω_s = 0.80, ω_{Hs} = 0.40, Relative ω = 0.50, H = 0.63, FD = 0.79), antisocial (ECV_s = 0.10, ω_s = 0.73, ω_{Hs} = 0.34, Relative ω = 0.46, H = 0.55, FD = 0.73), attention (ECV_s = 0.13, ω_s = 0.78, ω_{Hs} = 0.41, Relative ω = 0.52, H = 0.65, FD = 0.81), and mood (ECV_s = 0.13, ω_s = 0.83, ω_{Hs} = 0.39, Relative ω = 0.46, H = 0.62, FD = 0.72).

Table 5.7

Differences in Information Criteria Between the Calibration and Test Groups for Each Within-Person CFA Model

	Order A			Order B		
Model	ΔAIC	ΔBIC	ΔaBIC	ΔAIC	ΔBIC	ΔaBIC
Single Factor	1493	1794	1603	1585	1286	1477
Correlated Factors	1289	1621	1411	1401	1073	1282
Bifactor	1325	1726	1472	1464	1065	1319
Bifactor (revised)	1361	1782	1515	1497	1079	1346

Note. aBIC = sample size adjusted Bayesian information criterion. Order A and B reflect the sequence that each half of the sample was allocated as the calibration or test group. Negative values indicate that the calibration sample showed a lower (better) fit compared to the test sample.

5.3.2 Reliability and Concurrent Validity of the Bifactor Dimensions

Reliability. The common variance in the revised bifactor model was equally split between the *p* factor (ECV = .50) and specific factors (total ECV_s = .50). Therefore, both general and specific factors were needed to represent the multidimensional latent structure. By contrast, the variance in raw total scores was largely explained by the *p* factor (ω_H = .72). All items showed moderate-to-strong loadings on the *p* factor, except for SDQ item 22 'I take things that are not mine' (λ = .28). SDQ items 13 'I am often unhappy' (λ = .68), item 5 'I get very angry' (λ = .65), and MFQ item 5 'I felt I was no good anymore' (λ = .63) showed the strongest *p* factor loadings.

The specific factors each explained roughly one-third of the variance in raw subscale scores (ω_H = .32-.43); therefore, the inter-relatedness between specific groups of items was mainly explained by the *p* factor. The specific anxiety factor (ω_H = .34) showed the highest mean parameter change (*M* = .30, *SD* = .14), with SDQ items 13 'I am often unhappy' (Δλ = .51), item 16 'I am nervous in new situations' (Δλ = .33), and item 3 'I get a lot of headaches' (Δλ = .3) showing the largest decline in loading strength from the anxiety factor in the correlated factor model to the specific anxiety factor in the bifactor model (and hence, were most influenced by the common variance). Items that loaded most strongly onto the specific anxiety factor included SDQ item 21 'I think before I do things' (λ = .53), item 7 'I usually do as I am told' (λ = .48), and item 8 'I worry a lot' (λ = .48).

The specific antisocial factor (ω_H = .32) showed the lowest but most variable mean parameter change (*M* = .17, *SD* = .26). SDQ item 5 'I get very angry' showed a large decline in loading strength between the correlated factor and bifactor models ($\Delta\lambda$ = .53), along with items 18 'I often get accused of lying or cheating' ($\Delta\lambda$ = .27) and item 7 'I [do not] usually do as I am told' ($\Delta\lambda$ = .2). By contrast, SDQ items 12 'I fight a lot' ($\Delta\lambda$ = -.07) and 22 'I take things that are not mine' ($\Delta\lambda$ = -.08) loaded slightly higher onto the specific antisocial factor compared to the correlated factors antisocial factor, meaning they were least affected by the common variance. Consequently, SDQ items 12 (λ = .61) and 22 (λ = .50) loaded most strongly onto the specific antisocial factor.

The specific attention factor (ω_H = .35) and mood factor (ω_H = .43) showed moderate mean parameter changes (attention = .22, *SD* = .13; mood = .19, *SD* = 11). SDQ items 21 'I do not think before I do things' ($\Delta\lambda$ = .38) and item 2 'I am restless' ($\Delta\lambda$ = .06) showed the largest and smallest decrease in loading strength, respectively, from the correlated attention factor to the specific attention factor. Furthermore, SDQ items 2 'I am restless' (λ = .68) and item 10 'I am constantly fidgeting' (λ = .66) showed the strongest specific attention factor loadings. MFQ items 33 'I felt I was no good anymore' ($\Delta\lambda$ = .33) and item 2 'I didn't enjoy anything at all' ($\Delta\lambda$ = .06) showed the largest and smallest decrease in loading strength, respectively, from the correlated mood factor to the specific mood factor. Furthermore, MFQ items 2 'I didn't enjoy anything' (λ = .61) and item 5 'I felt I was no good anymore' (λ = .53) showed the strongest specific mood factor loadings.

Concurrent Validity. Within-person variability in antisocial BPVs positively predicted variability in the number of offences committed (β = .12, p = .043, 95% CI [.01, .24]). That is, higher (or lower) antisocial scores co-occurred with more (or less) offences for each adolescent over time. Moreover, within-person variability in anxiety BPVs negatively predicted variability in the number of school exclusions (β

= -.13, p = .040, 95% CI [-.25, -.01]). Therefore, higher (or lower) anxiety scores co-occurred with less (or more) exclusions for each adolescent over time.

The number of school exclusions significantly declined over time (β = -.14, p = .037, 95% CI [-.26, -.01]), as well as the number of offences, albeit marginally (β = -.11, p = .052, 95% CI [-.21, .00]). The only factor whose effect over time predicted within-person changes in the external outcomes was *p*. Specifically, reductions in *p* (see 'Multilevel Growth Model' for slope) positively predicted the number of offences committed (β = .13, p = .012, 95% CI [.03, .22]). In other words, decreases in *p* predicted decreases in offences. Moreover, reductions in *p* marginally predicted the number of exclusions (β = .06, p = .064, 95% CI [.00, .22]). That is, decreases in *p* marginally predicted decreases in school exclusions.

5.3.3 Multilevel Growth Model

Within-person change. Figure 5.2 shows the observed and predicted within-person growth curves pooled across adolescents for the *p* factor and specific anxiety, mood, antisocial, and attention factor BPVs. Both the *p* factor (β = -.28, *p* < .001, 95% CI [-.41, -.16]) and specific antisocial factor (β = -.27, *p* = .002, 95% CI [-.43, -.10]) decreased over time for each adolescent. Furthermore, the *p* factor (β = .03, *p* = .087, 95% CI [-.01, .07]), but not the antisocial factor (β = .01, *p* = .836, 95% CI [-.04, .05]), showed a marginally significant quadratic growth term. That is, within-person decline in the *p* factor decelerated during the follow-up period. The specific anxiety factor showed a significant linear increase over time for each adolescent (β = .17, *p* = .021, 95% CI [.03, .32]), which occurred at a steady pace (quadratic slope: β = .00, *p* = .984, 95% CI [-.05, .05]). Finally, the specific mood factor did not deviate from baseline (β = -.06, *p* = .42, 95% CI [-.20, .08]), while the specific attention factor maintained an elevated level throughout (linear slope: β = -.01, *p* = .89, 95% CI [-.12, .14]).

Between-person differences in within-person change. Adolescents significantly varied in their baseline BPVs for all bifactor dimensions, but not in their rate of change over time (see Table 5.8 for random intercepts and slopes). Moreover, the correlations among random effects within and between factors were weak and did not reach significance. In a model with clinical and demographic covariates predicting the random intercepts, linear slopes, and quadratic slopes, adolescent girls showed higher baseline *p* factor scores than boys (β = .54, *p* < .001, 95% CI [.36, .71]), while White British adolescents showed marginally lower baseline *p* factor scores (β = -.21, *p* = .068, 95% CI [-.44, -.02]), but caution is warranted since the majority of adolescents were White British boys. Treatment arm did not predict between-person differences in any factor at baseline or over time (see Table C3 for coefficients).

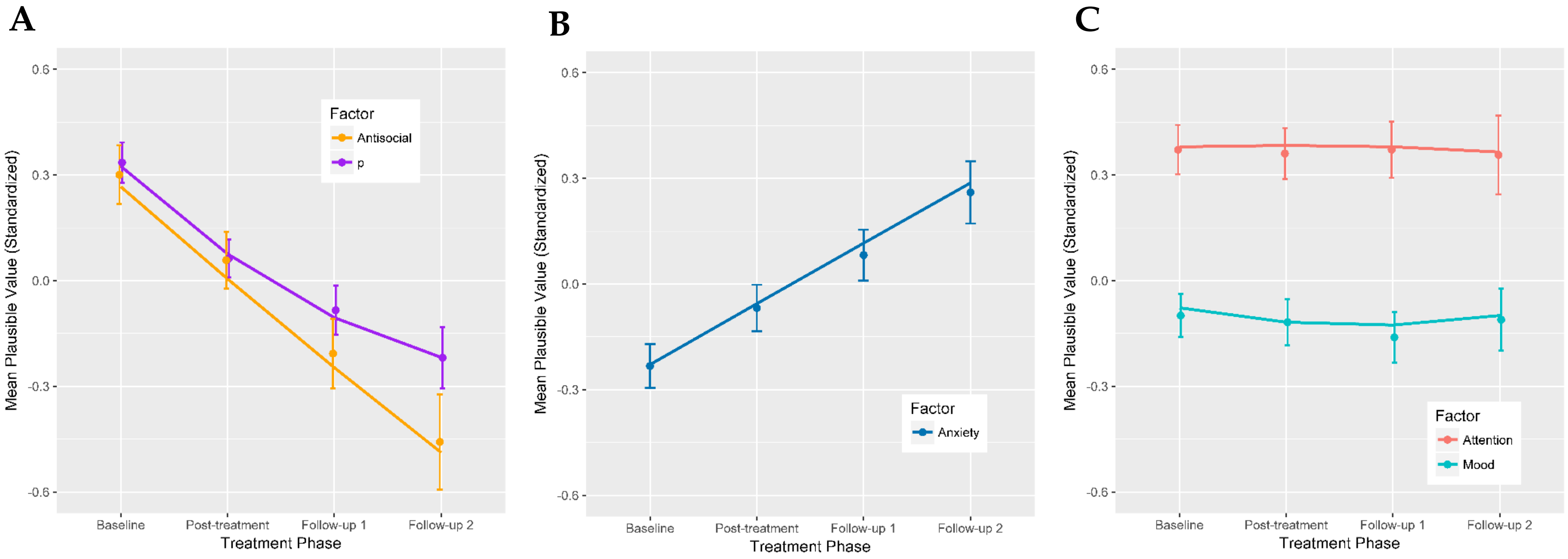

Figure 5.2. Average predicted trajectories (curves) and observed means (data points with error bars) for the bifactor growth model, including (A) general psychopathology and specific antisocial factors, (B) specific anxiety factor, and (C) specific mood and attention factors. The zero-point on the Y-axis reflects the standardized factor mean. Error bars indicate 95% CIs.

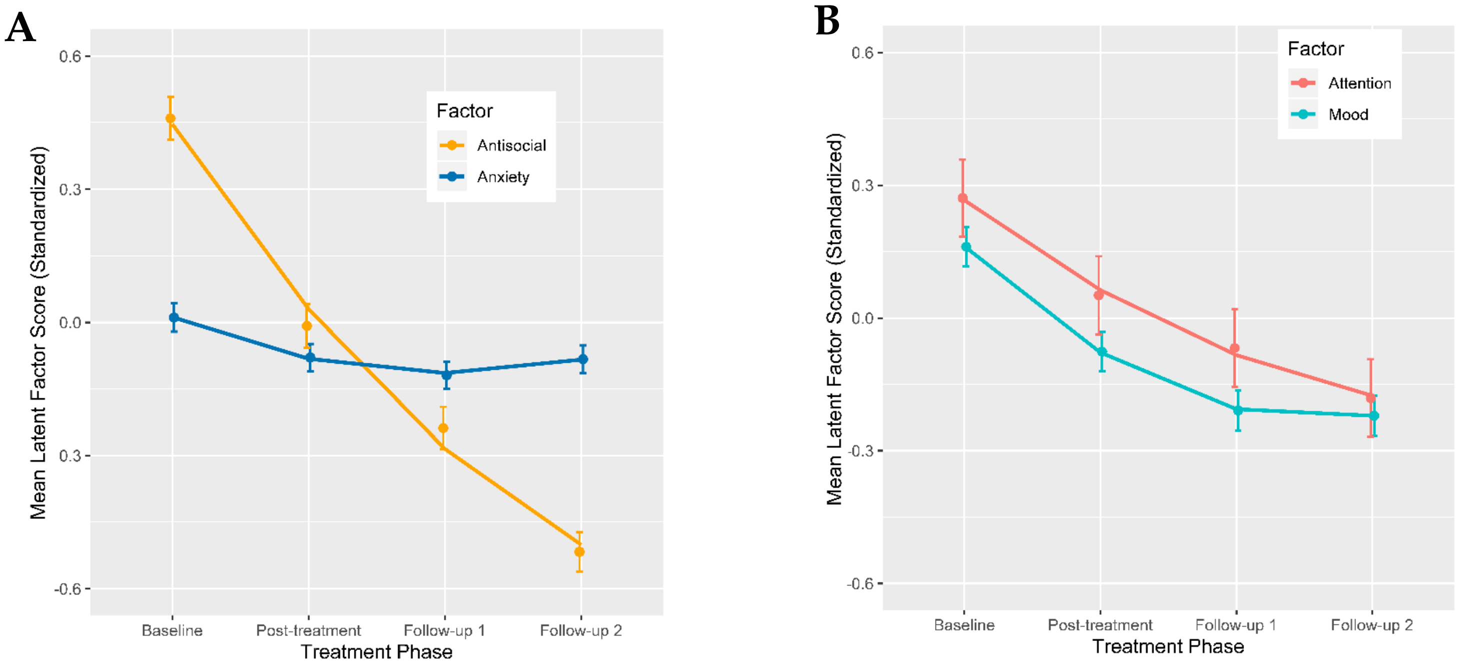

Figure 5.3. Average predicted trajectories (curves) and observed means (data points with error bars) for the correlated factors model, including (A) the antisocial and anxiety factors, and (B) attention and mood factors. The zero-point on the Y-axis reflects the standardized factor mean. Error bars indicate 95% CIs.

Table 5.8

Correlations Between Random Intercepts, Random Linear Slopes, and Random Quadratic Slopes for the General (p) and Specific Psychopathology Factors

R. Effect	1.	2.	3.	4.	5.	6	7.	8.	9.	10.	11.	12.	13.	14.	15.
1. $U_{0i}^{(p)}$	0.38***														
2. $U_{1i}^{(p)}$	-0.11	0.26													
3. $U_{2i}^{(p)}$	0.02	-0.08	0.03												
4. $U_{0i}^{(anxiety)}$	-0.02	0.02	0.00	0.22***											
5. $U_{1i}^{(anxiety)}$	0.07	-0.07	0.02	-0.14	0.27										
6. $U_{2i}^{(anxiety)}$	-0.01	0.02	-0.01	0.03	-0.08	0.03									
7. $U_{0i}^{(mood)}$	-0.01	0.07	-0.02	0.06	-0.05	0.01	0.16*								
8. $U_{1i}^{(mood)}$	0.09	-0.14	0.04	-0.04	0.09	-0.02	-0.14	0.27							
9. $U_{2i}^{(mood)}$	-0.02	0.04	-0.01	0.01	-0.02	0.01	0.03	-0.08	0.03						
10. $U_{0i}^{(anti)}$	-0.01	0.05	-0.01	-0.03	-0.01	0.00	-0.01	0.00	0.00	0.16*					
11. $U_{1i}^{(anti)}$	0.04	-0.07	0.02	-0.01	0.02	-0.01	0.03	-0.01	0.00	-0.14	0.31				
12. $U_{2i}^{(anti)}$	-0.01	0.02	-0.01	0.01	-0.01	0.00	-0.01	0.00	0.00	0.03	-0.09	0.03			
13. $U_{0i}^{(atten)}$	0.02	0.06	-0.02	-0.04	0.02	0.00	-0.03	0.02	-0.01	0.02	-0.03	0.01	0.24***		
14. $U_{1i}^{(atten)}$	0.00	-0.11	0.04	-0.01	0.05	-0.02	0.00	-0.01	0.00	0.00	0.05	-0.02	-0.17	0.33	
15. $U_{2i}^{(atten)}$	0.00	0.03	-0.01	0.00	-0.02	0.01	0.00	0.00	0.00	0.00	-0.01	0.01	0.04	-0.09	0.03

Note. Variances are on the diagonal. anti = specific antisocial factor; atten = specific attention factor; p = general psychopathology; U_{0i} = random intercept; U_{1i} = random linear slope; U_{2i} = random quadratic slope. $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

A sensitivity analysis was run of within-person change in the anxiety, mood, antisocial, and attention factors from a correlated factors model. Each factor is a weighted equivalent of a disorder-specific subscale that is uncontrolled for the common variance. Like in the primary analysis using disorder-specific subscales, all correlated factors declined over the study period. The antisocial factor showed the strongest linear decline ($\beta = -.44$, $p < .001$, 95% CI [-.57, -.31]), followed by the mood factor ($\beta = -.29$, $p < .001$, 95% CI [-.41, -.18]), but decline in the mood factor slowed with time (quadratic slope: $\beta = .06$, $p = .005$, 95% CI [.02, .09]). The attention and anxiety factors also showed significant linear declines ($\beta_{\text{attention}} = -.23$, $p < .001$, 95% CI [-.35, -.11]; $\beta_{\text{anxiety}} = -.13$, $p = .039$, 95% CI [-.24, -.01]; see Figure 5.3).

In another sensitivity analysis, the growth models were rerun using BPVs estimated from a bifactor model without cross-loadings (see Table C4 for factor loadings). The purpose was to determine the influence of cross-loadings on the direction and significance of the growth curves, particularly for the specific anxiety and antisocial factors, since the decline in antisocial scores might have been driven by an increase in the negatively weighted anxiety item that cross-loaded. Similarly, anxiety scores might have increased because of a decrease in the negatively weighted antisocial item or attention items which cross-loaded.

Figure C1 shows the observed and predicted within-person growth curves pooled across adolescents for the *p* factor and specific anxiety, mood, antisocial, and attention factor BPVs from a model without cross-loadings. The specific anxiety factor continued to show a significant linear increase over the study period ($\beta = .34$, $p < .001$, 95% CI [.18, .51]). The increase was stronger in magnitude than in the model with cross-loadings, most likely because of SDQ item 16's boost in loading strength on the anxiety factor after no longer cross-loading on the antisocial factor.

Therefore, it does not appear that the antisocial and attention items that cross-loaded on the anxiety factor drove its increase over time.

By contrast, the antisocial factor still declined over the study period (β = -.05, *p* = .614, 95% CI [-.22, .13]) but at a weaker magnitude that was no longer significant. Hence, it appears that the negatively weighted SDQ item 16 ('I am [not] nervous in new situations') contributed to the decline in antisocial scores. As for the other factors, the *p* factor continued to decline over time (β = -.47, *p* < .001, 95% CI [-.60, -.34]), which, like the anxiety factor, was stronger in magnitude than the model featuring cross-loadings. Removing the cross-loadings might have strengthened changes in the general variance because the *p* factor absorbs unmodelled covariance (Murray & Johnson, 2013). Moreover, the quadratic slope for the *p* factor was now significant, albeit just (β = .04, *p* = .045, 95% CI [.01, .08]). The mood (β = -.04, *p* = .638, 95% CI [-.21, .13]) and attention (β = .02, *p* = .779, 95% CI [-.12, .16]) factors both continued to show little change over time.

5.4 Discussion

Clinical outcomes are typically assessed and analysed by disorder, but the high rate of comorbidity between disorders suggests that this is confounded by characteristics shared across disorders. As a solution to this issue, the current study separated out the variance in clinical outcomes attributed to specific disorders from the variance shared across disorders. A bifactor model with a general *p* factor and specific antisocial, anxiety, mood, and attention factors, summarized the covariation best among behavioural and emotional symptoms in adolescents undergoing a psychosocial intervention for conduct problems. Furthermore, the *p* factor and specific antisocial factor declined over the intervention and follow-up period, while the specific anxiety factor increased. The mood and attention factors showed little change. The antisocial and anxiety factors predicted within-person variation in the number of offences and school exclusions recorded over the study period, respectively, while changes in the *p* factor predicted declines in both outcomes. Finally, girls showed higher *p* factor scores at baseline and less decline over the study period.

5.4.1 Which Model Summarized the Within-Person Covariation in Symptoms Best?

Within-person covariation among behavioural and emotional problems was best summarized by a revised bifactor model with a *p* factor and uncorrelated specific factors (including antisocial, attention, anxiety, and mood with cross-loadings) compared to a standard bifactor model without cross-loadings, a correlated factors model (with antisocial, attention, anxiety, and mood factor), and a single-factor model. Therefore, a single dimension explained the positive co-occurrences in symptoms experienced by a given adolescent over time, as well as specific factors capturing shared characteristics of groups of symptoms. In other words, increases (or decreases) in the way that a given adolescent rated symptom X were accompanied by increases (or decreases) in their ratings on symptom Y (with some symptoms co-occurring more strongly than other).

Most bifactor studies of psychopathology in adolescents describe between-person differences in symptom covariation (Carragher et al., 2016; Castellanos-Ryan et al., 2016; Laceulle et al., 2016; Lahey et al., 2015; Neumann et al., 2016; Snyder et al., 2017; Stochl et al., 2015; Tackett et al., 2013). The current study extends this literature by supporting bifactor model’s applicability to within-person levels of analysis (Kim & Eaton, 2017). There is also evidence that personality disorder symptoms positively co-occur within adults over time, which can be explained by a general *pd* factor (Wright et al., 2016). Nonetheless, it is possible that the within-person positive manifold in symptoms demonstrated in the current study and Wright et al.'s study was an artifact of undergoing an intervention, which biased symptoms in a similar direction of change (e.g., decline). If this were the case, however, we would also expect to see the specific factors follow a similar pattern of change, yet they diverged.

The bifactor model showed an improvement in fit after the cross-loadings between specific factors were freed. In fact, the standard bifactor without cross-loadings showed a slightly poorer fit than the correlated factors model, both of which were suboptimal according to conventional fit criteria (Hu & Bentler, 1999). This is surprising because the bifactor model included more freely estimated parameters than the correlated factors model, which should give it an advantage in maximizing the likelihood function (Brown, 2014). Nonetheless, several studies have reported that the bifactor model shows a near-equivalent or slightly better fit than the correlated factors model (Brodbeck et al., 2014; Caspi et al., 2014; Conway et al., 2019; Deutz et al., 2018; Lahey et al., 2012; Fernandez de la Cruz et al., 2018; Gomez et al., 2019; Haltigan et al., 2018; Laceulle et al., 2016; Liu et al., 2017; Miller et al., 2019; Patalay et al., 2015; Pettersson et al., 2018; Snyder et al., 2017; St Clair et al., 2017; Tackett et al., 2013; Weissman et al., 2019), most likely because they imply similar variance-covariance matrices (van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017). The near-equivalent fit, in addition to the unmodelled covariances (i.e. cross-loadings) that weaken the bifactor model's fit when mis-specified (Greene et al., 2019), likely contributed to the standard bifactor model's poorer fit relative to the correlated factors model.

There is still some concern over freeing cross-loadings in the bifactor model (see section 3.5.6). For example, shared variance beyond the *p* factor risks the interpretation of general and specific factors and implies that the model is mis-specified (Markon, 2019). Furthermore, cross-loadings are easily accommodated by the bifactor model due to its high fitting propensity; the improved fit observed after freeing the cross-loadings might be due to statistical rather than substantive reasons (Greene et al., 2019). However, cross-loadings were freed because they supported the personality style interpretation of specific factors (see below; Caspi et al., 2014). Freeing these cross-loadings might have also prevented an inflation of *p* factor loadings (Reise, Moore, & Maydeu-Olivares, 2011).

5.4.2 How Reliable were the Bifactor Dimensions?

The revised bifactor model showed a multidimensional data structure, with a 50-50 split in the common variance between the general and specific factors. Nonetheless, the *p* factor explained the majority of variance in raw total scores (72%), while specific factors failed to explain more than 43% of the variance in raw subscale scores. Therefore, the latent structure of psychopathology at the within-person level was multidimensional–requiring both general and specific dimensions–while its measurement was attributable to a single dimension, similar to between-person analyses (see Chapter 3 and 4).

The *p* factor showed the highest loadings from items associated with depression, including SDQ item 13 'I am often unhappy', item 5 'I get very angry[17]', and MFQ item 5 'I felt I was no good anymore'. Between-person analyses have also shown that depression symptom/disorder ratings load most strongly onto the *p* factor in adolescents (Calkins et al., 2015; Carragher et al., 2016; Haltigan et al., 2018; Jones et al., 2018; Lahey et al., 2015; Liu, Mustanski, Dick, Bolland, & Kertes, 2017; Rytila-Manninen et al., 2016; Preti, Carta, & Petretto, 2019; Schaefer et al., 2018; Tackett et al., 2013; Wade, Fox, Zeanah, & Nelson, 2018; see also Chapter 4). Therefore, the *p* factor might reflect depression as a global cause or consequence of co-occurring mental health problems (see section 2.2). The fact that within-person covariation in symptoms can be summarized by the overall level of distress in a diagnostically homogeneous group highlights the limits of defining a sample categorically.

Many items associated with anxiety and mood preferentially loaded onto the *p* factor rather than the specific anxiety factor. A similar finding was reported in Chapter 4, where anxious-depressed items in the ASR loaded preferentially onto the *p* factor, leaving behind an imprecise specific internalizing factor. In the current study, the specific anxiety showed strong loadings (and cross-loadings) from items associated with inhibition, compliance, and worry. Therefore, removing the common variance from the anxiety factor might have created a compensatory boost in variance associated with behavioural control and conscientiousness. This would explain why within-person variability in anxiety scores negatively but modestly

[17] Anger and irritability are commonly featured in depressed patients (Judd, Schettler, Coryell, Akiskal, & Fiedorowicz, 2013). It might therefore be more appropriate to discuss problems in 'dysregulation', which is explicit in covering a range of emotional states, rather than 'depression'.

predicted official records of school exclusions over time. The specific internalizing factor in between-person analyses also modestly predicts better academic performance and attendance (Lahey et al., 2015; Patalay et al., 2015; Sallis et al., 2019).

Some conduct problem items loaded preferentially onto the *p* factor while others loaded preferentially onto the antisocial factor. As described above, anger loaded preferentially onto the *p* factor, as did items associated with lying and disobedience, albeit to a lesser extent. By contrast, items associated with fighting and stealing loaded most strongly onto the specific antisocial factor and hence were least affected by the common variance. This pattern of loadings, including the way that nervousness negatively cross-loaded onto the specific antisocial factor, suggests that antisociality overlaps with callousness and fearlessness after removing the common variance (Lahey et al., 2017). Moreover, within-person variability in the antisocial factor positively predicted official records of criminal activity over time, although the association was modest. It is, however, uncertain why lying and cheating did not also load preferentially onto the antisocial factor, as these are characteristics of callous-unemotional traits (Frick, Ray, Thorton, & Kahn, 2013).

Attention-deficit hyperactivity items associated with disinhibition loaded preferentially onto the *p* factor, while items associated with attention problems loaded more strongly onto the specific attention factor. Attention problems are typically collapsed with hyper-activity/impulsivity, but they might be more distinct after controlling for the common variance. Mood items decreased in loading strength on average between the correlated factors model and bifactor model, but generally maintained healthy loadings on the specific mood factor. This might, in part, be a result of a subscale effect; MFQ items might have banded together due to their common method variance which is separate from the SDQ's common variance. Neither attention nor mood specific factors predicted external outcomes, limiting their interpretability as substantive factors.

5.4.3 How did the Bifactor Dimensions Change Over a Psychosocial Intervention?

p factor scores for each adolescent declined on average over the intervention and follow-up period and predicted reductions in criminal offences and school exclusions. Therefore, MST and MAU influenced a range of behavioural and emotional problems that co-occurred to varying degrees in each adolescent, despite targeting antisocial behaviour. While this is the first study to investigate changes in the *p* factor over a psychosocial intervention, prior studies suggest that family-based interventions for conduct problems influence a broad range of internalizing and externalizing problems, and hence, general psychopathology. For example, various family-based intervention studies for antisocial behaviour report declines in both internalizing and externalizing symptoms following treatment (Henggeler, Melton, Brondino, Scherer, & Hanley 1997; Huey et al., 2004; Hogue, Dauber, Samuolis, & Liddle, 2006; Ogden & Hagen, 2006; Rowland et al., 2005). While changes in internalizing scores are usually weaker (Bearman & Weisz, 2015; Riosa, McArthur, & Preyde, 2011; Sundell et al., 2008) or sometimes absent (Butler et al., 2011; Goodman, 2010; Henggeler et al., 1999; Letoureau et al., 2009; van der Stouwe et al., 2014), the stronger decline in externalizing problems might reflect both common and specific treatment effects. Indeed, this was observed when longitudinal change was analysed with correlated factors that do not control for the common variance: widespread declines in all symptom domains, but particularly strong decline in the antisocial domain.

Winiarski et al. (2017) found that increases in physiological markers of emotion dysregulation over the course of MST, such as changes in cortisol levels before and after a stressful performance task, predicted therapist-rated treatment non-response. Furthermore, increases in behavioural markers of emotion dysregulation (e.g., parent reports of children's self-regulatory problems) predicted post-treatment arrests. Therefore, the success of family-based interventions for conduct problems might rest on alleviating transdiagnostic markers of general psychopathology like emotion dysregulation (Carver, Johnson, & Timpano, 2017; see section 2.2.1). However, corroborating these behavioural and physiological markers with *p* factor scores, using multi-informant methods to assess treatment non-response, and comparisons with a control group are necessary to fully test this hypothesis.

Decreases in the *p* factor might reflect the common processes of psychological therapies because seemingly targeted interventions resulted in broad improvements across emotional and behavioural problems in the primary study (Fonagy et al., 2018) and studies reviewed above. This would also explain why MST and MAU showed few differences in the primary study, as they both influenced the *p* factor via common therapeutic mechanisms to a similar extent (both treatments were equally intensive and mainly differed in the extent that they featured an overarching theoretical framework). Wright et al. (2016) also reported large within-person declines in their general personality disorder factor in patients who received various treatments in an uncontrolled fashion–the general factor might have captured the common effects across treatments. Furthermore, Wade et al. (2018) reported decreases in the *p* factor in adolescents who had received foster care at an early age, but not in those who were institutionalized. The institutional care arm was unlikely to feature common (or specific) therapeutic processes because such care in Romania was notorious for being damaging to children's mental and physical development (and subsequently banned by the European Union; Nelson, Fox, & Zeanah, 2014).

The only covariate to predict between-person differences in baseline *p* factor scores and linear slopes was sex: adolescent girls reported higher initial *p* factor scores and less change over time. Winiarski et al. (2017) also found that increases in parent-reported emotional dysregulation predicted worse therapist-reported outcomes in girls only. The *p* factor is thought to be invariant across sex in population samples (Caspi et al., 2014), but antisocial girls might represent a more severe sub-population that is particularly prone to comorbid conditions (Keenan, Loeber, & Green, 1999). Nonetheless, the limited number of girls might have skewed the sex contrast. Further work is are needed to determine sex differences in the *p* factor estimated in clinical samples, with a careful consideration of the sample characteristics (e.g., Wade et al., 2018 reported higher *p* scores in boys who experienced early adversity).

Within-person change was not limited to the *p* factor. The specific antisocial factor also showed reductions over the intervention and follow-up period, which might reflect the specific aim of the interventions after separating out changes common to all symptoms. This is supported by the finding that within-person variability in the antisocial factor over time positively predicted official records of criminal activity. However, decline in the antisocial factor was no longer significant when cross-loadings were removed from the model. Some could argue that the decrease in anti-sociality was a function of increases in anxiety because the item that cross-loaded on the antisocial factor traditionally reflects anxiety (SDQ item 16: "I am nervous in new situations"). Nonetheless, item 16 loaded more strongly onto, and thus better reflects, the antisocial factor (and potentially fearlessness, see above) than the anxiety factor in the current sample. Furthermore, forcing SDQ item 16 to load exclusively onto the anxiety factor despite its affinity to the antisocial factor might have supressed the latter's growth curve in the parallel process growth model.

Surprisingly, the specific anxiety factor increased over the intervention and follow-up period, despite decreasing in the current analysis with correlated factors and in the primary study using observed subscale scores (Fonagy et al., 2018). The decrease in observed subscale scores was likely a function of the common variance, which tends to overpower the variance uniquely attributed to specific subscale sores (Rodriguez et al., 2016b; see Chapter 3). The increase in specific anxiety scores might reflect a facilitative effect, whereby adolescents regained some level of fearfulness that is characteristically decreased in adolescents with pronounced conduct problems (Frick, Stickle, Dandreaux, Farrell, & Kimonis, 2005). This is supported in part by the finding that within-person variability in anxiety scores negatively but modestly predicted official records of school exclusions over time.

Another possibility is that anxiety problems replaced anti-sociality because they were partial drivers of antisocial behaviour. Antisocial activity can serve to protect some young people from the social situations they find challenging and must confront once delinquent socializing is no longer available to support their avoidance (Aseltine, Gore, & Gordon, 2000). Alternatively, increases in anxiety might be an artifact of teasing apart the common and specific variance. It is interesting to note that Wade et al. (2018) also reported a modest increase in anxiety factor scores in adolescents who received foster care early in life, suggesting that the effect might be more artifactual than specific to the characteristics of the sample (although abandoned youth are at-a higher risk of delinquency; Van Wert, Mishna, Trocmé & Fallon, 2017).

The specific mood and attention factors showed little within-person change over time, yet significant decreases were observed in the correlated factors model and reported in the primary analysis of observed symptom subscales (Fonagy et al., 2018). Therapeutic change (or the lack thereof) in these problems might have been secondary to more common processes captured by the *p* factor. It is noteworthy that most outcome studies do not separate out the general and specific variance in outcome measures. Therefore, much of what is reported as disorder-specific change using disorder-based subscales might be underpinned by common processes such as decreases in general psychopathology.

5.4.4 Strengths and Limitations

A strength of using multilevel factor analysis is that the bifactor dimensions were estimated 'from the ground up'. That is, changes in the co-occurrences between symptoms were studied at the individual level and then aggregated over the sample, rather than studied and averaged between individuals. Studying within-person change, and between-person variation in within-person change, is naturally aligned with clinical practice which focuses on the individual but draws on knowledge from the population (Molenaar, 2009). Multilevel approaches thus offer a method for integrating both idiographic and nomothetic traditions in the study of the bifactor dimensions.

Another strength of the multilevel modelling approach is that fewer parameters were required to achieve stable growth factor estimates because the analysis was collapsed over 'time' rather than remodelled at each time-point. This is particularly important given the computational demands of bifactor models; a conventional single-level latent growth model, where the bifactor model was re-estimated at each time-point rather than across time-points, did not converge. However, a disadvantage is that it was not possible to test for conventional measurement invariance. Measurement invariance is still assumed within the parameters: an item with a strong within-level factor loading is inherently metric invariant, in that it consistently co-varies with other items over time. However, the extent to which the data support this modelling assumption is unknown. To minimise this issue, partial measurement invariance was demonstrated using conventional invariance tests, e.g., all but the mood factor showed metric invariance, and all but the second threshold showed scalar invariance between pre- and post-treatment. While this provides some confidence that the data generally upheld the assumption of measurement invariance, the results are not directly translatable to the multilevel approach.

Finally, Bayesian Plausible Values (BPVs) were used rather than latent variables to ease the computational demands of studying bifactor dimensions over time. A strength of BPVs is that they take into account the (un)reliability of factor scores and limit the number of type I errors when using factor scores as predictors or outcome variables (Laukaityte & Wiberg, 2017). However, BPVs are still an imperfect measure of latent variables. This is especially true of BPVs for specific factors; the factor determinacy and reliability of factor scores (used in lieu of BPVs) for specific factors was lower than the general factor and suboptimal in absolute terms. Consequently, the specific factor BPVs showed higher variability that likely increased type II error rates in the structural coefficients (e.g., the lack of significant between-person predictors). Therefore, the current growth curves might lack precision and require caution in their interpretation, particularly those for the specific factors.

5.4.5 Implications and Future Directions

Studies of differential stability show that people's rank-ordering on the *p* factor over time is as stable as personality traits (see section 2.3.2-2.3.4). Together with the hypothesis that the *p* factor represents a latent vulnerability to psychopathology (Caspi & Moffitt, 2018), there is an implicit assumption that *p* is immutable. However, the current findings demonstrate that absolute change in adolescents' *p* factor scores is malleable and amenable to a psychosocial intervention (see also Wade et al., 2018).[18] While people's relative standing on the *p* factor might be stable over the course of years, their individual expression might be more fluid over weeks and months (Mischel & Shoda, 1995). Not only does this have important implications for how we conceptualize the *p* factor (e.g., incorporating different timescales), but also how we understand change processes in psychotherapy. For instance, the initial drop in people's symptom ratings during psychotherapy–which is observed for various diagnoses (Hayes, Laurenceau, Feldman, Strauss, & Cardaciotto, 2007)–might reflect reductions in general distress as captured by the *p* factor. More lasting changes could be indexed by specific psychopathology factors that reflect stylistic patterns in the expression of distress (Caspi et al., 2014; Wright et al., 2016).

[18] Ideally, the MST/MAU groups would have been compared to a passive control group to rule out the possibility that change in the active arms was down to the natural passage of time.

It was hypothesised that the treatment (MST) and active control (MAU) arms in the current study would be similar in their rate of change on the *p* factor because both treatments were equally intensive and likely engaged common therapy processes to a similar extent. However, differences in the rate of change on specific factors were expected because the treatment arms differed in their therapeutic modalities (it being largely absent in the MAU condition). Yet, there were minimal differences between the treatment arms in the *p* factor and specific factor slopes. It is likely that the contrast was underpowered with roughly 300 adolescents per arm. Future studies with larger cohort should investigate the possibility that the treatment differences are observable in the specific factor slopes. It might simply be that specific interventions differ very little, even after controlling for common therapeutic effects, mirroring meta-analytic findings (Cuijpers, van Straten, Adnersson, & van Oppen, 2008; Wampold et al., 1997; Weisz et al., 2017). However, the focus need not be on treatment differences. By analysing outcome measures with the bifactor model, we might also find that different predictors of treatment efficacy (e.g., specific tasks, therapeutic stance, and therapeutic alliance) are associated with different levels of the psychopathology hierarchy (e.g., syndromal, spectral, and general factors), enabling more precise investigations of “what works for whom” question (Fonagy et al., 2015).

Chapter 6 Clarifying the Prognostic Value of Personality Disorders for Depression Outcomes Using the Bifactor Model

6.1 Introduction

In Chapter 5, clinical outcomes were analysed with the bifactor model to determine therapeutic changes specific to certain disorders as well as changes common to all disorders. The main take-home was that disorder-specific outcome measures are obscured by the substantial overlap between disorders; controlling for such overlap provides a more nuanced picture of therapeutic change. The current chapter extends these findings to personality disorder (PD) research, where the prognostic value of PDs for depression outcomes is uncertain but potentially confounded by disorder-general variance.

I will first review the mixed findings regarding the prognostic value of PDs for depression outcomes and argue that the substantial overlap between PDs masks their unique prognostic value. I will then compare the bifactor model to alternative models for how well it summarizes the covariation in self-reported PD symptoms, and examine the benefits of predicting changes in depression ratings over an inpatient treatment using the general and specific PD factors compared to standard PD factors that conflate the general and specific variance.

6.1.1 Do Personality Disorders Predict Adverse Outcomes for Depression?

Depressive disorders are a common and debilitating set of mental health problems that affect over 300 million people world-wide (World Health Organization, 2017). The most common depressive disorder, Major Depressive Disorder (henceforth, depression), is defined in the Diagnostic and Statistical Manual of Mental Disorders-5 (DSM-5) by changes in affect (e.g., low mood, loss of interest or pleasure, feelings of worthlessness or excessive guilt), cognition (e.g., poor concentration or indecisiveness, suicidal thoughts), and physical functioning (e.g., weight gain or loss, insomnia or hypersomnia, fatigue; American Psychiatric Association, 2013). Depression is associated with several adverse outcomes, including lower educational attainment, poorer work performance, comorbid mental and physical health conditions, and a higher risk of suicide (Ferrari et al., 2013; Kessler & Bromet, 2013). It is unsurprising that the World Health Organization (2017) ranked depression as the world's leading cause of disability.

One would hope that treatments for depression were highly effective to prevent a global epidemic, but gold-standard treatments such as cognitive-behavioural therapy and antidepressant medication achieve response rates of around 50% (Hofmann, Asnaani, Vonk, Sawyer, & Fang, 2012; Papakostas & Fava, 2009; Furukawa et al., 2016). Effect sizes for psychotherapy (*d* = .42; van Straten, Geraedts, Leeuw, Andersson, & Cuijpers, 2010) and pharmacotherapy (*d* = .40; Kirsch et al., 2008; Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008) are modest compared to wait-list or placebo controls, respectively, suggesting that there are several barriers to treatment (David, Cristea, & Hofmann, 2018).

One aspect of depression that may complicate treatment is that it frequently coincides with other conditions (Kessler et al., 2003). For example, epidemiological studies show that at least 50% of people reporting depressive symptoms also report anxiety symptoms or phobias (Adewuya et al., 2018; Choy, Fyer, & Goodwin, 2007; Kessler et al., 2008; Lamers et al., 2011). People reporting depressive symptoms are also at a heightened risk of alcohol, nicotine, and substance dependencies (Grant, 1995; Grant & Hardford, 1995; Swendson & Merikangas, 2000). The high rates of comorbidity among depression and other psychiatric disorders is also found in prospective longitudinal studies that are free from the biases associated with retrospective reports (Boschloo et al., 2011; Murphy et al., 2004).

Depressive disorders are also highly comorbid with personality disorders (PD; Grant et al., 2005; Friborg et al., 2014). Patients diagnosed with both depression and PDs show depressive symptoms that are more severe, persistent, recurrent, and have an earlier age of onset (Agosti, Hellerstein, & Stewart, 2009; Cyranowski et al., 2004; Fava et al., 1996; Sheets, Duncan, Bjornsson, Craighead, & Craighead, 2014; Skodol et al., 2011). The long-standing intuition since Kraeplin is that PDs hinder treatments for depression (Clarkin, Petrini, & Diamond, 2019; Ilardi and Head, 1995), but this clinical wisdom has been questioned over the last 30 years. Some systematic reviews and meta-analyses support the idea that PDs interfere with treatments for depression (Newton-Howes, Tyrer, & Johnson, 2006; Newton-Howes et al., 2014; Reich, 2003) while others show that PDs make little difference to treatment outcomes (Kool et al., 2005; Mulder, 2002).

Methodological heterogeneity is one of the main issues that contribute to the mixed prognostic value of PDs for depression outcomes. Studies differ in whether they are naturalistic or controlled; whether the treatment was psychotherapy, pharmacotherapy, or polytherapy; the way in which depression was defined and assessed; the inclusion criteria of participants; the length of treatment and assessment intervals; the definition of recovery (binary remission or continuous depression scores); and variable sample sizes, to name a few (French, Turner, Dawson, & Moran, 2017).

In his influential review, Mulder (2002) argued that the best controlled studies showed the weakest association between PD and depression outcomes. Many cite this finding as fact, but few recognise that it was derived from seven studies that were too heterogeneous to meta-analyze. More recently, Newton-Howes et al. (2014) pooled the findings from 58 studies and concluded that having a PD was associated with double the odds of a poor response to treatment for depression (OR = 2.16, 95% CI [1.83, 2.56]). Nonetheless, when the studies were split by design, the odds of a poorer outcome associated with having a PD were lower in controlled studies (OR = 1.52, 95% CI [1.23, 1.87]) compared to naturalistic studies (OR = 2.68, 95% CI [2.08, 3.46]). In both cases, however, the association between having a PD and poorer depression outcomes was significant. Therefore, PDs do appear to hinder treatment for depression, but as Mulder claimed, the association weakens with greater methodological control.

Patients with at least one PD often report higher depression scores at baseline and post-treatment than those without a PD (Casey et al., 2004; Craigie, Saulsman, & Lampard, 2007; Hardy et al., 1995; Newton-Howes et al., 2006; Shea et al., 1990; Tyrer, Tyrer, Yang, & Guo, 2016; van den Hout, Brouwers, & Oomen, 2006). However, controlling for baseline depression severity negates the association between PDs and poorer depression outcomes (De Bolle et al., 2010; Erkens et al., 2018; Harley et al., 2006; Kelly, Nur, Tyrer, & Casey, 2009; Spinhoven et al., 2011;

van Bronswijk et al., 2018). Unless the baseline differences between PD and non-PD groups are controlled for (which would naturally occur with baseline randomization in RCTs), differences in depression outcomes might be the result of differences in baseline severity rather than differential responses to treatment (Fowler et al., 2018; Moradveisi, Huibers, Renner, Arasteh, & Arntz, 2013; Unger, Hoffmann, Köhler, Mackert, & Fydrich, 2013). In other words, PD patients may do worse because they start off worse, not because they are less responsive to treatment. It may even be that PD patients' greater baseline severity gives the clinical impression that minimal progress has been made during treatment, but their progress is similar to patients without a PD diagnosis in absolute terms.

Some studies show that the negative prognosis associated with PDs remains even after controlling for baseline depression severity, and in some cases, baseline severity loses its predictive role (Fournier et al., 2008; Gorwood et al., 2010; Sasso & Strunk, 2013; Strandholm et al., 2014). Furthermore, some observational (i.e. non-treatment) studies show that certain personality disorder traits, such borderline and schizotypal traits, predict recurrence, persistence, and delayed remission of depression, even after controlling for baseline depression severity (Grilo et al., 2005; Grilo et al., 2010; Sheets et al., 2014; Skodol et al., 2011; Viinamaki et al., 2006). These findings suggest that certain aspects of PDs that do not overlap with overall illness severity might still predict differential responses to treatments, but novel methods are required to tease these aspects apart.

6.1.2 How Should Personality Disorders Be Classified?

The studies reviewed indicate that PDs do interfere with treatments for depression, but not to the extent that was originally thought. They also suggest that certain aspects of PDs are stronger predictors of depression outcomes than others, but these are obscured when assessing PDs as distinct entities. In a similar line of thought, Mulder (2002) concluded that "The inconsistent and unexpected findings [about the prognostic value of PDs for depression outcomes] may be due to the diagnostic system rather than measurement or sampling error." (p. 366).

Personality disorders, like other psychiatric disorders, co-occur more frequently than expected by chance and thus might be underpinned by broader underling processes (Oldham et al., 1992; Tyrer, Reed, & Crawford, 2015). Factor analytic studies show little support for the independence of DSM-IV PDs (Bastiaansen et al. 2011; Cox, Clara, Worobec, & Gant, 2012; Fossati et al., 2000; Huprich, Schmitt, Richard, Chelminski, & Zimmerman, 2010; Sharp et al., 2015; Trull, Vergés, Wood, & Sher, 2013). Instead, PDs can be classified by five broad factors that mirror the five-factor model of personality (Krueger et al., 2011; Krueger et al., 2012; Livesley, 2012; Skodol, 2012; Widiger, 2011).

There is widespread support for classifying PDs in terms of maladaptive traits, including negative affectivity (akin to neuroticism), detachment (vs. extraversion), antagonism (vs. agreeableness), disinhibition (vs. conscientiousness), and psychoticism (akin to openness; Anderson et al., 2012; Bagby et al., 2013; Hopwood et al., 2013; Sellbom, Smid, Saeger, Smit, & Kamphius, 2014; Thomas et al., 2012; Quilty, Ayearst, Chmielewski, Pollock, & Bagby, 2013; Wright & Simms, 2014; Wright et al., 2012). While there are some variations in the number and nature of traits specified, such as a four-factor alternative (Bastiaansen et al., 2011; Widiger, Livesley, & Clark, 2009) and six-factor alternative (Ashton, Lee, de Vries, Hendrickse, & Born, 2012; Ashton et al., 2004), most models are grounded in the five-factor model of personality that has been tested by hundreds if not thousands of empirical studies (Widiger & Costa, 2013).

The five-factor model of adaptive and maladaptive personality traits reflects the spectral level of a hierarchy (e.g., the covariation among disorders can be explained by higher level dimensions; Kotov et al., 2017). Markon, Krueger, and Watson (2005) ran a meta-analytic factor analysis of 44 adaptive and maladaptive personality trait scales with up 52,879 respondents and found support for two-to-five hierarchically organized factors, e.g., spectral level factors split into sub-spectral factors that were in turn split into subfactors, etc.

More recently, researchers have investigated super spectral or general factors that explain the covariation among all broadband traits. The two main methods that have been used by PD researchers to achieve this are the bifactor method and the bass-ackwards method. The former separates out the covariance among personality disorder symptom/disorder ratings into general and specific components (see Chapter 1). In other words, the bifactor model summarizes the problematic characteristics shared by all PDs (e.g., overall personality dysfunction) as well as characteristics specific to individual PDs or maladaptive personality traits (Sharp et al., 2016; Wright et al., 2016). By contrast, the bass-ackwards approach decomposes the variance of a single factor into finer components until no new factors are apparent (Boudreaux, South, & Oltmanns, 2019). This chapter focuses on the bifactor model as it explicitly models the hierarchical (or in this case, heterarchical) relationship between factors, while the bass-ackwards approach treats hierarchy as a construct and is more of a variance decomposition method[19] (Goldberg, 2006).

It should be noted that PD classification systems have progressed since the publication of the DSM-IV and ICD-10. For example, the DSM-5's Personality and Personality Disorder Work Group revised the PD diagnostic model to include three criteria that capture the shared and specific characteristics of PDs (Oldham, 2018; Skodol, 2012).[20] Criterion A is a dimensional measure of personality disorder severity based on transdiagnostic impairments in self-functioning (e.g., the degree of coherence and directedness experienced in one's identity) and interpersonal functioning (e.g., the extent that one can empathise and share intimacy with others). Criterion B includes 25 maladaptive traits that cluster into five broadband trait domains mirroring the five-factor model (e.g., negative affectivity, detachment, antagonism, disinhibition, and psychoticism). Criterion C involves six (rather than ten) PD categories (antisocial, avoidant, borderline, schizotypal, obsessive-compulsive, and narcissistic), with the PD not-otherwise-specified category replaced by a 'trait specified' category (e.g., the patient meets the criteria for PD but not any of the six PD diagnoses).

Patients qualify for a PD diagnosis (Criterion C) if they show at least moderate impairment in general personality functioning (Criterion A). Additionally, the way that patients express their impairment can be summarized

[19] It is surprising that the higher-order model, which explicitly estimates a hierarchy, has not been more widely adopted by PD researchers. It is, however, popular among personality researchers who have attempted to validate a 'general factor of personality' in addition to lower-order spectral factors (van der Linden, Nijenhius, & Bakker, 2010).

[20] The ICD-11 personality disorder taskforce proposed a similar model that includes a dimensional index of 'personality disturbance' and specific maladaptive traits reminiscent of the five-factor model that reflect the ways in which this disturbance is expressed (Tyrer et al., 2015).

with the maladaptive trait domains (Criterion B). In other words, a patient's overall level of personality impairment (Criterion A) and style of symptomatic expression (Criterion B) are used in concert to inform a specific PD diagnosis (Criterion C). Unlike the standard categorical model, the alternative model integrates information about PDs at different levels of specificity to arrive at a more holistic yet sensitive diagnosis.

The alternative model of PDs has several advantages over the standard categorical model. For instance, Criterion A overcomes the problem of comorbidity by including an explicit marker of severity that would otherwise obscure assessment by introducing multiple co-occurring PDs (Fowler & Oldham, 2013). Moreover, Criterion B overcomes problems with heterogeneity since most presentations can be mapped out on the trait markers (Hopwood, 2018). Despite these clear benefits, the alternative model of PDs was not approved by the APA's board of trustees on the grounds of there being insufficient evidence (Oldham, 2015), but also due to dynamics within the governing bodies (Skodol, Morey, Bender, & Oldham, 2013). Therefore, there has never been a more pressing time to test a 'binomial' classification of PDs that includes markers of severity and style. The bifactor model is poised to be instrumental in this endeavour, as the parsing of shared and specific variance aligns with the distinction between severity and style. I will now evaluate the handful of innovative studies that have used the bifactor model to examine the latent structure of PDs.

6.1.3 Bifactor Studies of Personality Disorders

The first study to investigate the general and specific aspects of PDs with the bifactor model was by Sharp et al. (2015). They demonstrated that the covariation in

PD symptom ratings did not neatly group into six factors reflecting DSM-IV diagnoses (while there are 10 DSM-IV PDs, dependent, histrionic, schizoid, and paranoid PDs were not sampled due to their low prevalence). Instead, symptoms belonging to a given PD cross-loaded onto multiple PD factors. An exploratory bifactor model showed that the covariation among symptoms was best explained by a general PD factor, as well as six specific factors that each resembled at least one disorder. Furthermore, borderline PD symptoms loaded almost exclusively onto the general factor, suggesting that problems characteristic of borderline PD (e.g., identity disturbance and interpersonal problems) are shared across PDs (Clark, Nuzum, & Ro, 2018; Sharp et al., 2015).

In a similar study by Williams, Scalco, and Simms (2017), the covariation among PD symptoms was best explained by an exploratory bifactor model with one general PD factor and four specific factors that resembled the five-factor model (e.g., neuroticism, extraversion, disinhibition vs. constraint, and psychoticism). Like Sharp et al.'s (2015) study, borderline PD symptoms loaded strongly and exclusively onto the general PD factor. Unlike Sharp et al.'s study, however, the full range of DSM-IV PDs were assessed, which likely produced different specific factors. Wright et al. (2016) also reported that borderline PD subscale ratings loaded strongly and exclusively onto a general PD factor estimated at the within-person level. However, they too reported different specific factors, including detachment, dependency, compulsivity, dominance, and disinhibition.

Other studies applying the bifactor model to PDs support both general and specific PD dimensions but differ in whether borderline PD ratings saturate onto the general factor. For instance, Conway, Hammen, and Brennan (2016) found that while BPD symptom counts loaded strongly onto a general PD factor, they also contributed to a specific factor reflecting instability vs. rigidity (e.g., positive loadings from borderline and dependent PDs and negative loadings from obsessive-compulsive PD). Furthermore, Jahng et al. (2011) found that in a large population cohort, the only specific factor to show significant (albeit weak) loadings after accounting for the general PD factor was one reflecting cluster B diagnoses (e.g., borderline, antisocial, histrionic, and narcissistic PDs). Therefore, caution is warranted when attributing theoretical significance to the saturation of borderline PD symptoms on the general PD factor, as changes in methodology (e.g., diagnostic- vs. symptom-level indicators) appear to influence this (a similar result for psychosis indicators saturating onto the *p* factor is described in section 2.2.1).

The general PD factor is thought to reflect the severity of personality impairment regardless of the specific PDs present (Jahng et al., 2011), consistent with theories that define PD as dysfunction in the overall structure of personality (Livesley, 2011). Indeed, the general PD factor predicts problems in a range of functional domains that would require a cohesive personality, including social and occupational functioning (Conway et al., 2016). By contrast, specific PD factors are thought to reflect stylistic expressions of symptoms, consistent with theories that separate out severity from style in personality dysfunction (Hopwood et al., 2011; Livesley, 2011). Few have validated the specific PD factors against personality traits, but one study by Hengartner, Ajdacic-Gross, Rodgers, Müller, and Rössler (2014) reported that personality disorders and personality traits formed theoretically relevant factors, such as avoidance/schizoid vs. extraversion and antisociality vs. conscientiousness, but their interpretation is limited by their weak factor loadings.

6.1.4 Study Aims

The studies reviewed demonstrate that PD measures capture the shared characteristics of PDs despite being designed to assess their unique features. Therefore, PD measures will be influenced by the principle of intwined generality (Gustafsson, 2002). That is, the predictive influence of specific PD markers, including PD status, PD count, or latent variables representing PDs, will be conflated with the shared variance among PDs unless it is explicitly controlled for. As Mulder (2002) put it, "Classification problems mean that it remains unclear whether personality disorder categories are a general measure of personality pathology affecting outcome or whether individual categories, or clusters, predict different outcomes." (p. 366). Examining the role of a specific PD marker on depression outcomes neglects the fact that this measurement is artificially derived from a broader construct (i.e. overall personality dysfunction). To examine the influence of specific PDs, one must first control for the variance common to all PDs inherent in any single 'slice' of the measure.

The current study investigated the prognostic value of PDs for depression outcomes. Unlike past studies that conflate general and specific aspects of PDs, the current study explicitly separated out these sources of variance using the bifactor model. The first part of the analysis compares the bifactor model to the correlated factors model and single-factor model for how well they describe the covariation among self-reported PD symptoms assessed after admission to psychiatric care. In the second part of the analysis, general and specific PD factors were used to predict initial depression scores and their rate of change over a 6-8 week inpatient intervention. The strength and direction with which general and specific PD factors predicted depression outcomes was compared to that of the correlated PD factor dimensions, which represent each PD but conflate the general and specific variance.

If the general PD factor reflects the severity of personality dysfunction, then it should predict higher depression scores overall. When the common variance is not controlled for, it might drive the association between "specific" PD markers and poorer depression outcomes. However, if general PD or its sequelae are controlled for (e.g., by covarying for baseline depression severity), then depression scores might normalize, giving the impression that PDs do not predict poorer outcomes. In either case, general PD predicts the overall severity of depression, not the rate of change (i.e. treatment responsiveness). By contrast, if specific PD factors reflect stylistic expressions of maladjustment, then those associated with antagonistic tendencies (e.g., borderline traits) should interfere with treatment engagement and predict slower rates of change (i.e. flatter slopes). However, these effects might be masked in studies that do not control for the common variance in PDs. We would therefore expect that the negative prognostic influence of specific PDs on depression outcomes should not be observed in the correlated factors model, which conflates the common and specific variance in PDs. Instead, correlated PD factors should mirror the effect of general PD, predicting higher baseline depression scores but not differential rates of change.

6.2 Method

6.2.1 Participants

The sample consisted of 2,352 inpatients admitted to the Menninger Clinic, Houston, between December 2012 and June 2015. Full demographics are presented in Table 6.1. Patients were mostly White/Caucasian American (89%), middle aged

(*M* = 35, *SD* = 15), and a mix of sexes (48% female). Most participants underwent some form of higher education, including some college (35%), completing a Bachelor's, Technical or Associates Degree (33%), or attaining a postgraduate degree or doctorate (21%). There were no exclusion criteria; participants of all diagnoses and severity levels were recruited and included in the analysis. Over half (56%) of patients reported moderately severe or severe depression on the Patient Health Questionnaire-9 (PHQ-9). Data were collected as part of the hospital's ongoing Adult Outcomes Project, which aims to integrate research and routine clinical practice (Allen et al., 2009). Data collection and analysis was approved by Baylor College of Medicine's Institutional Review Board.

Rates of DSM-IV PDs were as follows: borderline personality disorder (19%), avoidant personality disorder (16%), obsessive-compulsive personality disorder (9%), antisocial personality disorder (3%), narcissistic personality disorder (2%), and schizotypal personality disorder (0.4%). Histrionic, schizoid, dependent and paranoid PDs were not assessed as they showed prevalence rates of < .01% in a pilot sample (*N* = 1,200). This is also consistent is also consistent with the main PDs included in the DSM-5 Section III (American Psychiatric Association, 2013). Of the 31% of patients meeting the criteria for any PD, 34% met the criteria for at least one other PD.

Table 6.1

Clinical and Demographic Characteristics of the Inpatient Sample (N = 2,352)

Sample Characteristic	*M* or *N*	*SD* or %
Clinical		
PHQ-9 (admittance)	15	7
Minimal or none (0-4)	233	10%
Mild (5-9)	327	14%

Moderate (10-14)	455	18%
Moderate severe (15-19)	575	24%
Severe (20-27)	762	32%
Length of Stay (weeks)	6	3
Episode Number		
First admission	2055	87%
>1 admissions	297	13%
Program		
Hope	641	27%
CPAS	379	16%
Compass	758	32%
PIC	574	24%
Demographic		
Age	35	15
Sex		
Female	1120	48%
Male	1232	52%
Racial Background		
White or Caucasian	2096	89%
Other[a]	255	11%
Highest Level of Education		
Some schooling	56	2%
High School Diploma or Equivalent	211	9%
Some College	814	35%
Bachelors, Technical, or Associates Degree	761	33%
Postgraduate (Masters, Doctoral, or Professional Degree)	481	21%
Marital Status		
Married	1760	75%
Never married/separated	592	25%

Note. Compass = Compass Program for Young Adults (18-30); Hope = Hope Program for Adults; CPAS = Comprehensive Psychiatric Assessment Service; PIC = Professionals in Crisis program.

[a]Includes Asian, Black or African-American, Native American or Other Pacific Islander, and Multiracial.

6.2.2 Measures

Personality disorder symptoms were assessed within 72 hours of admission using the Structured Clinical Interview II for DSM-IV Personality Disorders Screening Questionnaire (SCID-II; First et al., 1994). Seven-to-nine symptoms for antisocial, avoidant, borderline, narcissistic, obsessive-compulsive, and schizotypal personality disorders were rated by patients with a 'yes' (threshold or true) or 'no' (subthreshold, false or absent). Internal consistency was acceptable or near acceptable for most disorders ($\alpha_{narcissistic}$ = .66, $\alpha_{avoidant}$ = .74, $\alpha_{borderline}$ = .75, $\alpha_{antisocial}$ = .86), except for two ($\alpha_{obsessive}$ = .56, $\alpha_{schizotypal}$ = .51). Antisocial behaviour items after the age of 15 were used.

Depression symptoms were assessed at admission and every fortnight until discharge with the Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spizter, & Williams, 2001). The PHQ-9 is a screening questionnaire of the DSM-IV criteria for major depressive disorder. Patients rated the frequency of depressive symptoms over the past fortnight on a Likert scale ranging from 0 (not at all) to 3 (nearly every day). Responses were then summed to form total depression scores. The PHQ-9 shows excellent criterion validity, with sensitivity and specificity rates for detecting depression of 88% or more depending on the cut-off used (Kroenke, Spitzer, Williams, & Löwe, 2010; Manea, Gilbody, & McMillan, 2015). The PHQ-9 also has excellent internal consistency (α = .89) and test-retest reliability (*r* = .84; Kroenke, Spitzer, & Williams, 2001), and is sensitive to change (Kroenke et al., 2010; Löwe, Kroenke, Herzog, & Gräfe, 2004). The internal consistency in the current sample averaged across the assessment periods was excellent (α = .90; range = .89-.91).

6.2.3 Intervention

Patients were admitted to one of four inpatient programs: Compass (31%) for young adults (18-24); Comprehensive Psychiatric Assessment Service (CPAS; 18%) for adults in crisis; Hope (27%) for adults with more chronic difficulties; and Professionals in Crisis (PIC; 24%) for professionals with long-standing disorders. All programs were multimodal and equally intensive, consisting of individual and group psychotherapy, psychoeducation, social and recreational activities, family work, psychopharmacology and medication management, general psychiatric and medical care, and continuous nursing care (Fowler et al., 2017). Patients were treated by multidisciplinary teams composed of psychiatrists, psychologists, social workers, psychiatric nurses, and rehabilitation specialists. Patients stayed for 6 weeks on average (*SD* = 3 weeks).

6.2.4 Data Quality Checks

Missing Data. The main cause of missing data was the length of inpatient stay; those who were discharged before the 8-week period showed missing responses up to that point. Specifically, 305 patients (12% of patients undergoing treatment after admission) were discharged within two weeks, a further 470 patients (22% of patients undergoing treatment at two weeks) were discharged within 4 weeks, a further 565 patients (33% of patients undergoing treatment at 6 weeks) were discharged within 6 weeks, and a final 638 patients (56% of patients undergoing treatment at 6 weeks) were discharged within 8 weeks. Missing data was assumed to be missing at random (i.e. the likelihood that a data point was missing depended on some observed or unobserved data other than the missing value itself, such as treatment length) and handled with full-information maximum likelihood (see Supplement C3). Length of inpatient stay was included as a covariate in all models.

Response distributions. Patients used the disagreement response option in the SCID-II screening questionnaire 87% of the time (*SD* = .12, range = .56-.99) and the agreement response option 13% of the time (*SD* = .12, range = .01-.44). Borderline and avoidant PD items showed the highest rates of agreement on average, while antisocial and schizotypal PD items showed the lowest (see Table 6.2).

Table 6.2

Response Frequencies on the SCID-II PD Screening Questionnaire by Personality Disorder (as Proportions)

Personality disorder	Mean	*SD*	Min	Max
Antisocial				
'No'	0.98	0.01	0.96	0.99
'Yes'	0.02	0.01	0.01	0.04
Avoidant				
'No'	0.79	0.10	0.65	0.87
'Yes'	0.21	0.10	0.13	0.35
Borderline				
'No'	0.74	0.07	0.61	0.83
'Yes'	0.26	0.07	0.17	0.39
Narcissistic				
'No'	0.91	0.08	0.71	0.97
'Yes'	0.09	0.08	0.03	0.29
Obsessive-compulsive				
'No'	0.83	0.12	0.56	0.96
'Yes'	0.17	0.12	0.04	0.44
Schizotypal				
'No'	0.96	0.03	0.89	0.99
'Yes'	0.04	0.03	0.01	0.11

There were few differences between the estimated and observed response distributions in the bifactor model (*M* = .003, *SD* = .002, range = -.002–.007), correlated factors model (*M* = -.004, *SD* = .003, range = -.01–.003), and single factor model (*M* = .001, *SD* = .005, range = -.01–.02; see 'Confirmatory Factor Analysis' for model specification).

Residual Correlation Matrix. The residual correlation matrix included 1,176 unique polychoric correlations between SCID-II PD items. In the bifactor model, the average negative residual (i.e. correlation over-estimated by the model) was -.09 (*SD* = .08) and the average positive residual (i.e. correlation under-estimated by the model) was .09 (*SD* = .08), both falling under the standard criteria of .20 (Christensen, Makransky, & Horton, 2017) and stricter criteria of .10 (Goodboy & Kline, 2017). There were 38 'problematic' residuals (3%) that fell above .29 or below -.29 (i.e. average residual +/- .2), four of which fell above .39 or below -.39 (0.3%, i.e. i.e. average residual +/- .3; Pallant & Tennant, 2007). Many problematic residuals were associated with antisocial PD items (24/38 or 63%).

In the correlated factors model, the average negative residual was -.10 (*SD* = .09) and the average positive residual was .08 (*SD* = .07), both falling within standard and stricter criteria (.20 and .10, respectively). There were 35 problematic residuals (3%) that fell above .29 or below -.29 (i.e. average residual +/- .2), 11 of which fell above .39 or below -.39 (0.9%, i.e. average residual +/- .3). Unlike the bifactor model, residuals were more diffuse and less associated with antisocial PD items (11/35 or 31%).

In the single factor model, the average negative residual was -.16 (*SD* = .12) and the average positive residual was .12 (*SD* = .11), both falling within the standard cut-off (.20) but not stricter cut-off (.10). There were 91 problematic residuals (8%) that fell above .34 or below -.34 (i.e. average residual +/- .2), 30 of which fell above .44 or below -.44 (3%, i.e. average residual +/- .3). Like the bifactor model, many residuals were associated with antisocial PD items (55/91 or 60%).

6.2.5 Statistical Analysis

Confirmatory Factor Analysis (CFA). Three item-level CFA models were compared for how well they described the covariation among SCID-II PD symptoms. The first model included a single factor upon which all symptoms loaded. The second model included six correlated factors each representing a PD with no cross-loadings. The third model included a general factor upon which all items loaded, as well as six specific factors that each represented a PD. Two versions of the bifactor model were tested: a traditional version where the specific factors were uncorrelated (Holzinger & Swineford, 1937), and a revised version where the correlations among specific factors were freed. In both versions, the general and specific factors were uncorrelated. All solutions were standardized (e.g., the first loading of each factor was freed, and factors had a mean of zero and variance of one). All models were estimated in Mplus 8.0 (Muthén & Muthén, 2017).

Model Comparison. Models were estimated using the robust maximum-likelihood estimator (MLR) and compared using Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and sample-size adjusted Bayesian Information Criteria (aBIC). A difference of 2 (AIC/BIC/aBIC) between models was considered negligible; a difference of 2-7 (AIC) or 2-6 (BIC/aBIC) suggested some evidence favouring the competing model; a difference of 7-10 (AIC) or 6-10 (BIC/aBIC) suggested strong evidence favoring the competing model, and a difference greater than 10 (AIC/BIC/aBIC) suggested very strong evidence favouring the competing model (Raftery, 1995). The difference in BIC values was also formally tested with the Vuong test, a likelihood-based test statistic corrected for the number or freely estimated parameters in each model (Vuong, 1986; see section 4.2.5 for equation).

Models were re-estimated using the weighted least squares means and variances adjusted estimator to assess their global fit. Acceptable fit was defined by Comparative Fit Index (CFI) values ≥ .90, Tucker-Lewis Index (TLI) values ≥ .90, and root mean squared error of approximation (RMSEA) values ≤ .08, whilst excellent fit was indicated by CFI values ≥ .95, TLI ≥ .95 and RMSEA ≤ .06 (Hu & Bentler, 1999).

Finally, models were assessed for overfitting with double cross-validation (Cudeck & Browne, 1983). The sample was randomly split into a calibration group and a test group. Parameters for the bifactor, correlated factor, or single-factor model were freely estimated in the calibration group and used to fix the parameters in the test group. Substantial differences between the calibration and test models, determined by the difference in information criteria difference values above, suggest that the model parameters are sensitive to peculiarities of the group used to estimate them, which is a symptom of overfitting (i.e. capitalizing on noise in the dataset). The process is then repeated, with participants who served as the calibration group now used as the test group and vice versa.

Reliability analysis. Model-based reliability indices were calculated from the MLR factor loading matrix using Dueber's (2017) Bifactor Indices Calculator. Reliability indices included omega hierarchical/hierarchical subscale (ω_H/ω_{Hs}, e.g., the proportion of variance in raw total or subscale scores explained by a the general and specific factors, respectively), construct reliability (H, e.g., the proportion of variance in the indicators explained by a given factor or the reliability of latent factor scores), explained common variance/explained common variance subscale (ECV/ ECVs, e.g., the strength of a given factor relative to all other factors in describing the common variance among items or the degree of multidimensionality; see section 3.2.3 for further details).

ECV/ECVs values ≥ .70 indicate that the majority of common variance is explained by a single factor and is hence 'essentially unidimensional'; ω_H/ω_{Hs}

values ≥ .80 indicate that the majority of variance in raw total or subscale scores is explained by the general or specific factors, respectively; H values ≥ .70 indicate that latent factor scores are represented well by a given set of indicators and are hence reliable (Hancock & Mueller, 2001; Rodriguez et al., 2016a).

The mean parameter change (MPC) and standard deviation of the parameter change (SDPC) were also computed to determine the extent that factor loadings for a given disorder decreased (positive MPC values) or increased (negative MPC values) from the correlated factors model to the bifactor models, and hence captured more or less of the common variance.

Latent Growth Model. Latent growth curve models (LGCM) were used to estimate changes in PHQ-9 total scores over the course of inpatient treatment (Clapp et al., 2013; Duncan & Duncan, 2009). Unlike multilevel growth models that represent time as an observed variable at the within-person level (see Chapter 5 and Appendix C), LGCMs represent time as latent parameters estimated in a factor analytic context (see Appendix D for a detailed description). In brief, the data are arranged in 'wide' format, where different variables represent the outcome variable at each time-points (see Table 6.3). Each iteration of the outcome variable is then loaded onto a growth factor at fixed values that represent the estimated growth process (e.g., 0, 1, 2, 3 reflects linear growth). Each loading reflects the predicted value of the outcome variable at a given time-point (in the multilevel approach, we would regress a single outcome variable that includes values for all time-points onto a variable with time-scores). The latent slope factor mean reflects the direction and steepness of the growth curve.

The outcome variables also load to equality onto an intercept factor, which reflects a constant predicted outcome value when the time value is zero. Consequently, the latent mean of the intercept factor reflects the predicted outcome value at the designated baseline time-point. The latent variances for the intercept and slope factors reflect inter-individual differences in person-specific baseline values and growth curves, respectively (mirroring the between-person and within-person levels of a multilevel growth model).

Table 6.3

Outcome Variable Scores at Each Time-Point Structured in 'Wide' Format

ID	y_1	y_2	y_3	…	y_T
1	21	14	3		0
2	17	13	10		10
3	15	5	12		14
4	24	19	15		7
5	18	16	11		3
6	16	8	3		0
7	19	15	12		8
8	11	5	0		0
⋮					
2,352	20	18	17		21

Note. y_t reflects the outcome variable at time-point *t*, where *t* = 0, 1, 2, 3, … *T*-1. Data are simulated for demonstration purposes.

Three models were estimated: an unconditional model with growth factors only (i.e. an intercept and slope factor), a part conditional model with growth factors, PD factors, and clinical covariates, and a full conditional model with growth factors, PD factors, clinical covariates, and demographic covariates. In the unconditional growth model, an intercept factor was estimated with loadings from PHQ-9 total scores at weeks 2-8 fixed to one and a linear slope factor with loadings reflecting a linear increase in time (week 2 scores = 0, week 4 scores = 1, week 6 scores = 2, week 8 scores = 3). PHQ-9 scores at week 0 (i.e. admission) were not part of the growth model but used a time-invariant covariate given the importance of controlling for baseline depression severity (see section 6.1.1).

A quadratic slope factor, with loadings that reflected non-linear increments in time (e.g., week 2 = 0, week 4 = 1, week 6 = 4, week 8 = 9), was added to the model and evaluated using the information criteria difference values described above. Growth factor variances and covariances were freely estimated or all growth factors.

In the part conditional growth model with PD factors and clinical covariates, the best-fitting growth factors from the unconditional model were regressed onto the bifactor dimensions or correlated factor dimensions estimated from SCID-II PD responses at admission. Growth curves and PD factors were estimated within the same structural equation model to avoid the use of factor scores. Growth factors were also regressed onto clinical covariates, including PHQ-9 scores at admittance, length of inpatient stay, number of prior admissions (first admission vs. one or more prior admissions), and inpatient program (HOPE vs. Compass; CPAS vs. Compass; PIC vs. Compass).

In the full conditional model, the growth factors were regressed onto the PD factors from the bifactor or correlated factor model, clinical covariates, and demographic variables, including age at admittance, sex, ethnicity (White/Caucasian vs. all other ethnic groups), highest level of education obtained (up to some college vs. bachelor's degree or beyond), and marital status (married vs. not marred/separated). Growth models were ran in Mplus 8.0 using the MLR estimator (Muthén & Muthén, 2017) and all covariates were centred. Partially standardized regression coefficients, which are standardized on the x-axis (e.g., latent factor scores) but not on y-axis (e.g., original PHQ-9 metric) are reported for all growth models.

6.3 Results

6.3.1 Model Comparison

The single factor model fit the data poorly (see Table 6.4) but all PD items loaded healthily, demonstrating their unidimensionality (see Table 6.5). The correlated factors model–with factors representing antisocial, avoidant, borderline, narcissistic, obsessive-compulsive, and schizotypal PDs–showed a good fit that improved on the single factor model (ΔAIC = 2,843; ΔBIC = 2,756; ΔaBIC = 2,804; *z* = 58.19, *p* < .001; see Table 6.5). All factors showed healthy positive loadings and were positively and uniformly inter-correlated (aside from the antisocial factor), suggesting the presence of a higher-order factor (see Table 6.5).

The traditional bifactor model–with a general factor and uncorrelated specific factors representing each PD–showed a good fit that improved on the correlated factors model (ΔAIC = 443; ΔBIC = 247; ΔaBIC = 355; *z* = 8.18, *p* < .001) and single-factor model (ΔAIC = 3,286; ΔBIC = 3,003; ΔaBIC = 3,159; *z* = 66.36, *p* < .001; see Table 6.4). The revised bifactor model–with a general factor and correlated specific factors–did not converge, so only the traditional model is taken forward to further analyses. All models showed poor cross-validation and differed substantially between the calibration and test groups (see Table 6.6).

To estimate the extent of bias in the general factor loadings due to unmodelled covariances, a sensitivity analysis was run comparing general PD factor loadings from a traditional bifactor model[21] to those from an exploratory bifactor structural equation model where unmodelled covariances are freed in the form of cross-loadings. The mean absolute difference in standardized factor loadings was .06 (*SD* = .05), suggesting that the general PD factor in the traditional bifactor model was minimally affected by unmodelled covariances. Another sensitivity analysis was run to determine the extent of bias introduced by estimating the covariances between binary indicators with robust maximum likelihood rather than robust weighted least squares. The mean absolute difference in standardized factor loadings between MLR and WLSMV factors was .04 (*SD* = .02, range = .02-.08), demonstrating that the extent of bias was minimal.

[21] The traditional bifactor model with uncorrelated specific factors was run with the robust weighted-least squares estimator rather than maximum likelihood because its comparator, the bifactor exploratory structural equation model, cannot be estimated with maximum likelihood when numerical integration (and hence maximum likelihood) was required.

Table 6.4

Model Fit Values for each CFA Model of the SCID-II Screening Questionnaire

	Fit Statistic							
Model	χ^2	*df*	CFI	TLI	RMSEA	AIC	BIC	aBIC
Single Factor	5670	1127	.79	.78	.04	68,846	69,410	69,099
Correlated Factors	2992	1112	.91	.91	.03	66,003	66,654	66,295
Bifactor (traditional)	2661	1078	.93	.92	.03	65,560	66,407	65,940
Bifactor (revised)				No convergence				

Note. χ^2 = chi-square statistic; aBIC = sample size adjusted Bayesian information criterion; CFI = Comparative Fit Index; *df* = degrees of freedom; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation.

Table 6.5

Standardized Factor Loadings for the Single Factor, Correlated Factor, and Bifactor Models of the SCID-II Screening Questionnaire

		Correlated Factors						Bifactor (Traditional)						
SCID-II Item	Single	AS	AV	BL	NS	OC	ST	G_{PD}	AS	AV	BL	NS	OC	ST
Antisocial														
Failure to conform	.86	.97						.46	.85					
Deceitfulness	.93	.94						.61	.74					
Impulsivity	.94	.98						.58	.80					
Irritable, aggressive	.83	.91						.32	.86					
Disregard for safety	.84	.96						.38	.89					
Irresponsible	.92	.94						.53	.78					
Lacks remorse	.91	.94						.45	.83					
Avoidant														
Avoids social work	.58		.70					.57		.39				
Must be liked	.62		.75					.58		.46				
Restraint in intimacy	.54		.61					.51		.31				
Preoccupied with rejection	.69		.81					.73		.34				
Socially inhibited	.60		.81					.59		.62				
Views self as inept	.63		.80					.65		.47				
No risks or new activities	.53		.66					.52		.41				
Borderline														

Avoids abandonment	.58	.66			.55	.41		
Interpersonal instability	.52	.62			.49	.48		
Identity disturbance	.64	.69			.68	.14		
Self-harming impulsivity	.59	.60			.51	.34		
Suicidality	.53	.60			.48	.38		
Affective instability	.67	.78			.64	.49		
Empty	.66	.70			.73	.07		
Intense anger	.60	.64			.50	.46		
Transient dissociation	.59	.59			.62	.07		
Narcissistic								
Grandiose	.50		.81		.23		.81	
Preoccupied with fantasies	.56		.67		.41		.52	
Believes s/he is special	.44		.79		.14		.85	
Needs admiration	.55		.72		.52		.50	
Entitlement	.54		.80		.36		.72	
Exploitative	.58		.80		.33		.73	
Lacks empathy	.60		.70		.41		.56	
Envious	.56		.58		.58		.25	
Arrogant	.53		.79		.36		.72	
Obsessive-compulsive								
Orderly	.41			.61	.37			.52
Perfectionistic	.44			.61	.46			.40
Workaholic	.22			.46	.20			.54
Moral inflexibility	.35			.52	.31			.44
Hoarding	.35			.46	.34			.31
Reluctant to delegate	.49			.73	.49			.52

Miserly	.28					.48		.25					.46	
Rigidity	.45					.56		.43					.28	
Schizotypal														
Ideas of reference	.62						.71	.67						.22
Odd beliefs	.55						.81	.41						.77
Odd perceptions	.55						.81	.45						.73
Odd thinking/speech	.40						.82	.16						.90
Suspicious	.68						.78	.63						.44
Constricted affect	.59						.85	.37						.77
Odd behavior/appearance	.35						.74	.13						.82
Lacks close friends	.41						.44	.42						.10
Social anxiety	.65						.62	.66						.13
Mean	.58	.95	.73	.65	.74	.55	.73	.46	.82	.43	.32	.63	.43	.54
SD	.16	.02	.08	.06	.08	.09	.13	.15	.05	.10	.17	.19	.10	.32
Factor correlations		AS	AV	BL	NS	OC	ST							
	AS	–												
	AV	.27	–											
	BL	.49	.70	–										
	NS	.57	.40	.61	–									
	OC	.27	.59	.60	.53	–								
	ST	.36	.61	.69	.51	.47	–							
Model-Based Reliability														
ECV/ECV$_s$								.42	.17	.05	.04	.14	.06	.12
ω/ω_s								.97	.99	.89	.88	.92	.79	.92

ω_H/ω_{Hs}	.79	.74	.31	.20	.68	.47	.56
Relative Omega	.81	.75	.34	.23	.74	.60	.61
H	.95	.94	.64	.59	.90	.67	.91
FD	.96	.98	.84	.79	.95	.83	.96

Note. AS = Antisocial; AV = Avoidant; BL = Borderline; ECV/ECV$_s$ = Explained Common Variance/Explained Common Variance-Subscale; FD = Factor Determinacy; G$_{PD}$ = General personality disorder; NS = Narcissistic; OC = Obsessive-compulsive; ω/ω_s = Omega/Omega-subscale; ω_H/ω_{Hs} = Omega hierarchical/Omega hierarchical-subscale; ST = Schizotypal.

Table 6.6

Differences in Information Criteria Between the Calibration and Test Groups for Each CFA Model of the SCID-II Screening Questionnaire

	Order A			Order B		
Model	ΔAIC	ΔBIC	ΔaBIC	ΔAIC	ΔBIC	ΔaBIC
Single Factor	648	1145	834	715	218	529
Correlated Factors	616	1189	830	588	15	374
Bifactor (traditional)	914	1659	1192	325	-420	47

Note. aBIC = sample size adjusted Bayesian information criterion. Order A and B reflect the sequence that each half of the sample was allocated as the calibration or test group. Negative values indicate that the calibration sample showed a lower (better) fit compared to the test sample.

6.3.2 Reliability Analysis

The common variance in the traditional bifactor model was roughly split between the general PD factor (42%) and uncorrelated specific factors (58%), favouring the latter. By contrast, a large proportion of the variance in raw total scores was explained by the general PD factor (ω_H = .79). The general PD factor also explained a substantial proportion of the variance in raw subscale scores for all PDs apart from the antisocial and narcissistic subscales, which were largely explained by the antisocial factor (ω_{Hs} = .74) and narcissistic factor (ω_{Hs} = .68), respectively.

Most items in the traditional bifactor model had moderate general PD factor loadings ($\bar{\lambda}$ = 0.46, SD = 0.15), the strongest being AVPD 4 'pre-occupied with rejection' (λ = .73), BPD 7 'empty' (λ = .73), and BPD 3 'identity disturbance' (λ = .68; see Table 6.5). On average, narcissistic PD items loaded strongly on the specific narcissistic factor ($\bar{\lambda}$ = 0.63, SD = 0.19) and weakly on the general PD factor ($\bar{\lambda}$ = 0.37, SD = 0.13). Similarly, antisocial PD items loaded strongly on the specific antisocial factor ($\bar{\lambda}$ = 0.82, SD = 0.05) and moderately on the general PD factor ($\bar{\lambda}$ = 0.48, SD = 0.11). The stable narcissistic and antisocial factor loadings meant that neither showed large changes from the correlated factors model to the bifactor model ($\text{MPC}_{\text{narcissistic}}$ = .11, *SD* = .12; $\text{MPC}_{\text{antisocial}}$ = .12, *SD* = .06).

By contrast, the specific borderline and avoidant PD factors explained the least amount of common variance (ECVs = .04 and .05, respectively) and raw subscale score variance (ω_{Hs} = .20 and .31, respectively). They also showed weak and moderate specific factor loadings, respectively (borderline: $\bar{\lambda}$ = 0.32, SD = 0.17; avoidant: $\bar{\lambda}$ = 0.43, SD = 0.10) and the strongest general factor loadings (borderline:

$\bar{\lambda}$ = 0.58, SD = 0.09; avoidant: $\bar{\lambda}$ = 0.59, SD = 0.08). Avoidant and borderline PD items showed notable declines in loading strength between the correlated factor and bifactor models ($MPC_{avoidant}$ = .31, *SD* = .09; $MPC_{antisocial}$ = .34, *SD* = .18) and hence were influenced by the common variance.

Some schizotypal PD items loaded preferentially onto the specific schizotypy factor ($\bar{\lambda}$ = 0.54, SD = 0.32), particularly STPD 2 'odd thinking/speech' (λ = .90) and STPD 7 'odd behaviour/appearance' (λ = .82). Other items loaded preferentially onto the general PD factor ($\bar{\lambda}$ = 0.43, SD = 0.20), including STPD 5 '(λ = .68) and STPD 9 'social anxiety' (λ = .65). Consequently, the mean parameter change (MPC) between the correlated factor and bifactor models was small but variable (MPC = .15, *SD* = .23). Obsessive-compulsive PD items showed weak general PD factor loadings ($\bar{\lambda}$ = 0.36, SD = 0.10) and moderate obsessive-compulsive specific factor loadings ($\bar{\lambda}$ = 0.43, SD = 0.10), which changed little between the correlated factor and bifactor models (MPC = .12, *SD* = .12).

H values generally followed a similar pattern to ECV and omega values. That is, factor scores for the general PD factor, as well as the specific PD factors that were least influenced by the shared variance (e.g., antisocial, narcissistic, and to a lesser degree, obsessive-compulsive and schizotypal), had H values that exceeded or neared the suggested cut-off of .70 for acceptable reliability (see Table 6.5). By contrast, avoidant (H = .64) and borderline (H = .59) factor scores both fell below the cut-off and hence less adequately represented by their items.

6.3.3 Latent Growth Curve Models

An unconditional growth model with an intercept factor and linear slope factor showed a good-to-excellent fit (CFI = .96, TLI = .95, RMSEA = .07, SRMR = .02). Adding a quadratic slope factor improved the information explained (ΔAIC = 68; ΔBIC = 45; ΔaBIC = 58; model fit: CFI = 1, TLI = 1, RMSEA = 0, SRMR = 0). The intercept mean (i.e. predicted PHQ-9 score at week two pooled across patients) fell just under the PHQ-9's clinical threshold of 10 (*b* = 9.56, *z* = 66.39, *p* < .001, 95% CI [9.27, 9.84]), but patients varied substantially around the mean (37.56, *z* = 13.13, *p* < .001, 95% CI [31.94, 43.16]). On average, patients showed a linear decline in PHQ-9 scores over the treatment period (*b* = -2.40, *z* = -16.87, *p* < .001, 95% CI [-2.68, -2.12]; see Figure 6.1a, 'overall' growth curve), but varied in the steepness of their individual slopes (13.76, *z* = 3.81, *p* < .001, 95% CI [6.67, 20.83]). The rate of decline in PHQ-9 scores slowed with time (*b* = 0.34, *z* = 6.81, *p* < .001, 95% CI [0.24, 0.43]; see Figure 6.1c, 'overall' growth curve), but again, patients varied in the extent of this quadratic pattern of change (1.09, *z* = 3.28, *p* < .001, 95% CI [0.44, 1.74]).

Bifactor growth model. In the part conditional growth model with the general and uncorrelated specific PD factors and clinical covariates, higher intercept values (i.e. week two PHQ-9 scores) were predicted by higher general PD factor scores at admission (*b* = 1.16, *z* = 6.49, *p* < .001, 95% CI [0.81, 1.51]), and marginally lower borderline scores (*b* = -0.49, *z* = -1.86, *p* = .062, 95% CI [-1.00, 0.03]), lower antisocial scores (*b* = -0.55, *z* = -2.14, *p* = .032, 95% CI [-1.05, -0.05]), and lower narcissistic scores (*b* = -0.38, *z* = -1.96, *p* = .050, 95% CI [-0.77, 0]). The general PD factor did not significantly predict individual differences in the rate of linear decline (*b* = -0.09, *z* = -0.42, *p* = .678, 95% CI [-0.53, 0.35]). By contrast, higher specific borderline scores predicted flatter linear slopes and thus slower decline (*b* = 0.58, *z* =

1.97, *p* = .049, 95% CI [0.01, 1.16]), while higher antisocial scores predicted a stronger quadratic (i.e. U-shaped) pattern of growth (*b* = 0.25, *z* = 2.26, *p* = .024, 95% CI [0.03, 0.46]). Regression coefficients for the clinical covariates matched those in the full conditional growth model (see below).

In the full conditional model with general and uncorrelated specific PD factors, clinical covariates, and demographic covariates, higher intercept values were still predicted by higher general PD scores (*b* = 1.14, *z* = 6.36, *p* < .001, 95% CI [0.79, 1.49]), lower borderline scores (*b* = -0.64, *z* = -2.47, *p* = .013, 95% CI [-1.14, -0.13]), and lower antisocial scores (*b* = -0.51, *z* = -1.99, *p* = .047, 95% CI [-1.02, -0.01]). The negative association between general PD and the linear slope factor was stronger but did not reach significance (*b* = -0.22, *z* = -0.96, *p* = .340, 95% CI [-0.66, 0.23]). Moreover, the positive association between borderline scores and linear slopes decreased slightly and was now marginal (*b* = 0.52, *z* = 1.75, *p* = .08, 95% CI [-0.06, 1.11]), while the positive association between antisocial scores and quadratic slopes increased slightly and remained significant (*b* = 0.26, *z* = 2.36, *p* = .018, 95% CI [0.04, 0.47]). Figure 6.1 shows the growth curves predicted by the general, borderline, and antisocial factors, and Table 6.7 shows the regression coefficients for the remaining PD factors, clinical covariates, and demographic covariates.

The strongest clinical covariate was PHQ-9 scores at admittance; higher initial PHQ-9 scores predicted higher intercept scores at week 2 (*b* = 0.51, *z* = 28.13, *p* < .001, 95% CI [0.47, 0.54]) and faster rates of linear decline (*b* = -0.12, *z* = -4.90, *p* < .001, 95% CI [-0.17, -0.07]). Influential demographic covariates included sex, with males reporting lower PHQ-9 scores at week two than females (*b* = -0.96, *z* = -3.92, *p* < .001, 95% CI [-1.43, -0.48]), and age, with older patients showing steeper declines in PHQ-9 scores (*b* = -0.04 *z* = -2.58, *p* = .01, 95% CI [-0.07, -0.01]). The full conditional model explained 57% of the variance in the intercept factor, 24% of the variance in the linear slope factor, and 16% of the variance in the quadratic slope factor.

Correlated factors growth model. In the part conditional growth model with correlated PD factors (including antisocial, avoidant, borderline, narcissistic, obsessive-compulsive, and schizotypal PD factors) and clinical covariates, the schizotypal PD factor at admission predicted higher intercept scores (i.e. week 2 PHQ-9 scores; b = .80, z = 2.03, p = .043, 95% CI [0.03, 1.57]) and steeper linear declines in PHQ-9 scores (b = -0.94, z = -2.14, p = .033, 95% CI [-1.80, -0.08]). Higher borderline factor scores predicted stronger inverted U-shaped changes in PHQ-9 scores (b = -0.40, z = -2.48, p = .013, 95% CI [-0.71, -0.08]), while higher antisocial factor scores predicted marginally stronger U-shaped changes (b = 0.25, z = 1.90, p = .058, 95% CI [-0.01, 0.51]). Regression coefficients for the clinical covariates matched those in the full conditional growth model (see below).

In the full conditional model with correlated PD factors, clinical covariates, and demographic covariates, higher schizotypal PD factor scores continued to predict higher intercept values (b = .80, z = 2.03, p = .043, 95% CI [0.03, 1.58]) and steeper linear declines (b = -0.98, z = -2.23, p = .026, 95% CI [-1.83, -0.12]). Moreover, higher borderline scores continued to predict stronger inverted U-shaped quadratic growth (b = -0.36, z = -2.25, p = .024, 95% CI [-0.68, -0.05]), while higher antisocial scores significantly predicted stronger U-shaped growth (b = 0.26, z = 1.99, p = .046, 95% CI [0, 0.52]). Figure 6.2 shows the growth curves predicted by the schizotypal, borderline, and antisocial correlated factors, and Table 6.8 shows the regression coefficients for the remaining PD factors, clinical covariates, and demographic covariates. The full conditional model explained 55% of the variance in the intercept factor, 23% of the variance in the linear slope factor, and 13% of the variance in the quadratic slope factor.

For reference, models with correlated PD factors or bifactor PD factors predicting the growth factors without clinical or demographic covariates were ran (see Table 6.9 for regression coefficients). In the correlated factor growth model, all PD factors apart from schizotypal and obsessive-compulsive PDs significantly predicted the intercept factor, with higher avoidant and borderline scores predicting higher intercept scores and lower antisocial and narcissistic scores predicting lower intercept scores. Contrary to the correlated factors growth model with covariates, no PD factor predicted variation in the linear slope factor. However, the borderline and antisocial factors continued to negatively and positively predict the quadratic slope factor, respectively. The bifactor growth model without covariates produced similar estimates to the bifactor growth model with covariates, except that the general PD factor's negative association with the linear slope was now significant.

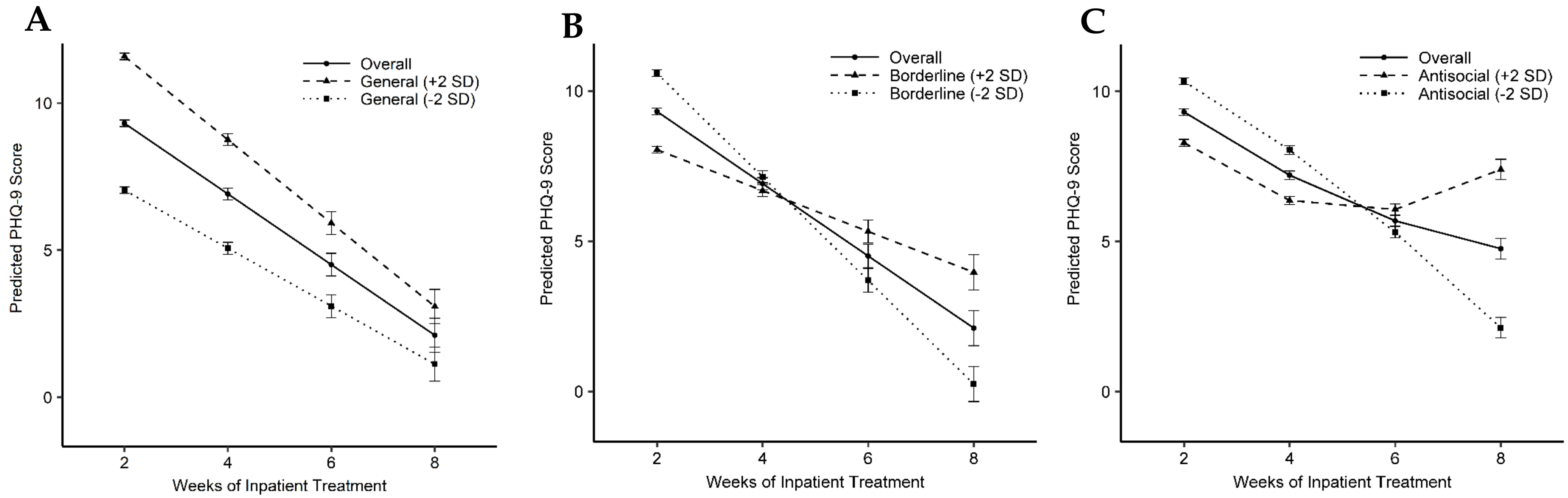

Figure 6.1. Growth curves for PHQ-9 scores predicted by the general and specific PD factors. (A) The linear slope factor for general PD scores +/- 2 standard deviations (SDs) from the mean; (B) The linear slope factor for specific borderline factor scores +/- 2 SDs from the mean; (C) The quadratic slope factor for specific antisocial factor scores +/- 2 SDs from the mean. The 'Overall' slope in each sub-figure reflects the linear or quadratic slope holding the general and specific factors constant. All growth factors are controlled for centred clinical and demographic covariates. Error bars reflect standard errors of the predicted means.

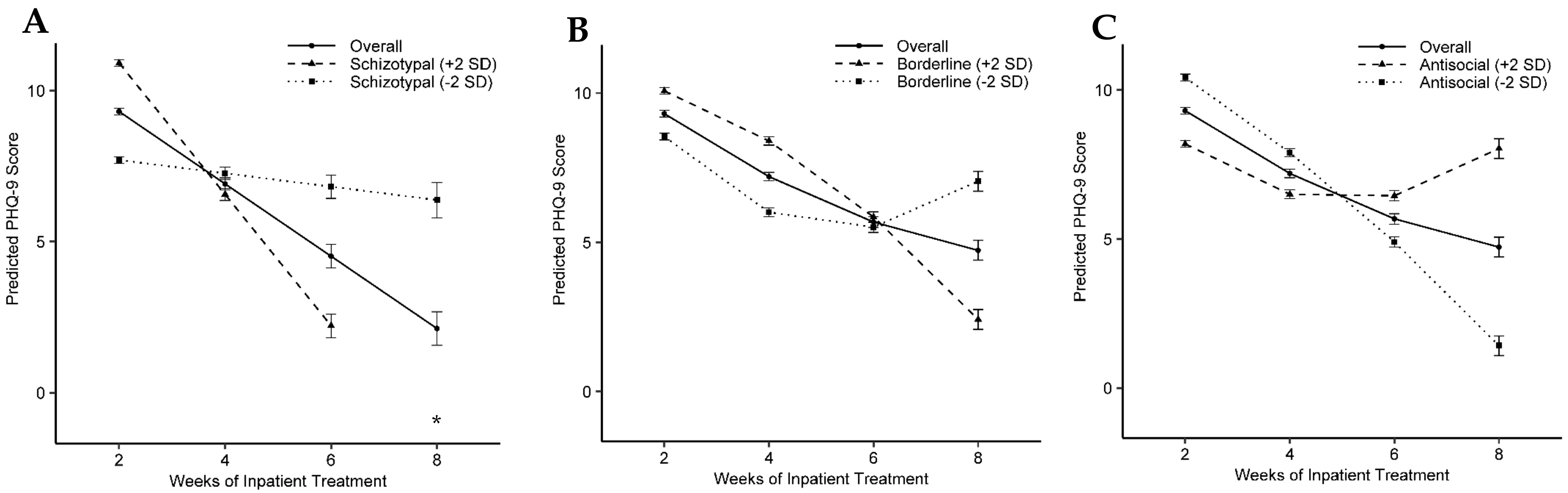

Figure 6.2. Growth curves for PHQ-9 scores predicted by the correlated PD factors. (A) The linear slope factor for schizotypal factor scores +/- 2 standard deviations (SDs) from the mean; (B) The quadratic slope factor for borderline factor scores +/- 2 SDs from the mean; (C) The quadratic slope factor for antisocial factor scores +/- 2 SDs from the mean. The 'Overall' slope in each sub-figure reflects the linear or quadratic slope holding all PD factors constant. All growth factors are controlled for centred clinical and demographic covariates. Error bars reflect standard errors of the predicted means. *PHQ-9 scores at week 8 predicted by schizotypal PD +2 SDs was not plotted as it fell outside the admissible range (e.g., -2.13).

Table 6.7

Standardized (b) and Unstandardized (B) Regression Coefficients for the Personality Disorder Factors, Clinical Covariates, and Demographic Covariates Predicting the Intercept, Linear Slope, and Quadratic Slope Factors in the Bifactor Growth Model

	Estimate			
Variable	*b* (95% CI)	*B* (95% CI)	*z*	*p*
Intercept				
Mean	**9.31 (9.09, 9.53)**	**1.64 (1.58, 1.70)**	**82.8**	**< .001**
Variance	**13.95 (12.38, 15.51)**	**0.43 (0.39, 0.48)**	**17.44**	**< .001**
PD Factor				
General	**1.14 (0.79, 1.49)**	**0.20 (0.14, 0.26)**	**6.36**	**< .001**
Antisocial	**-0.51 (-1.02, -0.01)**	**-0.09 (-0.18, 0)**	**-1.99**	**0.047**
Avoidant	-0.32 (-0.75, 0.12)	-0.06 (-0.13, 0.02)	-1.43	0.153
Borderline	**-0.64 (-1.14, -0.13)**	**-0.11 (-0.20, -0.02)**	**-2.47**	**0.013**
Narcissistic	-0.27 (-0.65, 0.12)	-0.05 (-0.12, 0.02)	-1.35	0.176
Obsessional	-0.28 (-0.7, 0.14)	-0.05 (-0.12, 0.02)	-1.32	0.186
Schizotypal	0.27 (-0.24, 0.78)	0.05 (-0.04, 0.14)	1.02	0.306
Clinical				
PHQ-9 Baseline	**0.51 (0.47, 0.54)**	**0.64 (0.60, 0.68)**	**28.13**	**< .001**
Length of stay	**0.05 (0.04, 0.06)**	**0.19 (0.14, 0.24)**	**7.69**	**< .001**
Episode Number	**1.45 (0.77, 2.13)**	**0.09 (0.04, 0.13)**	**4.16**	**< .001**
Unit (Hope v Compass)	-0.33 (-1.04, 0.38)	-0.03 (-0.08, 0.03)	-0.91	0.365
Unit (CPAS v Compass)	0.62 (-0.19, 1.43)	0.04 (-0.01, 0.09)	1.51	0.132
Unit (PIC v Compass)	-0.32 (-1.16, 0.53)	-0.02 (-0.09, 0.04)	-0.73	0.463
Demographic				
Sex	**-0.96 (-1.43, -0.48)**	**-0.08 (-0.13, -0.04)**	**-3.92**	**< .001**

Age	-0.01 (-0.03, 0.01)	-0.03 (-0.09, 0.03)	-0.88	0.381
Ethnic group	-0.55 (-1.25, 0.15)	-0.03 (-0.07, 0.01)	-1.53	0.125
Education	-0.20 (-0.74, 0.34)	-0.02 (-0.07, 0.03)	-0.73	0.466
Marital Status	0.20 (-0.42, 0.82)	0.02 (-0.03, 0.07)	0.64	0.524
Linear Slope				
Mean	**-2.40 (-2.80, -2.01)**	**-0.72 (-0.90, -0.54)**	**-11.93**	**< .001**
Variance	**8.39 (4.84, 11.93)**	**0.76 (0.61, 0.90)**	**4.64**	**< .001**
PD Factor				
General	-0.22 (-0.66, 0.23)	-0.07 (-0.20, 0.07)	-0.96	0.340
Antisocial	-0.17 (-0.83, 0.49)	-0.05 (-0.25, 0.15)	-0.50	0.620
Avoidant	0.17 (-0.35, 0.69)	0.05 (-0.10, 0.21)	0.64	0.522
Borderline	**0.52 (-0.06, 1.11)**	**0.16 (-0.01, 0.33)**	**1.75**	**0.080†**
Narcissistic	0.14 (-0.35, 0.63)	0.04 (-0.10, 0.19)	0.55	0.580
Obsessional	0.38 (-0.11, 0.88)	0.12 (-0.03, 0.26)	1.51	0.131
Schizotypal	-0.55 (-1.19, 0.10)	-0.16 (-0.36, 0.03)	-1.59	0.111
Clinical				
PHQ-9 Baseline	**-0.12 (-0.17, -0.07)**	**-0.26 (-0.37, -0.15)**	**-4.90**	**< .001**
Length of stay	**0.04 (0.02, 0.06)**	**0.24 (0.12, 0.36)**	**4**	**< .001**
Episode Number	-0.18 (-1.13, 0.78)	-0.02 (-0.11, 0.08)	-0.36	0.719
Unit (Hope v Compass)	0.59 (-0.31, 1.49)	0.08 (-0.04, 0.20)	1.28	0.199
Unit (CPAS v Compass)	**2.57 (0.72, 4.42)**	**0.29 (0.09, 0.49)**	**2.73**	**0.006**
Unit (PIC v Compass)	0.60 (-0.54, 1.75)	0.08 (-0.07, 0.22)	1.03	0.302
Demographic				
Sex	-0.04 (-0.68, 0.60)	-0.01 (-0.10, 0.09)	-0.13	0.897
Age	**-0.04 (-0.07, -0.01)**	**-0.19 (-0.32, -0.05)**	**-2.58**	**0.010**
Ethnic group	0.31 (-0.67, 1.29)	0.03 (-0.06, 0.11)	0.62	0.533
Education	-0.24 (-0.97, 0.50)	-0.04 (-0.15, 0.08)	-0.62	0.532
Marital Status	-0.06 (-0.89, 0.77)	-0.01 (-0.12, 0.10)	-0.14	0.887

Quadratic Slope				
Mean	**0.30 (0.13, 0.46)**	**0.27 (0.10, 0.44)**	**3.54**	**< .001**
Variance	**0.99 (0.37, 1.61)**	**0.84 (0.66, 1.01)**	**3.13**	**0.002**
PD Factor				
General	-0.06 (-0.21, 0.10)	-0.05 (-0.20, 0.09)	-0.72	0.472
Antisocial	**0.26 (0.04, 0.47)**	**0.24 (0.03, 0.45)**	**2.36**	**0.018**
Avoidant	0.07 (-0.11, 0.24)	0.06 (-0.10, 0.22)	0.77	0.441
Borderline	-0.11 (-0.31, 0.09)	-0.10 (-0.28, 0.08)	-1.08	0.278
Narcissistic	0.01 (-0.17, 0.19)	0.01 (-0.16, 0.18)	0.09	0.930
Obsessional	-0.04 (-0.22, 0.14)	-0.03 (-0.20, 0.13)	-0.41	0.681
Schizotypal	0.15 (-0.08, 0.39)	0.14 (-0.08, 0.36)	1.29	0.199
Clinical				
PHQ-9 Baseline	0.01 (-0.01, 0.03)	0.07 (-0.04, 0.19)	1.21	0.226
Length of stay	**-0.01 (-0.02, 0)**	**-0.17 (-0.30, -0.03)**	**-2.42**	**0.016**
Episode Number	0.17 (-0.15, 0.50)	0.05 (-0.05, 0.16)	1.04	0.301
Unit (Hope v Compass)	-0.15 (-0.46, 0.17)	-0.06 (-0.19, 0.07)	-0.92	0.360
Unit (CPAS v Compass)	**-0.62 (-1.31, 0.07)**	**-0.21 (-0.45, 0.03)**	**-1.75**	**0.080†**
Unit (PIC v Compass)	-0.11 (-0.53, 0.31)	-0.04 (-0.21, 0.12)	-0.51	0.609
Demographic				
Sex	0.04 (-0.19, 0.27)	0.02 (-0.09, 0.12)	0.33	0.743
Age	**0.01 (0, 0.02)**	**0.13 (-0.03, 0.28)**	**1.68**	**0.093†**
Ethnic group	-0.10 (-0.45, 0.25)	-0.03 (-0.12, 0.07)	-0.55	0.583
Education	0.16 (-0.11, 0.42)	0.07 (-0.05, 0.19)	1.17	0.240
Marital Status	0.06 (-0.24, 0.36)	0.02 (-0.10, 0.15)	0.38	0.704

Note. PD = personality disorder; PHQ-9 = Patient Health Questionnaire. Significant coefficients are in bold.
†Marginal result ($p < .1$)

Table 6.8

Standardized (b) and Unstandardized (B) Regression Coefficients for the Personality Disorder Factors, Clinical Covariates, and Demographic Covariates Predicting the Intercept, Linear Slope, and Quadratic Slope Factors in the Correlated Factors Growth Model

	Estimate			
Variable	*b* (95% CI)	*B* (95% CI)	*z*	*P*
Intercept				
Mean	**9.30 (9.08, 9.53)**	**1.64 (1.58, 1.71)**	**82.69**	**< .001**
Variance	**14.29 (12.87, 15.71)**	**0.45 (0.40, 0.49)**	**19.73**	**< .001**
PD Factor				
Antisocial	-0.56 (-1.29, 0.17)	-0.1 (-0.23, 0.03)	-1.50	0.135
Avoidant	0.08 (-0.51, 0.67)	0.01 (-0.09, 0.12)	0.27	0.79
Borderline	0.38 (-0.38, 1.15)	0.07 (-0.07, 0.2)	0.98	0.328
Narcissistic	-0.03 (-0.68, 0.62)	-0.01 (-0.12, 0.11)	-0.10	0.924
Obsessional	0.06 (-0.54, 0.66)	0.01 (-0.09, 0.12)	0.21	0.834
Schizotypal	**0.80 (0.03, 1.58)**	**0.14 (0.01, 0.28)**	**2.03**	**0.043**
Clinical				
PHQ-9 Baseline	**0.52 (0.48, 0.55)**	**0.66 (0.62, 0.69)**	**29.32**	**< .001**
Length of stay	**0.05 (0.04, 0.06)**	**0.19 (0.14, 0.24)**	**7.72**	**< .001**
Episode Number	**1.47 (0.78, 2.15)**	**0.09 (0.05, 0.13)**	**4.18**	**< .001**
Unit (Hope v Compass)	-0.35 (-1.06, 0.37)	-0.03 (-0.08, 0.03)	-0.95	0.342
Unit (CPAS v Compass)	0.57 (-0.24, 1.37)	0.04 (-0.01, 0.09)	1.37	0.170
Unit (PIC v Compass)	-0.48 (-1.32, 0.36)	-0.04 (-0.1, 0.03)	-1.13	0.260
Demographic				
Sex	**-0.87 (-1.34, -0.39)**	**-0.08 (-0.12, -0.03)**	**-3.55**	**< .001**
Age	-0.01 (-0.03, 0.01)	-0.03 (-0.09, 0.04)	-0.77	0.439

Ethnic group	-0.57 (-1.28, 0.13)	-0.03 (-0.07, 0.01)	-1.58	0.113
Education	-0.19 (-0.73, 0.35)	-0.02 (-0.06, 0.03)	-0.69	0.490
Marital Status	0.23 (-0.39, 0.85)	0.02 (-0.03, 0.07)	0.72	0.471
Linear Slope				
Mean	**-2.39 (-2.79, -2)**	**-0.74 (-0.93, -0.55)**	**-11.82**	**< .001**
Variance	**8.17 (4.69, 11.64)**	**0.77 (0.63, 0.91)**	**4.61**	**< .001**
PD Factor				
Antisocial	-0.06 (-0.86, 0.75)	-0.02 (-0.26, 0.23)	-0.13	0.894
Avoidant	-0.07 (-0.74, 0.59)	-0.02 (-0.23, 0.18)	-0.22	0.828
Borderline	0.58 (-0.34, 1.50)	0.18 (-0.11, 0.46)	1.23	0.219
Narcissistic	0.08 (-0.68, 0.83)	0.02 (-0.21, 0.26)	0.19	0.846
Obsessional	0.27 (-0.38, 0.92)	0.08 (-0.12, 0.28)	0.81	0.417
Schizotypal	**-0.98 (-1.83, -0.12)**	**-0.30 (-0.56, -0.04)**	**-2.23**	**0.026**
Clinical				
PHQ-9 Baseline	**-0.13 (-0.17, -0.08)**	**-0.28 (-0.38, -0.17)**	**-5.16**	**< .001**
Length of stay	**0.04 (0.02, 0.06)**	**0.25 (0.13, 0.37)**	**3.95**	**< .001**
Episode Number	-0.17 (-1.13, 0.79)	-0.02 (-0.12, 0.08)	-0.35	0.728
Unit (Hope v Compass)	0.58 (-0.32, 1.48)	0.08 (-0.04, 0.20)	1.27	0.203
Unit (CPAS v Compass)	**2.60 (0.75, 4.45)**	**0.30 (0.09, 0.51)**	**2.75**	**0.006**
Unit (PIC v Compass)	0.71 (-0.43, 1.84)	0.09 (-0.06, 0.24)	1.22	0.225
Demographic				
Sex	-0.09 (-0.73, 0.55)	-0.01 (-0.11, 0.08)	-0.28	0.782
Age	**-0.04 (-0.07, -0.01)**	**-0.19 (-0.33, -0.05)**	**-2.61**	**0.009**
Ethnic group	0.30 (-0.67, 1.27)	0.03 (-0.06, 0.12)	0.60	0.548
Education	-0.24 (-0.97, 0.50)	-0.04 (-0.15, 0.08)	-0.63	0.526
Marital Status	-0.08 (-0.91, 0.76)	-0.01 (-0.12, 0.10)	-0.18	0.858
Quadratic Slope				
Mean	**0.29 (0.13, 0.45)**	**0.26 (0.10, 0.43)**	**3.48**	**0.001**

Variance	**1.04 (0.44, 1.65)**	**0.87 (0.72, 1.01)**	**3.36**	**0.001**
PD Factor				
Antisocial	**0.26 (0, 0.52)**	**0.24 (-0.01, 0.49)**	**1.99**	**0.046**
Avoidant	0.12 (-0.11, 0.35)	0.11 (-0.10, 0.31)	1.04	0.299
Borderline	**-0.36 (-0.68, -0.05)**	**-0.33 (-0.64, -0.03)**	**-2.25**	**0.024**
Narcissistic	-0.07 (-0.36, 0.22)	-0.06 (-0.33, 0.20)	-0.47	0.638
Obsessional	-0.04 (-0.27, 0.2)	-0.04 (-0.25, 0.18)	-0.32	0.750
Schizotypal	0.19 (-0.12, 0.50)	0.17 (-0.12, 0.46)	1.18	0.237
Clinical				
PHQ-9 Baseline	0.01 (-0.01, 0.03)	0.07 (-0.04, 0.18)	1.19	0.236
Length of stay	**-0.01 (-0.02, 0)**	**-0.16 (-0.29, -0.02)**	**-2.33**	**0.02**
Episode Number	0.16 (-0.17, 0.49)	0.05 (-0.05, 0.15)	0.96	0.336
Unit (Hope v Compass)	-0.14 (-0.46, 0.17)	-0.06 (-0.18, 0.07)	-0.87	0.382
Unit (CPAS v Compass)	**-0.62 (-1.31, 0.06)**	**-0.21 (-0.45, 0.02)**	**-1.78**	**0.075†**
Unit (PIC v Compass)	-0.12 (-0.54, 0.29)	-0.05 (-0.21, 0.11)	-0.58	0.561
Demographic				
Sex	0.05 (-0.18, 0.28)	0.02 (-0.08, 0.13)	0.42	0.671
Age	**0.01 (0, 0.02)**	**0.13 (-0.02, 0.28)**	**1.72**	**0.085†**
Ethnic group	-0.09 (-0.44, 0.26)	-0.02 (-0.12, 0.07)	-0.50	0.616
Education	0.16 (-0.11, 0.42)	0.07 (-0.05, 0.19)	1.16	0.246
Marital Status	0.06 (-0.24, 0.36)	0.02 (-0.10, 0.15)	0.38	0.703

Note. PD = personality disorder; PHQ-9 = Patient Health Questionnaire. Significant coefficients are in bold.
†Marginal result (p < .1)

Table 6.9

Standardized (b) and Unstandardized (B) Regression Coefficients for the Personality Disorder Factors Alone Predicting the Intercept, Linear Slope, and Quadratic Slope Factors in the Correlated Factors Growth Model Followed by the Bifactor Growth Model

	Estimate			
Variable	*b* (95% CI)	*B* (95% CI)	*z*	*P*
Correlated Factors Growth Model				
Intercept	**9.46 (9.18, 9.74)**	**1.69 (1.61, 1.76)**	**65.95**	**< .001**
Mean	**22.70 (20.55, 24.84)**	**0.72 (0.66, 0.79)**	**20.70**	**< .001**
Variance				
PD Factor				
Antisocial	**-1.25 (-2.33, -0.18)**	**-0.22 (-0.41, -0.04)**	**-2.29**	**0.022**
Avoidant	**0.92 (0.15, 1.70)**	**0.17 (0.03, 0.30)**	**2.34**	**0.019**
Borderline	**1.85 (0.79, 2.90)**	**0.33 (0.14, 0.52)**	**3.44**	**0.001**
Narcissistic	**-1.10 (-2.09, -0.10)**	**-0.20 (-0.37, -0.02)**	**-2.16**	**0.031**
Obsessional	0.65 (-0.16, 1.46)	0.12 (-0.03, 0.26)	1.57	0.116
Schizotypal	0.85 (-0.31, 2)	0.15 (-0.05, 0.35)	1.44	0.15
Linear Slope				
Mean	**-2.41 (-2.68, -2.13)**	**-0.87 (-1.12, -0.63)**	**-16.89**	**< .001**
Variance	**7.22 (3.59, 10.84)**	**0.95 (0.86, 1.04)**	**3.90**	**< .001**
PD Factor				
Antisocial	0.35 (-0.46, 1.16)	0.13 (-0.16, 0.42)	0.85	0.395
Avoidant	-0.24 (-0.86, 0.38)	-0.09 (-0.31, 0.14)	-0.75	0.453
Borderline	0.30 (-0.61, 1.2)	0.11 (-0.22, 0.44)	0.64	0.520

Narcissistic	-0.08 (-0.83, 0.67)	-0.03 (-0.30, 0.24)	-0.20	0.839
Obsessional	0.20 (-0.45, 0.84)	0.07 (-0.16, 0.31)	0.61	0.545
Schizotypal	-0.67 (-1.49, 0.16)	-0.24 (-0.55, 0.06)	-1.58	0.114
Quadratic Slope				
Mean	**0.34 (0.24, 0.44)**	**0.32 (0.19, 0.46)**	**6.74**	**< .001**
Variance	**1.03 (0.41, 1.65)**	**0.94 (0.85, 1.04)**	**3.27**	**0.001**
PD Factor				
Antisocial	0.21 (-0.04, 0.47)	0.20 (-0.05, 0.46)	1.63	0.103
Avoidant	0.10 (-0.11, 0.32)	0.1 (-0.1, 0.3)	0.96	0.337
Borderline	**-0.37 (-0.67, -0.06)**	**-0.35 (-0.66, -0.04)**	**-2.34**	**0.019**
Narcissistic	0.02 (-0.27, 0.31)	0.02 (-0.25, 0.3)	0.16	0.876
Obsessional	-0.04 (-0.27, 0.2)	-0.04 (-0.26, 0.19)	-0.31	0.759
Schizotypal	0.11 (-0.19, 0.4)	0.1 (-0.19, 0.39)	0.71	0.480
Bifactor Growth Model				
Intercept				
Mean	**9.50 (9.22, 9.78)**	**1.64 (1.57, 1.71)**	**66.64**	**< .001**
Variance	**20.03 (17.09, 22.98)**	**0.60 (0.50, 0.69)**	**13.33**	**< .001**
PD Factor				
General	**2.99 (2.63, 3.35)**	**0.52 (0.46, 0.57)**	**16.26**	**< .001**
Antisocial	**-0.95 (-1.62, -0.28)**	**-0.16 (-0.28, -0.05)**	**-2.76**	**0.006**
Avoidant	-0.38 (-0.93, 0.18)	-0.07 (-0.16, 0.03)	-1.33	0.184
Borderline	**-1.21 (-1.93, -0.49)**	**-0.21 (-0.33, -0.09)**	**-3.31**	**0.001**
Narcissistic	**-1.44 (-1.93, -0.95)**	**-0.25 (-0.33, -0.17)**	**-5.74**	**< .001**
Obsessional	-0.21 (-0.71, 0.29)	-0.04 (-0.12, 0.05)	-0.83	0.408
Schizotypal	-0.05 (-0.69, 0.58)	-0.01 (-0.12, 0.10)	-0.17	0.868
Linear Slope				

Mean	**-2.43 (-2.70, -2.15)**	**-0.79 (-0.99, -0.59)**	**-17.07**	**< .001**
Variance	**8.54 (4.79, 12.28)**	**0.90 (0.79, 1.02)**	**4.46**	**< .001**
PD Factor				
General	**-0.41 (-0.78, -0.03)**	**-0.13 (-0.25, -0.02)**	**-2.13**	**0.034**
Antisocial	0.11 (-0.57, 0.79)	0.04 (-0.18, 0.25)	0.31	0.757
Avoidant	0.21 (-0.28, 0.70)	0.07 (-0.09, 0.23)	0.84	0.399
Borderline	**0.70 (0.13, 1.26)**	**0.23 (0.06, 0.40)**	**2.43**	**0.015**
Narcissistic	0.24 (-0.25, 0.74)	0.08 (-0.08, 0.24)	0.96	0.337
Obsessional	0.30 (-0.19, 0.79)	0.10 (-0.06, 0.26)	1.20	0.231
Schizotypal	-0.22 (-0.85, 0.40)	-0.07 (-0.28, 0.13)	-0.70	0.484
Quadratic Slope				
Mean	**0.35 (0.25, 0.45)**	**0.35 (0.2, 0.49)**	**7.04**	**< .001**
Variance	**0.95 (0.32, 1.57)**	**0.92 (0.8, 1.05)**	**2.97**	**0.003**
PD Factor				
General	-0.09 (-0.22, 0.04)	-0.09 (-0.23, 0.04)	-1.38	0.169
Antisocial	**0.21 (-0.01, 0.42)**	**0.20 (-0.02, 0.43)**	**1.86**	**0.063**
Avoidant	0.05 (-0.11, 0.21)	0.05 (-0.11, 0.21)	0.63	0.527
Borderline	-0.12 (-0.31, 0.06)	-0.12 (-0.3, 0.06)	-1.3	0.195
Narcissistic	0.03 (-0.15, 0.21)	0.03 (-0.15, 0.21)	0.35	0.728
Obsessional	-0.01 (-0.19, 0.17)	-0.01 (-0.18, 0.16)	-0.11	0.910
Schizotypal	0.10 (-0.12, 0.32)	0.10 (-0.13, 0.32)	0.86	0.387

Note. PD = personality disorder; PHQ-9 = Patient Health Questionnaire. Significant coefficients are in bold.
†Marginal result (*p* < .1)

6.4 Discussion

Findings have been mixed as to whether PDs predict differential responses to treatment for depression. A complicating factor is that current assessment measures conflate what is shared among PDs (i.e. severity) with what is specific to particular PDs (i.e. style; Hopwood et al., 2011). The shared and specific aspects of PDs might predict depression outcomes in opposite directions, contributing to the mixed findings. The current chapter investigated the unique contributions of the shared and specific components of PDs to depression outcomes by first separating out these two sources of variance with the bifactor model, and then using the resultant general and specific PD factors to predict changes in depression severity over an inpatient treatment.

Covariation in PD symptom reports was best explained by a general PD factor, as well as uncorrelated specific factors reflecting each PD assessed. The general PD factor predicted higher initial depression scores, but not differential rates of change. By contrast, the specific borderline factor predicted slower rates of decline over the over the treatment period, while the antisocial factors predicted a U-shaped pattern of change. Each finding is interpreted in turn.

6.4.1 Does the Latent Structure of Personality Disorders Follow a Bifactor Model?

Consistent with past studies, covariation in PD symptom responses was best explained by a bifactor model with a general PD factor and specific antisocial, avoidant, borderline, narcissistic, obsessive, and schizotypal PD factors (Conway et al., 2016; Jahng et al., 2011; Sharp et al., 2015; Williams et al., 2017; Wright et al., 2016). Informal and formal tests of information criteria supported the bifactor model over the correlated factors model and a single-factor model. Comparing models with information criteria that penalize for model complexity was important because the bifactor model has a highest fitting propensity (Bonifay & Cai, 2017; Murray & Johnson, 2013).

In absolute terms, the bifactor and correlated factor models both showed acceptable fit which did not differ substantially from each other. This is similar to bifactor studies of psychopathology that show near-equivalent fit between the bifactor and correlated factor models (Caspi et al., 2014; Conway et al., 2019; Lahey et al., 2012; Gomez et al., 2019; Haltigan et al., 2018; Laceulle et al., 2016; Patalay et al., 2015; Snyder et al., 2017; St Clair et al., 2017; Tackett et al., 2013). It was therefore important to compare these models with alternative means such as the residual correlation matrix and cross-validation. Problematic residual correlations in the bifactor and correlated factor models were minimal, but the bifactor model tended to misrepresent correlations involving antisocial PD items (further issues associated with antisocial PD items are discussed in the 'Strengths and Limitations' section). Moreover, cross-validation tests showed that parameters in all models were poorly replicated between the calibration and test halves of the sample, which likely reflects the short-comings of improper cross-validation tests (e.g., lack of an independent test sample and power reductions from splitting the sample).

Most bifactor studies of PD feature correlated specific factors (Conway et al., 2016; Sharp et al., 2015; Williams et al., 2017), yet such a model did not converge in the current study. This might be because prior studies used exploratory bifactor methods that are robust to oblique rotations (Morin, Arens, & Marsh, 2016); confirmatory bifactor models with correlated specific factors lack stability and converge less often (Greene et al., 2019). Fixing the correlations between specific factors simplifies their interpretation: specific PD factors reflect stylistic expressions of symptoms free from the attributes common to all PDs (Wright et al., 2016). The benefit of uncorrelated specific factors is particularly important to interpreting their unique influence on depression outcomes; correlating the specific PD factors would remove their unique prediction and contradict the goal of estimating specific factors residualized for the common variance.

A disadvantage of uncorrelated specific factors is that unmodelled covariances might be expressed through inflated general factor loadings (Greene et al., 2019; Murray & Johnson, 2013; Reise, Moore, & Maydeu-Olivares, 2011). In the current study, there were minor differences between the general PD factor loadings from a confirmatory bifactor model with orthogonal specific factors, and an exploratory bifactor structural equation model (ESEM) with cross-loadings freed. Had the general PD factor in the confirmatory model been influenced by unmodelled covariances, then it should have shown substantially weaker loadings when the covariances were freed in the exploratory model. Nonetheless, this sensitivity analysis provides only a rough approximation of bias because a different estimator was used to the main model (ESEM requires weighted least squares for complex models).

6.4.2 How should the General and Specific PD Factors Be Interpreted?

The bifactor model showed a multidimensional data structure, with the specific factors explaining more than half of the common variance. However, the variance in raw total scores was largely explained by the general PD factor.

Therefore, while the latent structure of PDs requires both general and specific factors to be adequately modelled, its measurement is mainly attributable to a single dimension like in bifactor studies of psychopathology (see Chapters 3, 4, and 5).

Consistent with past studies, borderline PD items loaded most strongly and preferentially onto the general PD factor compared to the specific borderline factor (Conway et al., 2016; Sharp et al., 2015; Williams et al., 2017; Wright et al., 2016). Items such as 'feels empty' and 'identity disturbance' reflect problems in self-functioning, i.e. stability and coherence in one's sense of identity. While not observed to the same extent in past studies, avoidant PD items such as 'preoccupied with rejection' preferentially loaded onto the general PD factor rather than the specific avoidant factor. Avoidant PD items reflect interpersonal dysfunction, i.e. problems in the ability to relate to and empathise with others. Taken together, this pattern of loadings supports the idea that general PD reflects dysfunction in self- and other-functioning, consistent with Criterion A of the DSM-5 Section III alternative model of personality disorders (Oldham, 2018).

While borderline PD items formed a specific borderline factor in the current study, others have shown that such items load to unity with the general PD factor (Sharp et al., 2015; Williams et al., 2017; Wright et al., 2016). Items load to unity when they account for little variance beyond the general factor (Gustafsson & Åberg-Bengttson, 2010). Nonetheless, specific factors can still be identified but lack reliability. Indeed, the borderline and avoidant factors explained (i) less than a third of the variance in raw borderline and avoidant subscale scores (e.g., omega - subscale hierarchical values), (ii) less than 5% of the common variance in the measurement model (e.g., explained common variance-subscale values), and (iii) were inadequately represented as latent variables by their indicators (e.g., H values

< .70). The specific borderline and avoidant factors might have been identified due to the confirmatory model's constraints; exploratory models with fewer constraints are associated with non-identified borderline factors (Sharp et al., 2015; Williams et al., 2017). It is important to note that the information explained by the specific borderline and avoidant factors was not lost per se but explained by the general PD factor.

By contrast, antisocial and narcissistic PD items loaded most strongly onto their respective specific factors, as has been reported by others (Sharp et al., 2015; Williams et al., 2017; Wright et al., 2016). The specific antisocial and narcissistic factors showed high reliability, changing least between the correlated factors and bifactor model (i.e. low mean parameter change); explaining most of the variance in raw narcissistic subscale scores (i.e. high omega hierarchical-subscale values); and being strongly represented by their indicators (i.e. H values). This follows the general trend in psychopathology research, whereby externalizing factors tend to show high reliability, stable specific factor loadings, and preferential loadings from antisocial items (Carragher et al., 2016; Caspi et al., 2014; Conway et al., 2019; Gomez et al., 2019; Haltigan et al. 2018; Hyland et al., 2018; Lahey et al., 2012; Lahey et al., 2015; Lahey et al., 2017; Martel et al., 2017; Olino et al., 2014). Unlike psychopathology studies, however, antisocial PD items still showed moderate general PD factor loadings, while narcissistic PD items showed weak general PD loadings. Therefore, it is not the case that all externalizing-type PD items are distinct from the general PD factor, but rather, some capture more reliable variance beyond the general variance than others.

Schizotypal PD items were split between the general PD factor and specific schizotypal factor in a pattern mirroring previous item-level analyses (Sharp et al., 2015; Williams et al., 2017). For example, items associated with ideas of reference, suspiciousness, and social anxiety loaded more strongly onto the general PD factor, whereas items associated with unusual perceptions, odd beliefs, and strange behaviours loaded more strongly onto the specific schizotypal factor. This pattern may be best understood as a divide between severity and style (Hopwood et al., 2011). Paranoid thinking and anxiety accompany a range of severe presentations (Caspi et al., 2014), while odd beliefs and behaviours are characteristic of personality traits that are not necessarily pathological (van Os & Reininghaus, 2016). The split between severity and style did not hinder the amount of common variance explained by the schizotypal PD factor (e.g., relatively high ECV_s values for an individual specific factor), or the extent that schizotypal PD items reliably represented the specific schizotypal construct (e.g., H values remained high), but it did reduce the proportion of variance in raw schizotypal subscale scores explained by the specific schizotypal factor (e.g., omega-hierarchical subscale was low).

In contrast to schizotypal PD items, obsessive-compulsive PD items were not predicted well by the general PD factor or the specific obsessive-compulsive factor. Furthermore, the specific obsessive-compulsive factor showed relatively weak reliability in terms of the proportion of variance explained in raw obsessive-compulsive subscale scores (e.g., low omega hierarchical-subscale value), and the extent that it was well represented by obsessive-compulsive items (e.g., low H value). Unlike the borderline and avoidant factors, whose poor reliability was attributable to the general factor, obsessive-compulsive factor loadings were minimally affected by the common variance, as denoted by the small mean parameter change. Therefore, obsessive-compulsive items performed poorly in estimating a reliable obsessive-compulsive PD factor.

6.4.3 How do the General and Specific PD Factors Predict Treatment Outcomes for Depression?

Higher general PD factor scores significantly predicted higher initial intercept values (e.g., depression scores at week two) but not variation in the linear or quadratic growth curves. In other words, individual differences in the severity of personality dysfunction predicted the overall severity of depression, which might be misinterpreted as an association between uncontrolled markers of PD (i.e. measures that conflate severity and style) and poorer depression outcomes. Importantly, PD severity did not in itself predict differential treatment responses; controlling for the general PD factor, like in the current study, or its sequalae, such as baseline depression severity, negates the association between uncontrolled measures of PD and poorer depression outcomes (De Bolle et al., 2010; Erkens et al., 2018; van Bronswijk et al., 2018). In sum, the mixed findings regarding the predictive value of PDs on depression outcomes might be largely explained by the extent that overall illness severity is controlled for (Mulder, 2002).

Higher specific borderline factor scores were associated with lower initial depression scores and flatter negative linear slopes. That is, once the effect of general PD severity and other stylistic tendencies was controlled for, borderline traits predicted slower treatment responses. This is particularly interesting given that in an overlapping dataset, a BPD diagnosis was associated with higher initial depression scores but not with differential treatment response rates (Fowler et al., 2018). If anything, patients with a BPD diagnosis showed better absolute outcomes, in that their depression scores dropped a larger amount to reach a similar end-point to those without a BPD diagnosis. The current study suggests that the higher severity of baseline depression scores associated with a BPD diagnosis was in fact a function of general PD severity. Only once the common variance in PD ratings was separated from the specific variance do we find that stylistic borderline traits are associated with poorer depression outcomes.

As alluded to above, the specific borderline factor might reflect personality tendencies such as a fragile (or malleable) identity and interpersonal sensitivity that interfere with trusting socially communicated information as a deferential source (Fonagy, Luyten, Allison, & Campbell, 2017a, 2017b). Mistrust towards the communicator's message and intentions would slow progress in treatment, be it directly (e.g., by disregarding therapeutic material or salubrious experiences in general) or indirectly (e.g., by not adhering to treatment recommendations such as medication or homework). One might argue that the specific borderline factor lacked reliability and therefore does not precisely reflect these tendencies. While this is true, the fact that it predicts slower treatment responses is evidence of its substantive nature.

Alternatively, the association between the specific borderline factor and slower treatment responses might be a by-product of controlling for the general PD factor, which lowered the initial depression scores and hence steepness of the slope. However, those with higher borderline factor scores had higher predicted depression scores at the final time-point compared to those with low borderline factor scores, suggesting that the flatter slopes were not purely a function of removing the baseline severity effect. Still, care should be taken not to over-interpret the association as it was rather weak and did not survive correction for demographic variables at the 5% level.

Higher specific antisocial factor scores were associated with lower initial depression scores and stronger quadratic (i.e. U-shaped) slopes. That is, once the effect of general PD severity and the stylistic tendencies was controlled for, antisocial traits predicted an initial decline followed by an upward inflection in depression scores. Few have documented the prognostic value of ASPD for depression outcomes, but an early prospective study reported higher depression recurrence rates associated with ASPD (and BPD) compared to bipolar disorder (Perry, 1988). More generally, ASPD is associated with high rates of recidivism (Bonta, Blais, & Wilson, 2014). The specific mechanisms that predict recurrence in offending and depression are unlikely to be the same, but the broader mechanisms associated with antisocial traits may contribute to both, such as disinhibition (Remster, 2014).

The prognostic value of the general and specific PD factors is most apparent when compared to the correlated factors. When the growth factors were regressed onto the correlated PD factors alone, most PDs predicted the intercept in ways that reflected their affinity to the common variance in PDs. For example, PDs that were most reflective of the common variance, such as borderline and avoidant PDs, predicted higher baseline depression scores, which was likely a function of the general PD factor's baseline severity effect (controlling for general PD resulted in negative predictions between the specific borderline and avoidant factors and baseline depression scores). By contrast, PDs that were least reflective of the common variance, such as antisocial and narcissistic PDs, maintained the negative predictions of baseline depression scores observed when general PD was controlled for in the bifactor model.

It is perhaps surprising that none of the PD factors in the correlated factors growth model without other covariates predicted variation in the linear slopes. Had we simply run a correlated factors model, which conflates the general and specific variance in PDs, we might have concluded that PDs have no prognostic effect on treatment responsiveness for depression like in some studies (Kool et al., 2005; Mulder, 2002). Controlling for baseline depression severity (at admittance) and other clinical covariates removed the effect of PDs on initial depression scores (at week 2) that was likely driven by a baseline severity effect.

Nonetheless, baseline depression severity did not capture all the variance in overall baseline severity. Hence, the uncontrolled common variance in PDs associated with personality disorder severity likely drove the positive and negative associations between schizotypal PD and the intercept and linear slope growth factors, respectively. That is, higher schizotypal PD scores in the correlated factors growth model likely predicted higher initial depression scores due to the baseline severity effect of general PD. Moreover, the faster decline in depression scores predicted by higher schizotypal PD scores was likely an effect of regression to the mean[22] underpinned by general PD.

It is interesting that higher borderline factor scores in the correlated factors growth model were associated with slower rates of decline in depression scores, albeit weakly and not significantly. Only when patients' overall illness severity was accounted for, by estimating the common variance in PDs and controlling for baseline severity in depressions scores, did we see the adverse effect of borderline

[22] Regression to the mean describes the empirical phenomenon that extreme scores at an initial measurement occasion normalize upon repeated measurement (Wise, 2004). In the current study, depression scores at admittance and the general PD factor both predicted a regression to the mean in depression scores, which started higher but ended lower.

PD on depression outcomes. It is also interesting that the antisocial factor predicted a U-shaped pattern of change in both the bifactor and correlated factor growth models, probably because the antisocial factor was largely independent of the common variance. Admittedly, the association between borderline PD factor scores and stronger inverted U-shaped change in the correlated factors growth model is not entirely clear, but it might be a conflated effect of the specific borderline variance, which predicted flatter slopes, and the common PD variance, which predicted steeper slopes.

6.4.4 Strengths and Limitations

A limitation of the current study is that most patients disagreed to experiencing PD symptoms. Therefore, item response distributions had a restricted range that could hinder the integrity of the factors estimated. For example, simulation studies have shown that the robust maximum likelihood estimator (which was used in the current study) introduces more bias compared to the robust weighted least squares estimator when analysing binary indicators (Beauducel & Herzberg, 2006). Yet, weighted least squares is not typically applied to continuous or mixed continuous and categorical indicators like in the current growth model; there are anecdotal reports of bizarre polyserial correlations between continuous and categorical indicators using weighted least squares (Rigdon, 2015). Robust maximum likelihood was chosen over robust weighted least squares because it is perhaps better to estimate binary indicators with an estimator that performs slightly worse but still adjusts the standard errors for bias, rather than estimate continuous outcomes with polyserial correlations without any adjustment for bias. A sensitivity analysis demonstrated that robust maximum likelihood and weighted least squares estimators produced similar factor loadings; growth factors could not be estimated with weighted least squares, demonstrating its instability.

The restricted item response distributions also risk a substantive interpretation of the PD factors. Items with heavily skewed response distributions can form ‘difficulty’ factors that reflect similarities in the likelihood of endorsing items rather than variation in an underlying trait (Guilford, 1941, but see McDonald & Ahlawat, 1974 for a refined explanation). Weighted least squares and robust maximum likelihood estimators are designed to minimize artifactual factors resulting from item distribution similarity (Wirth & Edwards, 2007), but they cannot change the underlying distributions. Therefore, a factor might be driven by a limited part of the sample.

Take for instance the specific antisocial factor. Only 2% of the sample endorsed antisocial PD items on average, most likely because of their ‘difficulty’ (e.g., few patients admitted for depression would be expected to commit [or disclose] antisocial behaviours, particularly criminal behaviour, compared to patients in a forensic setting). However, the specific antisocial PD factor is unlikely to be an artifact of item difficulty because it predicted depression outcomes. Still, a subset of the sample likely drove this prediction, contradicting the assumption that antisocial PD is well represented by a continuous latent trait; a categorical latent variable might be more appropriate in the current sample.

Another issue is that both PDs and depressive symptoms were reported by patients. Their relationship might thus be partly explained by response biases affecting both measures. Patients diagnosed with a PD often rate their depression as more severe than do clinicians (Unger, Hoffmann, Köhler, Mackert, & Fydrich, 2013). Hence, the slower rate of decline associated with borderline traits and the U-shaped pattern of change associated with antisocial traits might be a function of stylistic patterns in *reporting* rather than behaving, such as catastrophising one's symptoms or emphasising their severity as the fear of discontinuing treatment grows. Nonetheless, the two are unlikely to be distinct: negative response styles may in themselves reflect behavioural tendencies that confer risk to psychopathology (Lahey et al., 2012). Moreover, there is some evidence that clinician ratings of PDs capture general severity, while patient ratings reflect both severity and style (Woods, Edershile, Wright, & Lenzenweger, 2019).

6.4.5 Implications and Future Directions

The current findings demonstrate the value of assessing overall levels of personality impairment as well as unique styles of symptom expression (Hopwood et al., 2011). The severity-style framework is at heart of the DSM-5 Section III alternative model of PDs, which features a criterion for the overall level of personality impairment in 'self' and 'other' domains in addition to specific PD diagnoses (Oldham, 2018; Skodol et al., 2011). The current findings add to the growing evidence favouring the assessment of both general and specific aspects of personality impairment (Conway et al., 2016; Hengartner et al., 2014; Hopwood et al., 2011; Jahng et al., 2011; Morey & Benson, 2016; Morey et al., 2015; Morey, Benson, & Skodol, 2016; Morey et al., 2012; Morey, Skodol, & Oldham, 2014; Sharp et al., 2015; Skodol et al., 2012; Waugh et al., 2017; Williams et al., 2017; Wright et al., 2016), the previous lack of which encouraged the APA's board of trustees to retain the standard categorical model (Oldham, 2015).

Future studies could also incorporate intermediate levels of PD characteristics in their models, such as the maladaptive personality traits specified in Criterion B of the DSM-5 alternative model (e.g., Strickland et al., 2019). It would be interesting to compare the predictive strength and direction of each level of PD on clinical outcomes to inform novel approaches to triaging patients (e.g., general and specific PD characteristics might inform the intensity and type of intervention, respectively; Bach & First, 2018; Hopwood, 2018).

The current findings also favour the inclusion of a dimensional BPD qualifier, which is the topic of much debate, as it is uncertain whether BPD reflects a general or specific PD impairment (Reed, 2018). The specific borderline factor predicted poorer depression outcomes beyond the general PD factor but was poorly measured. Future studies should determine the characteristics that define specific borderline traits to improve their measurement (e.g., Fowler et al., 2018), bearing in mind that they might interact with the specific nature of depression presented by patients (Rogers, Widiger, & Krupp, 1995; Westen et al., 1992).

The current findings also highlight the importance of studying the unique contributions of general and specific PD components to depression outcomes. If these components are not separated out, their conflicting relationships might obscure treatment predictions. The bifactor model achieves this separation but requires sufficiently large sample sizes to estimate the increased number of parameters introduced by the general factor. Other methods that are more suited to clinical practice include carefully conducted clinical interviews using the Structured Clinical Interview for the DSM-5 Alternative Model of PDs (First, Skodol, Bender, & Oldham, 2017), or ipsatizing trait-domain scores measured using the Personality Inventory for DSM-5 (Krueger, Derringer, Markon, Watson, & Skodol, 2012) for general severity scores assessed using the Level of Personality Functioning Scale (Morey, 2017; see Hopwood, 2018, and Skodol, Morey, Bender, & Oldham, 2015, for case illustrations).

Finally, the current research question–if PDs predict differential responses to treatment for depression–is born from a dated tradition of dividing clinical disorders (axis I) and personality disorders (axis II). It was assumed that PDs are a primary feature of the clinical profile that shapes the course of depression (Tyrer, 2015) and there is some evidence supporting this. For instance, PDs in adolescence significantly increase the risk of depression in adulthood (Johnson et al., 1999), and improvements in PD precede improvements in depression, but not the reverse (Gunderson et al., 2004). Nonetheless, the presence of both a PD and depression in adolescence often outweighs the predictive strength of either one alone (Crawford et al., 2008; Kasen, Cohen, Skodol, Johnson, & Brook, 1999). Therefore, the relationship between PDs and depression may not be a simple, unidirectional one (Livesley, 2015).

While it was assumed that borderline and antisocial traits predicted differential responses to treatment because they themselves did not change, there may be bidirectional interactions between PDs and depression (Widiger, 2011). Future studies should take repeated measures of both depression and PDs–which may still reflect state and trait aspects of the clinical presentation, respectively–to determine the developmental processes by which PDs influence changes in depression and vice versa.

Chapter 7 General Discussion

Contrary to popular depictions of science as a series of breakthroughs, science progresses through a slow and steady accumulation of research findings (Foster, Rzhetsky, & Evans, 2015). However, scientific productivity often booms following a breakthrough. Consider the first direct evidence of Einstein's gravitational waves reported by Abbott and colleagues in 2016 that has already accrued over 6,000 citations and is set to pave a new era of gravitational wave physics. One could say that the bifactor model provides evidence for psychiatry's gravitational waves; a single dimension of mental health has been intuited for at least a century, but it is only since the "discovery" of the *p* factor that clinical scientists have a direct measure that is subject to scientific inquiry. Consequently, studies that analyse psychiatric data with a bifactor model have boomed since the seminal work of Lahey et al. (2012) and Caspi et al. (2014). The bifactor model has received positive attention from social and clinical scientists alike, but also negative attention from quantitative methodologists. The current thesis aimed to investigate 'both sides of the coin', with studies testing the methodological issues and clinical utility of the bifactor model.

I will begin this final chapter by summarizing the main findings from the preceding chapters. I will then evaluate these chapters based on recent criticisms of the bifactor model and draw inferences about the field more broadly. I will end by discussing the implications and future directions of the findings presented with an emphasis on clinical research and practice.

7.1 Summary of Findings

This thesis was structured into two parts, each with two empirical chapters. The first part focused on methodological issues associated with the bifactor model. Chapter 3 aggregated reliability estimates for bifactor studies of psychopathology published to date, including the explained common variance (i.e. the degree of multidimensionality in the factor solution) and omega hierarchical (i.e. the internal consistency among indicators predicted by a given factor).

On average, the common variance (i.e. variance in the indicators explainable by all factors modelled) was split between the *p* factor and specific psychopathology factors, favouring a multidimensional model. However, the amount of variance explained by the *p* factor was dependent on the study characteristics, particularly the informant (e.g., parent and teacher reports overestimated and underestimated the *p* factor strength, respectively, compared to self-reported problems). By contrast, the internal consistency of total and subscale scores was largely explained by the *p* factor.

In Chapter 4, the second methodological study, I examined the contribution of response biases to the *p* factor and specific psychopathology factors. Response biases are consistencies in responding on self-report measures that are unrelated to the construct assessed; the positive covariation among all symptoms could be a product of response biases rather than a substantive latent trait. The tendency to indiscriminately agree or disagree with a set of heterogeneous questions explained just 4% of the variance in the *p* factor, and even less in the specific psychopathology factors. Therefore, it is unlikely that response biases account for the systematic variance in the psychopathology factors, at least when measured as general preferences for certain response options.

The second part of this thesis explored the benefits of analysing clinical outcomes data with the bifactor model. In Chapter 5, longitudinal changes in the *p* factor and specific psychopathology factors were assessed over a psychosocial intervention for antisocial adolescents. An initial analysis showed widespread declines across problem areas when analysed as independent subscales (e.g., antisocial, attention, anxiety, and mood subscales). However, a different picture emerged when changes specific to each problem area were analysed whilst controlling for changes common to all problem areas, as captured by the specific factors and *p* factor, respectively.

As expected, the *p* factor declined over the intervention and follow-up period, reflecting the widespread changes across all problem areas observed in the initial analysis. However, the only specific factor that continued to decline over the study period was the antisocial factor, presumably because it captured the specific effect of the interventions on conduct problems. Surprisingly, the specific anxiety factor increased over the study period, which might reflect a facilitative effect of treatment. Lastly, the specific mood and attention factors showed little change over the study period, suggesting that their decline in the initial analysis was a function of the *p* factor.

Chapter 6 investigated whether changes in depression over an inpatient intervention could be predicted by the general and specific aspects of personality disorders from a bifactor model. Findings have been mixed as to whether personality disorders predict differential responses to treatment, but this might be because personality disorder measures conflate the overall severity of personality dysfunction (which might predict the overall severity of depression) and disorder-specific profiles (which might predict differential treatment responses). Supporting this hypothesis, the general personality disorder factor predicted higher baseline depression scores but not differential rates of change in depression scores, while the specific borderline and antisocial factors predicted slower or U-shaped declines in depression scores, respectively. By contrast, borderline personality disorder no longer predicted differential rates of change when personality disorders were analysed as distinct but correlated entities using a corelated factors model.

7.2 Study Evaluation

Bifactor models of psychopathology have received a new wave of criticism since starting this thesis (Greene et al., 2019; Sellbom & Tellegen, 2019; Watts, Poore, & Waldman, 2019). In this section, I evaluate the current chapters against these criticisms and highlight the implications for the field more broadly. While I have split these criticisms into three sections, they are not mutually exclusive.

7.2.1 "Bifactor Models Cannot Be Selected Simply Because They Fit the Data Better"

Most bifactor studies of psychopathology have relied almost exclusively on model fit indices when choosing the model that represents their data best. Yet several authors have shown that the bifactor model's ability to fit the data better than competing models is partly due to its greater complexity (Bonifay & Cai, 2017; Gignac, 2016) and overfitting tendencies (Greene et al., 2019; Murray & Johnson, 2013; Reise, Kim, Mansolf, & Widaman, 2016). This has led some researchers to conclude that “its current popularity notwithstanding, we believe these applications of the bifactor model need to be challenged and are actually troublesome… a bifactor model cannot be selected simply because it was found to fit the data better.” (Sellbom & Tellegen, 2019, p. 9-10).

In all chapters, a standard or revised bifactor model fit better than a correlated factors model and single factor model. Therefore, it could be argued that support for the bifactor model shown throughout this thesis reflects its higher fitting propensity, “which is a statistical feature, rather than a substantive argument for utilizing a bifactor model.” (Greene et al., 2019, p. 17). However, in addition to validating the bifactor dimensions against external criteria (see below), the current chapters included alternative tests of model integrity to avoid an over-reliance on fit indices.

As a minimum, information criteria were used to compare models throughout the thesis. Information criteria penalize for model complexity based on the number of freely estimated parameters (Morgan, Hodge, Wells, & Watkins, 2015). Therefore, the bifactor model’s superiority in each chapter should not be due to its over-parametrization. However, information criteria do not penalize for a model’s functional complexity; bifactor models fit better than competing models even when the number of parameters is held constant, indicating that the *way* that bifactor models specify their parameters is also important (Bonifay & Cai, 2017). Information criteria still show a pro-bifactor bias when fitted to data generated from a correlated factors model, particularly when the population model features misspecifications (e.g., correlated residuals or cross-loadings; Greene et al., 2019; Murray & Johnson, 2013).

Greene et al. (2019) suggested that information criteria be formally compared with the Vuong test (1986) to provide greater certainty in which model better approximates the true data generating model. In each chapter, the Vuong test generally aligned with the comparisons of information criteria (e.g., differences >|10| imply substantial differences between models). However, in Chapter 5, the Vuong test and comparison of information criteria diverged, e.g., the difference in information criteria between a traditional bifactor model (specific factors without cross-loadings) and correlated factors model was > 10 but the Vuong statistic was not significant. Information criteria are variable across different sample sizes (Preacher & Merkle, 2012), which is problematic when comparing models using set cut-offs. Therefore, statistics like the Vuong test can be helpful in quantifying the difference between models parametrically (Sayyareh, Obeidi, & Bar-Hen, 2010). It should be noted, however, that the Vuong test statistic was rather large in all chapters (e.g., *z* scores > 40), implying it is over-sensitive to model differences.

Models were also cross-validated by estimating each solution in half the sample and evaluating the model parameters in the remaining half. If the bifactor model is most sensitive to sample-specific noise (Greene et al., 2019), then it should have shown the poorest cross-validation. However, cross-validation was poor for all models in each chapter, suggesting that the influence of noise on model fit might not be unique to the bifactor model (it might simply be that the bifactor model handles noise better). Nonetheless, the 'split-half' approach to cross-validation has its weaknesses; the cut in sample size threatens the stability of estimates. Ideally, model parameters would have been cross-validated in independent samples, but most of the current chapters were secondary analyses of data that had already been collected.

Finally, bifactor models in each chapter were evaluated with model-based reliability indices. Rather than testing which model shows a better fit, reliability indices summarize the properties of a bifactor model, such as how multidimensional the variance explained is (e.g., explained common variance), or what proportion of the inter-relatedness between items is explained by a given factor (e.g., omega hierarchical; Rodriguez et al., 2016b). Reliability indices provided a more rounded analysis of the bifactor model and followed general trends, e.g., the common variance was equally split between the *p* factor and specific psychopathology factors and hence was multidimensional, while the inter-relatedness among all items or subgroups of items was mainly attributable to the *p* factor.

Model-based reliability indices offer a promising approach to evaluating bifactor models compared to fit indices (Sellbom & Tellegen, 2019), but they too have drawbacks. For example, much of the work examining the properties of model-based reliability indices has used simulation methods (Reise, Bonifay, & Haviland, 2013; Reise, Scheines, Widaman, & Haviland, 2013; Zinbarg, Yovel, Revelle, & McDonald, 2006). Certain predictions made in simulation studies are not upheld in empirical studies; for example, Reise, Scheines, Widaman, and Haviland (2013) found that the explained common variance was invariant to the number of test items and percentage of uncontaminated correlations (i.e. the number of correlations attributable to a general factor free from the influence of specific factors), but both of these characteristics predicted variability in the explained common variance between bifactor studies in Chapter 3.

Another limitation of model-based reliability indices is that they are dependent on the quality of the model estimated; if the model is mis-specified, then so will the reliability indices. One could argue that the reliability indices reported in the current chapters should accurately represent the distribution of variance, since the bifactor models fit well. However, fit indices are biased estimators of bifactor model fit (see above) and should probably not be relied upon for this argument. It will be important for future work to determine methods for testing the accuracy of reliability indices (e.g., developing confidence intervals), which take into account their sensitivity to a study's methodological characteristics (see Chapter 3).

Overall, the methodological quality of the chapters presented is ultimately subject to the tests used to evaluate them. Many tests are in their infancy or require special care when applied to datasets with polytomous items (e.g., model-based reliability indices; Reise, Bonifay, & Haviland, 2013). The question of whether bifactor models truly provide a better fit to competing models ultimately rests on the advancement of model comparison methods that incorporate the number of estimated parameters and a model's functional complexity, such as minimum description length approaches[23] (Markon & Jonas, 2016).

7.2.2 "Bifactor Models Do Not Reflect the Latent Structure of Psychopathology"

Throughout this thesis, the bifactor model has been compared to the correlated factors model and single factor model under the assumption that each represent a different underlying structure of psychopathology. However, comparing models based on statistical fit reflects their psychometric properties rather than their underlying structure. In fact, the mechanisms that produce

[23] Minimum description length approaches were not used to evaluate models in this thesis as they require advanced computations that are not currently available in common structural equation modelling packages.

correlations among item responses might be entirely different to latent traits (van Bork et al., 2017; van der Maas et al., 2006). Therefore, some researchers have concluded that "Bifactor models cannot be used, however, to specify an optimal internal structure unless also theoretically strongly justified" (Sellbom & Tellegen, 2019, p. 11; see also Bonifay et al., 2017, and Greene et al., 2019).

Investigating the underlying structure of psychopathology is no simple task but models can be compared against theoretically relevant external criteria (Greene et al., 2019; Sellbom & Tellegen, 2019). Each chapter in this thesis included external predictors or outcomes that were associated with the bifactor dimensions in theoretically relevant ways. For example, in Chapter 4, the *p* factor and specific psychopathology factors were weakly predicted by response biases, supporting their substantive validity. In Chapter 5, reductions in the *p* factor were associated with declines in criminal offences and school exclusions, supporting hypotheses of *p* as an index of general impairment (Caspi & Moffitt, 2018). Furthermore, higher levels of specific anxiety predicted fewer school exclusions, supporting hypotheses of internalizing-related factors as personality traits associated with obedience and inhibition (Lahey et al., 2015). Finally, in Chapter 6, the general and specific personality disorder (PD) factors predicted longitudinal changes in depression outcomes in a manner consistent with the hypothesis that general PD reflects the overall severity of personality dysfunction, whereas specific PDs reflect styles of maladjustment (Sharp et al., 2015; Wright et al., 2016).

Validating the bifactor dimensions against external criteria demonstrates their substantive basis, but it still might not be sufficient to demonstrate the latent structure of psychopathology (Bonifay et al., 2017; Watts et al., 2019). For instance, cross-sectional associations between the *p* factor and external variables suggest that *p* is just as much a product of illness severity than it is an underlying vulnerability factor. Moreover, conceptualizing the *p* factor as a broad vulnerability factor risks its ability to be falsified, since it will correlate with any and all forms of risk (Greene et al., 2019).

Bonifay et al. (2017) and Watts et al. (2019) proposed some 'riskier' tests of the bifactor model that place it under stronger theoretical scrutiny and hence provide better approximations of the underlying structure of psychopathology. For example, Bonifay et al. advised that structural models of psychopathology be validated against the hypothesised psychobiological structure, and that changes in the psychobiological structure should cause changes in the latent variables. None of the empirical chapters included psychobiological measures, but Chapter 5 examined changes in the bifactor dimensions that were preceded by, and hence potentially caused by, a psychosocial intervention. Therefore, Chapter 5 provides indirect evidence of causal shifts in the structure of psychopathology via changes in the environment that are inherently mediated by the psychobiological structure (Roiser, 2015; see also Wade, Fox, Zeanah, & Nelson, 2018).

In another 'risky' test, Watts et al. (2019) proposed that the bifactor model should be directly compared with competing models for differences in their relationship with external criteria. Most bifactor studies that include external criteria assume that the bifactor model improves on the external validity of the correlated factors model without explicitly testing this. The two chapters that assessed the predictive validity of the bifactor dimensions over a psychosocial intervention (Chapter 5) or for depression outcomes (Chapter 6) included a direct comparison with the correlated factors model. In Chapter 5, disorder-specific factors in the bifactor model showed more nuanced changes over a psychosocial intervention (e.g., factors declined, increased, or remained constant) compared to the correlated factors model (e.g., all factors declined). In Chapter 6, borderline personality disorder in the bifactor model significantly predicted poorer depression outcomes but not in the correlated factors model. Therefore, in both correlated factor models, theoretically and clinically important findings were masked by the variance common to all problems.

Even though the bifactor and correlated factor models diverged in their external predictions, Watts et al. (2019) would argue that the bifactor model should have also explained more variance in the external criteria compared to the correlated factors model for it to have exceeded its external validity. However, the bifactor and correlated factor models explained the same amount of variance in Chapter 6.[24] Watts et al. reasoned that "should bifactor and correlated factors models of psychopathology explain an equivalent amount of the variance in their constituent psychopathology indicators [or external criteria], it would indicate that bifactor models merely redistribute aspects of psychopathology into a greater or different number of factors." (p. 4). Nevertheless, redistributing the variance into common and specific components is not only the goal of bifactor models, it is also their greatest asset (Gustafsson & Åberg-Bengttson, 2010; Reise, 2012). Consider the findings from Chapters 5 and 6: had the variance common to all disorders remained mixed with the variance associated with specific disorders, then theoretically and clinically relevant differences would have been missed.

[24] R^2, or the proportion of variance explained in an outcome variable by its predictors, was not computable in Chapter 5 due to the statistical uncertainty in estimating R^2 in multilevel models with random effects (Nakagawa & Schielzeth, 2013).

In sum, the latent structure of psychopathology cannot be evaluated by model fit indices alone, or even by correlating factors with external criteria. Carefully conducted studies are needed that compare diverging predictions of the bifactor and correlated factors models grounded in theory. The current chapters go some way in achieving this, but further work is needed that advances the theoretical basis of these models (see Carver, Johnson, & Timpano, 2017; Del Giudice, 2014; Fonagy, Luyten, Allison, & Campbell, 2017a) and explicitly tests their predictions in prospective longitudinal designs.

7.2.3 "Bifactor Models of Psychopathology Are Difficult to Interpret"

Perhaps the most common criticism of bifactor models of psychopathology is that they are difficult to interpret (Bonifay et al., 2017; Greene et al., 2019; Sellbom & Tellegen, 2019; Watts et al., 2019). I have already mentioned how a latent vulnerability interpretation of the *p* factor might not reflect the true data-generating mechanisms and also risks falsification (see section 7.2.2). Specific factors also come under interpretational fire because of their assumption of orthogonality: what does it mean for internalizing and externalizing problems to be orthogonal to, or removed from, general psychopathology (Bonifay et al., 2017)? Furthermore, in studies that free cross-loadings or correlations between specific factors, what does it mean for there to be shared variance that is not explained by the *p* factor (Markon, 2019)? As a result, some authors have concluded that "a bifactor model of psychopathology is difficult to interpret" (Watts et al., 2019, p. 14) and have endorsed alternative models to avoid "unclear latent meanings of descriptors (items) and unclear descriptive meanings of latent variables." (Sellbom & Tellegen, 2019, p. 11).

While the bifactor model's constraints might appear removed and unjustified, they are in fact grounded in a binomial taxonomy of psychopathology. For example, a latent bifactor structure distinguishes between the severity and style of people's problems and weighs them equally (Caspi et al., 2014; Greene et al., 2019). Indeed, the common variance was generally split between the *p* factor and specific factors across bifactor models published to date and in each empirical chapter, supporting this assumption.

Nonetheless, the review of bifactor studies published to date also demonstrated that the degree of multidimensionality changed with different study characteristics (see Chapter 3). Therefore, it might be that the design of the current chapters favoured a multidimensional solution, but they might have favoured a more unidimensional solution with the *p* factor outweighing the specific factors under different conditions. Moreover, the *p* factor in each chapter[25] was weighted towards certain problems over others (e.g., internalizing vs. externalizing), despite interpretations of *p* as a liability towards 'any and all symptoms' (Watts et al., 2019).

The main challenge to interpreting the specific factors in the current chapters was that they tended to suffer a loss in loading strength compared to the correlated factors model (see also Watts et al., 2019). This questions whether the specific factors were necessary beyond improving the fit of the single factor model. In other words, do specific factors represent theoretically meaningful components of psychopathology? Validating the specific factors against external criteria in each chapter provided some support for their substantive value, but predictions were

[25] I refer to the general factor in Chapter 6 as a *p* factor for ease but acknowledge that it is a general personality disorder factor that is not synonymous, but overlaps, with the general psychopathology *p* factor (Widiger & Oltmanns, 2017).

often weak (which is also observed in prior bifactor studies; e.g., Jones et al., 2019; Lahey et al., 2015; Patalay et al., 2015; Sallis et al., 2019). Therefore, the specific factors might have been theoretically meaningful but psychometrically unreliable.

The contrast between the specific factors' validity and reliability raises an important question: did the specific factors represent the same underlying construct across each chapter and even compared to prior studies (Greene et al., 2019)? The H index provides an estimate of a factor's reliability given its indicators; how likely is it that we would replicate the underlying construct in another study using the same indicators? Aside from Chapter 4, where the H index was high for each specific factor (H > .70; Hancock & Mueller, 2001), H values were generally subpar for specific factors in Chapters 5 and 6 (surprisingly, the average H value across bifactor studies to date was .69; see Chapter 3). These findings indicate that the constructs assessed by each specific factor were not reliably represented by their indicators and likely differed to some degree across chapters.

There were, however, consistencies across chapters in the items whose specific factor loadings dropped in favour of their *p* factor loadings. For example, in Chapters 4 and 5, items associated with depression and anxiety loaded most strongly and sometimes exclusively onto the *p* factor rather than the specific internalizing factor, while antisocial items such as rule-breaking and intrusiveness (Chapter 4) and fighting and stealing (Chapter 5) loaded preferentially onto the specific externalizing-type factors. Chapter 6 included personality disorder (PD) items rather than clinical 'axis I' items, but a similar pattern of change was observed. For instance, borderline PD items loaded most strongly onto the general PD factor; depression, anxiety, and aspects of borderline PD represent the internalizing dimension (Kotov et al., 2017). Moreover, narcissistic and antisocial items loaded most weakly onto the specific PD factors that represent the externalizing dimension (Kotov et al., 2017). These findings add to the argument that specific factors are meaningful but limited in their psychometric properties.

In all, the interpretability of specific factors in the bifactor model is limited by their reduced loading strength. However, this does not rule out their meaningfulness as representing specific styles of coping. It might be that current assessment measures are not well designed to capture specific domains beyond the general factor (Gignac, 2016).

7.3 Implications and Future Directions

I now turn to implications of the current findings for existing and future research on the quantitative classification of psychopathology. Given the breadth of this topic and emphasis on methodological issues in the prior section, I will focus on clinical implications for psychiatric nosology, clinical assessment, and treatment.

7.3.1 Implications for Psychiatric Nosology

Classifying natural entities into discrete groups is an efficient means of making sense of the world, but it also has its costs. We create borders between countries and group people into different races, but neither reflect the common landscapes and overlapping gene pools that tie them together (Jorde & Wooding, 2004). Similarly, mental disorders share overarching problems in emotions and relationships that are overseen when treated as discrete pathological entities. Statistical issues aside, the current chapters demonstrate the richness of classifying psychiatric symptoms into general and specific components.

Of course, understanding what makes things different is also important; just like how neighbouring countries often differ in cultural and religious traditions, mental disorders differ in their qualities and characteristics. Natural entities are composed of many layers that need to be decomposed in order to be fully understood (Simon, 1973). This thesis supports the growing movement towards a hierarchical analysis of psychopathology that is mainstream in other disciplines of science (most obvious is the hierarchical taxonomy of organisms in biology, but also consider hierarchical theorems of space and time; Wu, 2013).

Perhaps the most comprehensive and ambitious nosology of mental disorders to date is the Hierarchical Taxonomy of Psychopathology (HiTOP; Kotov al., 2017). HiTOP organizes mental disorders into a hierarchy of several, inter-related dimensions (see Figure 7.1). Dimensions at higher levels of the hierarchy explain the associations among lower level dimensions. The further down the hierarchy one moves, the narrower the phenomenon explained by each dimension. At the top of the hierarchy is a 'super spectra' level that includes broad factors like the *p* factor which explain the co-occurrences among all spectral level dimensions at the level below, including internalizing, externalizing (both disinhibited and antagonistic forms), thought disorder, detachment, and somatoform. In turn, spectral level factors explain the co-occurrences among subfactors (e.g., internalizing summarizes the covariation among fear, distress, eating, and sexual problems), and the cascading process continues until the lowest level (e.g., individual symptoms). The HiTOP dimensions are not meant to be definitive but provide an initial framework for further research to build on.

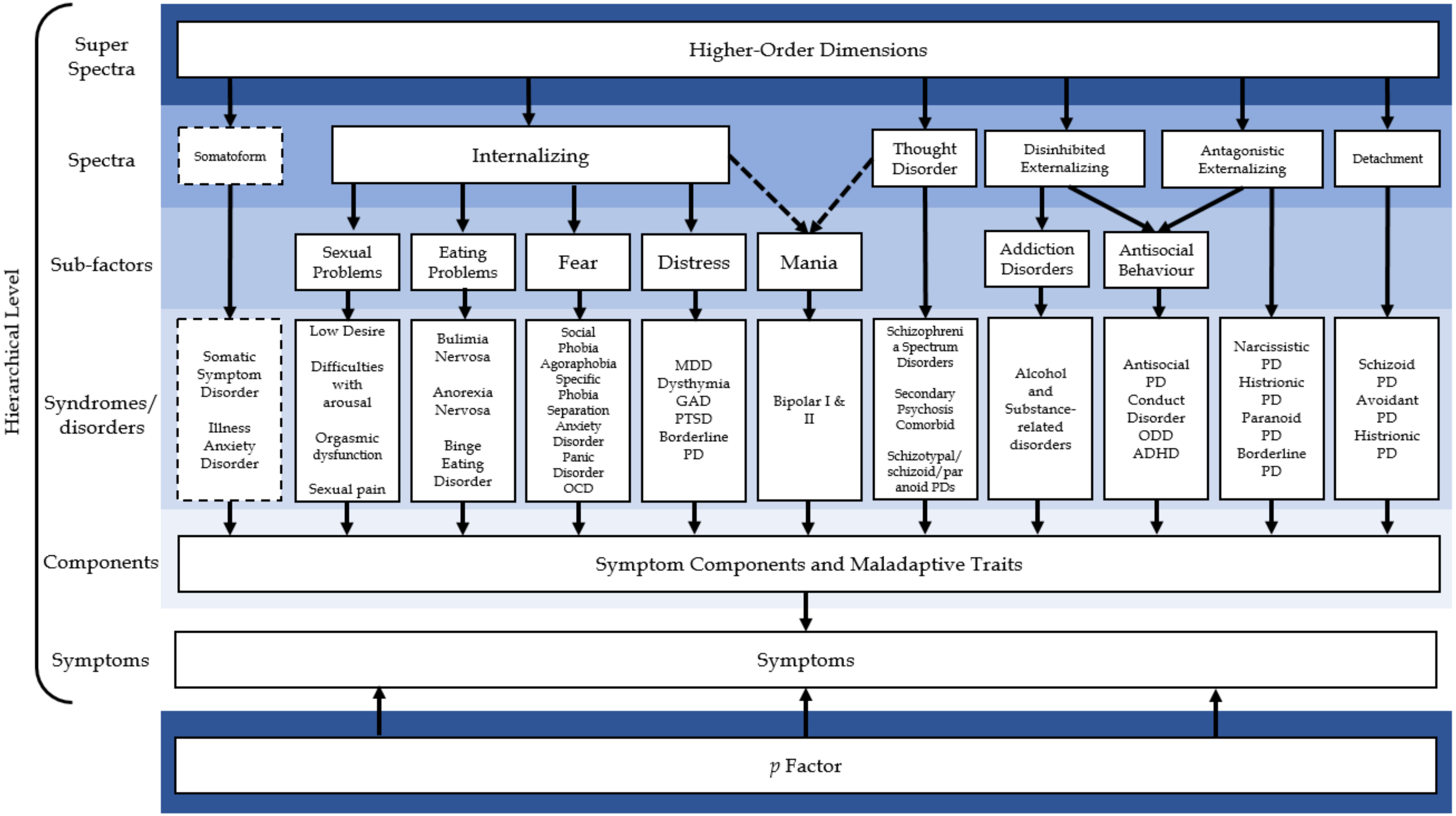

Figure 7.1. Schematic of the Hierarchical Taxonomy of Psychopathology (HiTOP) reproduced from Kotov et al. (2017). The *p* factor has been added below the symptom level to demonstrate its heterarchical relationship with specific dimensions.

The current chapters support the HiTOP nosology in several ways. For example, substantive evidence for a *p* factor in each chapter upholds the importance of super spectral factors in the hierarchy. The inclusion of super spectral factors is the topic of much debate: On the one hand, spectral level factors are positively correlated, implying the influence of a broader dimension (Lahey et al., 2012). On the other, spectral level correlations are not uniform, as would be expected if they were underpinned by a single dimension (Krueger et al., 2018). The only study to model spectral factors was Chapter 4 (Chapters 5 and 6 modelled disorder-level factors), which showed strong and uniform positive correlations. However, certain characteristics of the Achenbach Self-Report, such as its relatively higher number of items and percentage of uncontaminated correlations, might have inflated the common variance (see Chapter 3). Therefore, further studies are needed to determine the extent that the associations among spectral level factors are driven by methodological characteristics that would question the validity of an overarching super spectral factor (though narrower super spectral factors still might be relevant).

The current chapters also align with the spectra and subfactors identified in the HiTOP nosology. For example, in Chapters 5 and 6, disorder-specific factors from the internalizing domain showed stronger within-domain correlations (e.g., correlations between internalizing disorders) than between-domain correlations (e.g., correlations between internalizing and externalizing/cognitive disorders), implying the influence of a spectral level internalizing factor. Furthermore, in Chapter 4, internalizing items were split between the *p* factor and specific internalizing factor in a way that mirrored the bifurcation of internalizing problems into distress and fear subfactors, respectively (assuming that specific internalizing represented the fear subfactor, since the social withdrawal symptoms that defined it overlap with social phobia, agoraphobia, and separation anxiety disorder).[26]

The current chapters also support HiTOP's higher-level externalizing factors: Chapter 4 featured an externalizing factor and Chapters 5 and 6 showed stronger within-domain positive correlations between externalizing-type disorders than between-domain correlations. Externalizing-type specific factors also showed the greatest reliability beyond the *p* factor, consistent with the notion of a super spectral externalizing factor (Krueger & Markon, 2014). Furthermore, externalizing items followed a loading pattern that is partially consistent with the bifurcation of a super spectral externalizing dimension into antagonistic and disinhibited spectra (Kotov et al., 2017). For instance, disinhibited items associated with fighting and stealing loaded preferentially onto the externalizing-type factors, while antagonistic items such as anger and disobedience loaded preferentially onto the *p* factor. Nonetheless, it is uncertain why the *p* factor accounted for the antagonistic items and not a spectral level externalizing factor. Furthermore, Chapter 6 did not follow this pattern, with narcissistic PD items (i.e. antagonism; Kotov et al., 2017) loading preferentially onto the specific narcissistic factor rather than the general PD factor.

As hinted above, the current chapters also differ from the HiTOP model in important ways. For instance, the HiTOP model represents the shared and unique aspects of each dimension across multiple levels of a hierarchy. Aspects common to all problems are represented at the top of the hierarchy, while more distinct aspects are explained by dimensions at the lower levels. The current chapters support a

[26] The heavy weighting of anxious-depressed items on the *p* factor has led some to argue that *p* might simply reflect a broader distress factor (Kim & Eaton, 2015), which is another argument against the inclusion of a single super spectral factor that influences all problem domains, but perhaps not narrower super spectral factors.

bifactor model, which represents the shared and unique aspects of each *specific problem* at a single level of analysis. In other words, there is something shared across all problems that is distinct from their unique features. This difference is visualized in Figure 7.1, where the super spectral general factor in the HiTOP model sits at the top of the hierarchy, while the general factor in the bifactor model sits underneath the symptom level and hence beside the spectral level dimensions.

Some argue that the HiTOP model (which is based on a higher-order model) and bifactor model imply different realities and hence nosologies, which is particularly problematic given their statistical overlap (van Bork et al., 2017). However, the difference between models might not be as large as purported. For example, Kim and Eaton (2015) found that the *p* factor from the bifactor model was almost perfectly correlated with general factor from a hierarchical model[27] (r = .99). Furthermore, the specific factors in each model were strongly correlated, despite being parametized differently. Indeed, specific factors in the current chapters generally align with the HiTOP model (see above). Therefore, the bifactor and higher-order models might be more similar than different; the languages might differ but the meaning behind the phrase is the same. The advantage of 'speaking in bifactor' is that we can directly estimate the variance unique to the factors sitting below (or beside) the super spectral level (Gignac, 2008). Some spectral and sub-spectral level factors in the hierarchical/higher-order model often overlap strongly with, and hence explain little beyond, the general factor (Gustafsson & Åberg-Bengttson, 2010; Kim & Eaton, 2015; see also Chapter 3).

[27] Kim and Eaton (2015) used the hierarchical method, also known as the 'bass-ackwards' method, to derive a hierarchy of factors, rather than the higher-order model, but they imply a similar structure.

7.3.2 Implications for Clinical Assessment

A recurrent theme of this thesis is that mental disorders are not discrete entities as implied by current nosologies.[28] This is most apparent in Chapter 5, where a group of 'antisocial adolescents' were classified by a spectrum of severity spanning a range of emotional and behavioural problems. Furthermore, a cross-cutting spectrum of severity was estimated in a range of samples, from community volunteers (Chapter 4) to inpatients (Chapter 6), demonstrating its applicability to a range of populations encountered in research and practice.

The importance of disorder-general and disorder-specific assessment in clinical research is highlighted by the principle of intwined generality that pervades this thesis. To recap, this principle suggests that any measure captures general and specific aspects of a construct; to study the specific aspects, one must first control for the general aspects that will otherwise obscure one's observations (Gustafsson, 2002). The principle was evidenced throughout this thesis by comparisons between the bifactor and correlated factor models, the former controlling for the common variance and the latter conflating it with the specific variance. In Chapter 5, for instance, disorder-specific factors changed in nuanced ways over a psychosocial intervention when the general variance was controlled for with a *p* factor. By contrast, all disorder-specific factors declined in the correlated factors model. In Chapter 6, disorder-specific factors that were most representative of the general variance were only predictive of poorer depression outcomes once the general variance was explicitly controlled for.

[28] The foreword in the DSM-5 recognises the substantial overlap between diagnostic groups but disorder-specific assessments and treatments ultimately dictate.

The current chapters pose a daunting thought: our prior understanding of disorder-specific effects might be the result of broader psychological dimensions. However, it is questionable whether a disorder-specific approach has produced any significant advances to begin with. Consider, for instance, the limited evidence for biomarkers related to individual disorders (Insel, 2014), or the way in which risk factors relate non-specifically to multiple disorders (Keyes et al., 2012). The current chapters are a testament to the improved specificity in predictions after separating out the shared and specific aspects of mental disorders with the bifactor model. Disorder-specific research can be likened to navigating a long journey with road signs alone, while bifactor research is like using a road map which provides the specific junctions in addition to the overall route.

The bifactor model's implications for clinical research mainly concern the assessment of shared and specific aspects of mental disorders across people. However, Chapter 5 demonstrates that these shared and specific aspects can be summarized for a given individual, highlighting its applicability to clinical assessment in practice. Current assessment approaches aim to identify the DSM or ICD diagnoses that best fit an individual's presentation (Clark, Cuthbert, Lewis-Fernández, Narrow, & Reed, 2017). However, this approach falls short when clients present with multiple problems at different levels of severity (e.g., clinical, sub-threshold). 'Mixed' or 'unspecified' diagnoses might be offered, but these tend to misrepresent the overlap and continuity in people's problems, respectively (Krueger et al., 2018). Transdiagnostic approaches to assessment resolve these issues by focusing on dimensions that cut across disorders, which is what clinicians naturally do when formulating a client's problems in terms of their overarching causes (Rodriguez-Seijas, Eaton, & Krueger, 2015).

What might the assessment of general and specific psychopathology dimensions for a given client look like? Space limits a detailed discussion of this intricate and exciting question, but I will tease a vision based on the severity-style framework introduced in Chapter 6 and detailed by Bach and First (2018), Hopwood (2018), and Skodol, Morey, Bender, and Oldham (2015) in the context of personality pathology. To recap, the severity-style framework suggests that a client's presentation can be understood in terms of their overall impairment in addition to their specific personality characteristics or styles of coping (Hopwood et al., 2011).[29] These two components are distinct, such that one can present certain personality characteristics that are commonly associated with psychopathology (e.g., negative affectivity) but low levels of impairment (Livesley, 2011). Both components are necessary for painting a picture of a client's presentation with brush strokes that are neither too fine nor too coarse.

General psychopathological severity could be assessed by rating a client's level of impairment across several life domains, since general severity predicts key outcomes regardless of the actual disorders present, including academic and occupational functioning, social functioning, and clinical functioning (e.g., risk of relapse, hospitalization, and suicide; Conway et al., 2019, see also Chapter 5 where the *p* factor predicted delinquency and school attendance). Rather than inferring a client's level of severity based on their levels of comorbidity, a general severity dimension provides an explicit index of overall impairment that can be used to

[29] While personality disorder researchers are at the forefront of thought in applying quantitative models to clinical practice, their ideas apply to the *p* factor and specific psychopathology factors, which are thought to reflect severity and personality markers, respectively (Caspi et al., 2014).

define ranges of functioning across the spectrum of mental health (Kotov et al., 2017).

Specific personality or coping styles could be assessed by rating a client on measures of personality and emotion regulation. The resultant profiles could be used to predict the nature of current/future problems experienced and the kinds of interventions they are most suited to (Bach & First, 2018; see also Chapter 6 where specific personality disorder traits predicted particular treatment responses). It should be stressed that personality assessments do not preclude a focus on specific symptoms or disorders. Indeed, while clients typically present with several co-occurring issues, certain problems are usually more pertinent than others. However, the dimensional approach to assessment encourages the clinician to consider the relationships between problems, which are not isolated entities (Rodriguez-Seijas et al., 2015). It should also be noted that there is movement towards a dimensional assessment approach in the DSM-5, which includes severity measures in addition to diagnostic criteria for some disorders (American Psychiatric Association, 2013).

7.3.3 Implications for Treatment

A quantitative nosology promises to improve clinical practice, as comorbidity and diagnostic heterogeneity–products of the current diagnostic system–will no longer hinder our ability to predict treatment responses (Kotov et al., 2017). This can already be seen in Chapter 6, which produced a more refined picture of the personality styles that predicted poorer treatment outcomes after including both general and specific personality disorder dimensions as prognostic variables.

One could take the results in Chapter 6 a step further by triaging patients to different services and interventions based on their levels of general and specific psychopathology. A patient's overall severity might inform the intensity of the intervention they receive; for instance, mild, moderate, and severe levels of impairment might warrant self-help, outpatient services, or inpatient stay, respectively (Bach & First, 2018). Furthermore, a patient's specific psychopathology profile might inform the type of intervention they receive. For instance, internalizing traits might be best suited to interventions that target 'over-thinking' tendencies and the preceding/ensuing feelings of shame and guilt, such as exposure and response prevention with cognitive restructuring. By contrast, externalizing traits might be most suited to interventions that target difficulties in 'under-thinking' tendencies and the preceding/ensuing feelings of aggression and frustration, including mentalization-based and social skills training (see Hopwood et al., in press for similar ideas).

Hierarchical models might also help improve our understanding of therapeutic change. A key issue for future research is to explain why psychological interventions and pharmacotherapies designed to target specific disorders have widespread effects on multiple disorders (Barlow, Sauer-Zavala, Carl, Bullis, & Ellard, 2014; Hudson & Pope, 1990). Hierarchical models provide (at least) two answers. The first is that interventions show similar efficacy across disorders because they mainly target the *p* factor. By implication, the same intervention could be administered to all clients regardless of their specific disorders (Caspi & Moffitt, 2018; Meier & Meier, 2017).

This hypothesis could be tested by examining whether symptom changes are exclusive to, or strongest in, the *p* factor compared to disorder-specific factors. This was, in fact, tested in Chapter 5, but the *p* factor and specific factors both changed over a psychosocial intervention for antisocial behaviour. It may be that declines in the *p* factor reflected reductions in overall severity that would be achieved by any effective intervention, whereas changes in the specific factors reflected treatment-specific processes that would differ across interventions.

One could also test this hypothesis by applying the bifactor model to a meta-analysis of outcomes data belonging to various interventions or to psychotherapy Q-sort data. Strong effect sizes associated with the modality-general variance, and weak or non-existent effect sizes associated with the modality-specific variance, would support the hypothesis that interventions achieve change through shared mechanisms that influence a common target as captured by the *p* factor. Prior meta-analyses partially support this hypothesis as they emphasise non-specific treatment effects (Wampold & Imel, 2015). However, they might not accurately represent the specific effects of treatments as they do not control for the common variance.

A second hypothesis is that interventions show similar efficacy across disorders because they target spectral factors that lead to upstream changes in super spectral factors (Rodriguez-Seijas et al., 2015). Therefore, tailored interventions are necessary, but they may achieve their effect by targeting mechanisms associated with broadband dimensions rather than specific disorders. This would explain why there are subtle within-domain treatment effects; for example, CBT might be more effective for internalizing disorders than externalizing disorders in youth (Weisz, McCarty, & Valeri, 2006). It is also consistent with findings showing that common therapeutic mechanisms, such as a good working alliance, are necessary but not sufficient for clinically significant change; specific contents and/or tasks that fall within a coherent model of mind are also necessary (Wampold, 2015).

The second hypothesis can be tested by evaluating longitudinal changes in clinical outcomes with the higher-order model. The higher-order model is statistically equivalent to the correlated factors model, which was assessed over a psychosocial intervention in Chapter 5. While we can infer that the widespread declines observed in the disorder-specific correlated factors (synonymous with first-order factors) were partially driven by changes in an overarching factor (i.e. second-order factor, which was represented by the correlations between disorder-specific factors in the correlated factors model), we cannot determine by how much, since the variance predicted by first-order dimensions is a mix of first-order and second-order factors. Alternatively, we could determine the extent that spectral level factors change through super spectral factors by estimating the mediation terms between general and specific bifactor dimensions. Nonetheless, methods are needed to correctly orthogonalize the general and specific factors to avoid model mis-specification (e.g., by using factor scores; Koch et al., 2018).

The difference between these two hypotheses of treatment change mirrors the difference between the bifactor and higher-order models. As was discussed above, such differences might be a matter of style rather than substance (see section 7.3.1). Bifactor approaches emphasise change from the 'top-down', while higher-order approaches emphasise change from the 'bottom-up'. In fact, a proper comparison of these approaches might result in an intractable research agenda, since interventions that target broad dimensions are still communicated via specific means (Fonagy & Allison, 2014). The important point is that both approaches advocate transdiagnostic interventions that are beginning to gain traction (Barlow et al., 2014; Norton & Paulus, 2016). There is already evidence that transdiagnostic interventions produce similar, if not better, results compared to disorder-specific interventions and have added benefits such as lower attrition rates (Barlow et al., 2017; Newby et al., 2015).

7.4 Conclusions

The current thesis explored the methodological challenges and clinical utility of the bifactor model of psychopathology. Two main conclusions can be reached: First, self-report measures of psychopathology capture both shared and specific aspects of mental disorders that can be estimated with the bifactor model using general and specific psychopathology factors, respectively. These factors are substantive in nature but subject to the methodological conditions they are estimated in. Second, disorder-specific analyses of psychopathology measures conflate the shared and specific aspects of mental disorders, which can obscure inferences made about the predictors and components of therapeutic change. The bifactor model is a useful tool for separating out the shared and unique variance to gain a more nuanced understanding of clinical outcomes

References

Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., ... & Adya, V. B. (2016). Observation of gravitational waves from a binary black hole merger. *Physical review letters*, *116*(6), 061102. doi:10.1103/PhysRevLett.116.061102

Achenbach, T. M. (2006). As Others See Us. *Current Directions in Psychological Science, 15*(2), 94-98. doi:10.1111/j.0963-7214.2006.00414.x

Achenbach, T. M. (2009). *The Achenbach System of Empirically Based Assessment (ASEBA): Development, findings, theory, and applications*. Burlington: University of Vermont Research Center for Children, Youth, and Families.

Achenbach, T. M., Ivanova, M. Y., & Rescorla, L. A. (2017). Empirically based assessment and taxonomy of psychopathology for ages 1(1/2)-90+ years: Developmental, multi-informant, and multicultural findings. *Compr Psychiatry, 79*, 4-18. doi:10.1016/j.comppsych.2017.03.006

Achenbach, T. M., Ivanova, M. Y., Rescorla, L. A., Turner, L. V., & Althoff, R. R. (2016). Internalizing/Externalizing Problems: Review and Recommendations for Clinical and Research Applications. *J Am Acad Child Adolesc Psychiatry, 55*(8), 647-656. doi:10.1016/j.jaac.2016.05.012

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*(2), 213-232. doi:10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2003). *Manual for the ASEBA adult forms & profiles*. Research Center for Children, Youth, & Families, University of Vermont, Burlington, VT, USA.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch Rating Model and the Disordered Threshold Controversy. *Educational and Psychological Measurement, 72*(4), 547-573. doi:10.1177/0013164411432166

Adewuya, A. O., Atilola, O., Ola, B. A., Coker, O. A., Zachariah, M. P., Olugbile, O., . . . Idris, O. (2018). Current prevalence, comorbidity and associated factors for symptoms of depression and generalised anxiety in the Lagos State Mental Health Survey (LSMHS), Nigeria. *Compr Psychiatry, 81*, 60-65. doi:10.1016/j.comppsych.2017.11.010

Afzali, M. H., Sunderland, M., Carragher, N., & Conrod, P. (2017). The Structure of Psychopathology in Early Adolescence: Study of a Canadian Sample. *Can J Psychiatry*, 706743717737032. doi:10.1177/0706743717737032

Agosti, V., Hellerstein, D. J., & Stewart, J. W. (2009). Does personality disorder decrease the likelihood of remission in early-onset chronic depression? *Compr Psychiatry, 50*(6), 491-495. doi:10.1016/j.comppsych.2009.01.009

Aichholzer, J. (2014). Random intercept EFA of personality scales. *J Res Pers, 53*, 1-4. doi:10.1016/j.jrp.2014.07.001

Aitkin, M. & Aitkin, I. (2005). Bayesian inference for factor scores. In A. Maydeu-Olivares, & J. J. McArdle. (Eds.). *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald*, (pp. 207-22). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Alessandri, G., Vecchione, M., Fagnani, C., Bentler, P. M., Barbaranelli, C., Medda, E., . . . Caprara, G. V. (2010). Much More Than Model Fitting? Evidence for the Heritability of Method Effect Associated With Positively Worded Items of the Life Orientation Test Revised. *Structural Equation Modeling: A Multidisciplinary Journal, 17*(4), 642-653. doi:10.1080/10705511.2010.510064

Alpert, J. E., Fava, M., Uebelacker, L. A., Nierenberg, A. A., Pava, J. A., Worthington III, J. J., & Rosenbaum, J. F. (1999). Patterns of axis I comorbidity in early-onset versus late-onset major depressive disorder. *Biological Psychiatry*, *46*(2), 202-211.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Anderson, J. L., Sellbom, M., Bagby, R. M., Quilty, L. C., Veltri, C. O., Markon, K. E., & Krueger, R. F. (2013). On the convergence between PSY-5 domains and PID-5 domains and facets: implications for assessment of DSM-5 personality traits. *Assessment, 20*(3), 286-294. doi:10.1177/1073191112471141

Andrews, G., Slade, T., & Issakidis, C. (2002). Deconstructing current comorbidity: data from the Australian National Survey of Mental Health and Well-Being. *Br J Psychiatry, 181*, 306-314. doi:10.1192/bjp.181.4.306

Angold, A., Costello, E. J., Messer, S. C., Pickles, A., Winder, F., & Silver, D. (1995). The development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research*, *5*, 237-249.

Antonovsky, A. (1987). *Unraveling the Mystery of Health. How people manage stress and stay well*. San Francisco: Jossey-Bass.

Arias, V. B., & Arias, B. (2017). The negative wording factor of Core Self-Evaluations Scale (CSES): Methodological artifact, or substantive specific variance? *Personality and Individual Differences, 109*, 28-34. doi:10.1016/j.paid.2016.12.038

Arrindell, W. A., Urban, R., Carrozzino, D., Bech, P., Demetrovics, Z., & Roozen, H. G. (2017). SCL-90-R emotional distress ratings in substance use and impulse control disorders: One-factor, oblique first-order, higher-order, and bi-factor models compared. *Psychiatry Res, 255*, 173-185. doi:10.1016/j.psychres.2017.05.019

Aseltine, R. H., Gore, S., & Gordon, J. (2000). Life Stress, Anger and Anxiety, and Delinquency: An Empirical Test of General Strain Theory. *Journal of Health and Social Behavior, 41*(3), 256. doi:10.2307/2676320

Ashton, M. C., Lee, K., de Vries, R. E., Hendrickse, J., & Born, M. P. (2012). The maladaptive personality traits of the Personality Inventory for DSM-5 (PID-5) in relation to the HEXACO personality factors and schizotypy/dissociation. *J Pers Disord, 26*(5), 641-659. doi:10.1521/pedi.2012.26.5.641

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., . . . De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *J Pers Soc Psychol, 86*(2), 356-366. doi:10.1037/0022-3514.86.2.356

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. doi:10.1080/10705510903008204

Bach, B., & First, M. B. (2018). Application of the ICD-11 classification of personality disorders. *BMC Psychiatry, 18*(1), 351. doi:10.1186/s12888-018-1908-3

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*(2), 491-509.

Bagby, R. M., Sellbom, M., Ayearst, L. E., Chmielewski, M. S., Anderson, J. L., & Quilty, L. C. (2014). Exploring the hierarchical structure of the MMPI-2-RF Personality Psychopathology Five in psychiatric patient and university student samples. *J Pers Assess, 96*(2), 166-172. doi:10.1080/00223891.2013.825623

Balsters, J. H., Cussans, E., Diedrichsen, J., Phillips, K. A., Preuss, T. M., Rilling, J. K., & Ramnani, N. (2010). Evolution of the cerebellar cortex: the selective expansion of prefrontal-projecting cerebellar lobules. *Neuroimage, 49*(3), 2045-2052. doi:10.1016/j.neuroimage.2009.10.045

Barker, E. D., & Salekin, R. T. (2012). Irritable oppositional defiance and callous unemotional traits: is the association partially explained by peer victimization? *J Child Psychol Psychiatry, 53*(11), 1167-1175. doi:10.1111/j.1469-7610.2012.02579.x

Barlow, D. H., Farchione, T. J., Bullis, J. R., Gallagher, M. W., Murray-Latin, H., Sauer-Zavala, S., . . . Cassiello-Robbins, C. (2017). The Unified Protocol for Transdiagnostic Treatment of Emotional Disorders Compared With Diagnosis-Specific Protocols for Anxiety Disorders: A Randomized Clinical Trial. *JAMA Psychiatry, 74*(9), 875-884. doi:10.1001/jamapsychiatry.2017.2164

Barlow, D. H., Sauer-Zavala, S., Carl, J. R., Bullis, J. R., & Ellard, K. K. (2013). The Nature, Diagnosis, and Treatment of Neuroticism. *Clinical Psychological Science, 2*(3), 344-365. doi:10.1177/2167702613505532

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition, 21*(1), 37-46. doi:10.1016/0010-0277(85)90022-8

Bastiaansen, L., Rossi, G., Schotte, C., & De Fruyt, F. (2011). The structure of personality disorders: comparing the DSM-IV-TR Axis II classification with the five-factor model framework using structural equation modeling. *J Pers Disord, 25*(3), 378-396. doi:10.1521/pedi.2011.25.3.378

Baumgartner, H., & Steenkamp, J.-B. E. M. (2018). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*(2), 143-156. doi:10.1509/jmkr.38.2.143.18840

Baumgartner, H., & Weijters, B. (2015). Response biases in crosscultural measurement. In S. Ng, & A. Y. Lee (Eds.), *Handbook of culture and consumer psychology* (p. 150-180), Oxford University Press.

Baumgartner, J. N., Schneider, T. R., & Capiola, A. (2018). Investigating the relationship between optimism and stress responses: A biopsychosocial perspective. *Personality and Individual Differences, 129*, 114-118. doi:10.1016/j.paid.2018.03.021

Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M. T., . . . Bjorgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychol Med, 46*(16), 3359-3369. doi:10.1017/S0033291716002300

Bearman, S. K., & Weisz, J. R. (2015). Review: Comprehensive treatments for youth comorbidity - evidence-guided approaches to a complicated problem. *Child Adolesc Ment Health, 20*(3), 131-141. doi:10.1111/camh.12092

Beauducel, A., & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(2), 186-203. doi:10.1207/s15328007sem1302_2

Beaujean, A. (2015). John Carroll's Views on Intelligence: Bi-Factor vs. Higher-Order Models. *Journal of Intelligence, 3*(4), 121-136. doi:10.3390/jintelligence3040121

Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattell-Horn-Carroll factor models: differences between bifactor and higher order factor models in predicting language achievement. *Psychol Assess, 26*(3), 789-805. doi:10.1037/a0036745

Beck, A., Steer, R. & Brown, G. (1996). *Manual for the Beck Depression Inventory-II (BDI-II)*. Psychological Corporation: San Antonio.

Ben-Porath, Y. S. (2013). Assessing personality and psychopathology with self-report inventories. In R. Graham, & J. A Naglieri, *Handbook of Psychology, Vol. 10: Assessment*, (pp. 622–44). Hoboken, NJ: Wiley.

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., . . . Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Sci Rep, 5*, 8923. doi:10.1038/srep08923

Benning, S. D., Patrick, C. J., Blonigen, D. M., Hicks, B. M., & Iacono, W. G. (2005). Estimating facets of psychopathy from normal personality traits: a step toward community epidemiological investigations. *Assessment, 12*(1), 3-18. doi:10.1177/1073191104271223

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606. doi:10.1037/0033-2909.88.3.588

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*(3), 186-204. doi:10.1037/h0031474

Bijl, R. V., Ravelli, A., & van Zessen, G. (1998). Prevalence of psychiatric disorder in the general population: results of the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Social Psychiatry and Psychiatric Epidemiology, 33*(12), 587-595. doi:10.1007/s001270050098

Billiet, J. B., & McClendon, M. J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 608-628. doi:10.1207/s15328007sem0704_5

Birmaher, B., Ryan, N. D., Williamson, D. E., Brent, D. A., Kaufman, J., Dahl, R. E., . . . Nelson, B. (1996). Childhood and adolescent depression: a review of the past 10 years. Part I. *J Am Acad Child Adolesc Psychiatry, 35*(11), 1427-1439. doi:10.1097/00004583-199611000-00011

Black, L., Panayiotou, M., & Humphrey, N. (2019). The dimensionality and latent structure of mental health difficulties and wellbeing in early adolescence. *PLoS One, 14*(2), e0213018. doi:10.1371/journal.pone.0213018

Block, J. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI.* East Norwalk, CT, US: Appleton-Century-Crofts.

Bloemen, A. J. P., Oldehinkel, A. J., Laceulle, O. M., Ormel, J., Rommelse, N. N. J., & Hartman, C. A. (2018). The association between executive functioning and psychopathology: general or specific? *Psychol Med, 48*(11), 1787-1794. doi:10.1017/S0033291717003269

Blum, D., & Holling, H. (2017). Spearman's law of diminishing returns. A meta-analysis. *Intelligence, 65*, 60-66. doi:10.1016/j.intell.2017.07.004

Bohnke, J. R., & Croudace, T. J. (2015). Factors of psychological distress: clinical value, measurement substance, and methodological artefacts. *Soc Psychiatry Psychiatr Epidemiol, 50*(4), 515-524. doi:10.1007/s00127-015-1022-5

Bohnke, J. R., & Croudace, T. J. (2016). Calibrating well-being, quality of life and common mental disorder items: psychometric epidemiology in public mental health research. *Br J Psychiatry, 209*(2), 162-168. doi:10.1192/bjp.bp.115.165530

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley. doi:10.1002/9781118619179

Bolt, D. M., & Johnson, T. R. (2009). Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style. *Applied Psychological Measurement, 33*(5), 335-352. doi:10.1177/0146621608329891

Bolt, D. M., Lu, Y., & Kim, J. S. (2014). Measurement and control of response styles using anchoring vignettes: a model-based approach. *Psychol Methods, 19*(4), 528-541. doi:10.1037/met0000016

Bonifay, W., & Cai, L. (2017). On the Complexity of Item Response Theory Models. *Multivariate Behav Res, 52*(4), 465-484. doi:10.1080/00273171.2017.1309262

Bonifay, W., Lane, S. P., & Reise, S. P. (2016). Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science, 5*(1), 184-186. doi:10.1177/2167702616657069

Bonta, J., Blais, J., & Wilson, H. A. (2014). A theoretically informed meta-analysis of the risk for general and violent recidivism for mentally disordered offenders. *Aggression and Violent Behavior, 19*(3), 278-287. doi:10.1016/j.avb.2014.04.014

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16*(1), 5-13. doi:10.1002/wps.20375

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu Rev Clin Psychol, 9*, 91-121. doi:10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychol Rev, 110*(2), 203-219. doi:10.1037/0033-295X.110.2.203

Boschloo, L., Schoevers, R. A., van Borkulo, C. D., Borsboom, D., & Oldehinkel, A. J. (2016). The network structure of psychopathology in a community sample of preadolescents. *J Abnorm Psychol, 125*(4), 599-606. doi:10.1037/abn0000150

Boschloo, L., van Borkulo, C. D., Rhemtulla, M., Keyes, K. M., Borsboom, D., & Schoevers, R. A. (2015). The Network Structure of Symptoms of the Diagnostic and Statistical Manual of Mental Disorders. *PLoS One, 10*(9), e0137621. doi:10.1371/journal.pone.0137621

Boschloo, L., Vogelzangs, N., Smit, J. H., van den Brink, W., Veltman, D. J., Beekman, A. T., & Penninx, B. W. (2011). Comorbidity and risk indicators for alcohol use disorders among persons with anxiety and/or depressive disorders: findings from the Netherlands Study of Depression and Anxiety (NESDA). *J Affect Disord, 131*(1-3), 233-242. doi:10.1016/j.jad.2010.12.014

Boudreaux, M. J., South, S. C., & Oltmanns, T. F. (2019). Symptom-level analysis of DSM-IV/DSM-5 personality pathology in later life: Hierarchical structure and predictive validity across self- and informant ratings. *J Abnorm Psychol, 128*(5), 365-384. doi:10.1037/abn0000444

Brodbeck, J., Stulz, N., Itten, S., Regli, D., Znoj, H., & Caspar, F. (2014). The structure of psychopathological symptoms and the associations with DSM-diagnoses in treatment seeking individuals. *Compr Psychiatry, 55*(3), 714-726. doi:10.1016/j.comppsych.2013.11.001

Brown, T. A. (2014). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford.

Bruner, G. C. (2013). *Marketing Scales Handbook: Multi-item Measures for Consumer Insight Research* (Vol. 7). Fort Worth, TX: GCBII Productions.

Burt, S. A., Krueger, R. F., McGue, M., & Iacono, W. (2003). Parent-child conflict and the comorbidity among childhood externalizing disorders. *Arch Gen Psychiatry, 60*(5), 505-513. doi:10.1001/archpsyc.60.5.505

Butler, S., Baruch, G., Hickey, N., & Fonagy, P. (2011). A randomized controlled trial of multisystemic therapy and a statutory therapeutic intervention for young offenders. *J Am Acad Child Adolesc Psychiatry, 50*(12), 1220-1235 e1222. doi:10.1016/j.jaac.2011.09.017

Calkins, M. E., Merikangas, K. R., Moore, T. M., Burstein, M., Behr, M. A., Satterthwaite, T. D., . . . Gur, R. E. (2015). The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. *J Child Psychol Psychiatry, 56*(12), 1356-1369. doi:10.1111/jcpp.12416

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. doi:10.1037/h0046016

Carragher, N., Teesson, M., Sunderland, M., Newton, N. C., Krueger, R. F., Conrod, P. J., . . . Slade, T. (2016). The structure of adolescent psychopathology: a symptom-level analysis. *Psychol Med, 46*(5), 981-994. doi:10.1017/S0033291715002470

Carver, C. S., & Scheier, M. F. (2017). Optimism, Coping, and Well-Being. In C. L. Cooper, & J. C. Quick (Eds.), *The Handbook of Stress and Health: A Guide to Research and Practice* (pp. 400–414). Chichester: John Wiley & Sons Ltd.

Carver, C. S., Johnson, S. L., & Timpano, K. R. (2017). Toward a Functional View of the P Factor in Psychopathology. *Clin Psychol Sci, 5*(5), 880-889. doi:10.1177/2167702617710037

Casey, P., Birbeck, G., McDonagh, C., Horgan, A., Dowrick, C., Dalgard, O., . . . Group, O. (2004). Personality disorder, depression and functioning: results from the ODIN study. *J Affect Disord, 82*(2), 277-283. doi:10.1016/j.jad.2003.11.009

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., . . . Moffitt, T. E. (2014). The p Factor: One General Psychopathology

Factor in the Structure of Psychiatric Disorders? *Clin Psychol Sci, 2*(2), 119-137. doi:10.1177/2167702613497473

Caspi, A., & Moffitt, T. E. (2018). All for One and One for All: Mental Disorders in One Dimension. *Am J Psychiatry, 175*(9), 831-844. doi:10.1176/appi.ajp.2018.17121383

Castellanos-Ryan, N., Briere, F. N., O'Leary-Barrett, M., Banaschewski, T., Bokde, A., Bromberg, U., . . . Consortium, I. (2016). The structure of psychopathology in adolescence and its common personality and cognitive correlates. *J Abnorm Psychol, 125*(8), 1039-1052. doi:10.1037/abn0000193

Cattell, R. B. (1965). A biometrics invited paper. Factor analysis: An introduction to essentials I. The purpose and underlying models. *Biometrics*, *21*(1), 190-215.

Cerda, M., Sagdeo, A., & Galea, S. (2008). Comorbid forms of psychopathology: key patterns and future research directions. *Epidemiol Rev, 30*, 155-177. doi:10.1093/epirev/mxn003

Chalmers, R. P. (2018). On Misconceptions and the Limited Usefulness of Ordinal Alpha. *Educ Psychol Meas, 78*(6), 1056-1071. doi:10.1177/0013164417727036

Chan, K. W., Yim, C. K., & Lam, S. S. (2010). Is Customer Participation in Value Creation a Double-Edged Sword? Evidence from Professional Financial Services Across Cultures. *Journal of Marketing*, *74*(3), 48–64. doi:10.1509/jmkg.74.3.48

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaivete among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav Res Methods, 46*(1), 112-130. doi:10.3758/s13428-013-0365-7

Chandler, J., & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annu Rev Clin Psychol, 12*, 53-81. doi:10.1146/annurev-clinpsy-021815-093623

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A Comparison of Bifactor and Second-Order Models of Quality of Life. *Multivariate Behav Res, 41*(2), 189-225. doi:10.1207/s15327906mbr4102_5

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255. doi:10.1207/s15328007sem0902_5

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2014). A Mixture Group Bifactor Model for Binary Responses. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 375-395. doi:10.1080/10705511.2014.915371

Choi, K. W., Batchelder, A. W., Ehlinger, P. P., Safren, S. A., & O'Cleirigh, C. (2017). Applying network analysis to psychological comorbidity and health behavior: Depression, PTSD, and sexual risk in sexual minority men with trauma histories. *J Consult Clin Psychol, 85*(12), 1158-1170. doi:10.1037/ccp0000241

Choy, Y., Fyer, A. J., & Goodwin, R. D. (2007). Specific phobia and comorbid depression: a closer look at the National Comorbidity Survey data. *Compr Psychiatry, 48*(2), 132-136. doi:10.1016/j.comppsych.2006.10.010

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Appl Psychol Meas, 41*(3), 178-194. doi:10.1177/0146621616677520

Clapp, J. D., Grubaugh, A. L., Allen, J. G., Mahoney, J., Oldham, J. M., Fowler, J. C., . . . Frueh, B. C. (2013). Modeling trajectory of depressive symptoms among psychiatric inpatients: a latent growth curve approach. *J Clin Psychiatry, 74*(5), 492-499. doi:10.4088/JCP.12m07842

Clark, L. A., Cuthbert, B., Lewis-Fernandez, R., Narrow, W. E., & Reed, G. M. (2017). Three Approaches to Understanding and Classifying Mental Disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychol Sci Public Interest, 18*(2), 72-145. doi:10.1177/1529100617727266

Clark, L. A., Nuzum, H., & Ro, E. (2018). Manifestations of personality impairment severity: comorbidity, course/prognosis, psychosocial dysfunction, and 'borderline' personality features. *Curr Opin Psychol, 21*, 117-121. doi:10.1016/j.copsyc.2017.12.004

Clarkin, J. F., Petrini, M., & Diamond, D. (2019). Complex depression: The treatment of major depression and severe personality pathology. *J Clin Psychol, 75*(5), 824-833. doi:10.1002/jclp.22759

Coan, R. W. (1964). Facts, factors, and artifacts: The quest for psychological meaning. *Psychological Review, 71*(2), 123-140.

Constantinou, M. P., Goodyer, I. M., Eisler, I., Butler, S., Kraam, A., Scott, S., . . . Fonagy, P. (2019). Changes in General and Specific Psychopathology Factors Over a Psychosocial Intervention. *J Am Acad Child Adolesc Psychiatry, 58*(8), 776-786. doi:10.1016/j.jaac.2018.11.011

Conway, C. C., Forbes, M. K., Forbush, K. T., Fried, E. I., Hallquist, M. N., Kotov, R., . . . Eaton, N. R. (2019). A Hierarchical Taxonomy of Psychopathology Can Transform Mental Health Research. *Perspect Psychol Sci, 14*(3), 419-436. doi:10.1177/1745691618810696

Conway, C. C., Hammen, C., & Brennan, P. A. (2016). Optimizing Prediction of Psychosocial and Clinical Outcomes With a Transdiagnostic Model of Personality Disorder. *J Pers Disord, 30*(4), 545-566. doi:10.1521/pedi_2015_29_218

Conway, C. C., Mansolf, M., & Reise, S. P. (2019). Ecological validity of a quantitative classification system for mental illness in treatment-seeking adults. *Psychol Assess, 31*(6), 730-740. doi:10.1037/pas0000695

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. doi:10.1037/0021-9010.78.1.98

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality, 54*, 13-29. doi:10.1016/j.jrp.2014.07.003

Cote, J. A., & Buckley, M. R. (1987). Estimating Trait, Method, and Error Variance: Generalizing across 70 Construct Validation Studies. *Journal of Marketing Research, 24*(3), 315. doi:10.2307/3151642

Cox, B. J., Clara, I. P., Worobec, L. M., & Grant, B. F. (2012). An Empirical Evaluation of the Structure ofDSM-IVPersonality Disorders in a Nationally Representative Sample: Results of Confirmatory Factor Analysis in the National Epidemiologic Survey on Alcohol and Related Conditions Waves 1 and 2. *Journal of Personality Disorders*, 1-12. doi:10.1521/pedi_2012_26_039

Craigie, M. A., Saulsman, L. M., & Lampard, A. M. (2007). MCMI-III personality complexity and depression treatment outcome following group-based cognitive-behavioral therapy. *J Clin Psychol, 63*(12), 1153-1170. doi:10.1002/jclp.20406

Cramer, A. O., Waldorp, L. J., van der Maas, H. L., & Borsboom, D. (2010). Comorbidity: a network perspective. *Behav Brain Sci, 33*(2-3), 137-150; discussion 150-193. doi:10.1017/S0140525X09991567

Crawford, T. N., Cohen, P., First, M. B., Skodol, A. E., Johnson, J. G., & Kasen, S. (2008). Comorbid Axis I and Axis II disorders in early adolescence: outcomes 20 years later. *Arch Gen Psychiatry, 65*(6), 641-648. doi:10.1001/archpsyc.65.6.641

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cudeck, R., & Browne, M. W. (1983). Cross-Validation Of Covariance Structures. *Multivariate Behav Res, 18*(2), 147-167. doi:10.1207/s15327906mbr1802_2

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *J Consult Clin Psychol, 76*(6), 909-922. doi:10.1037/a0013075

Cyranowski, J. M., Frank, E., Winter, E., Rucci, P., Novick, D., Pilkonis, P., . . . Kupfer, D. J. (2004). Personality pathology and outcome in recurrently depressed women over 2 years of maintenance interpersonal psychotherapy. *Psychol Med, 34*(4), 659-669. doi:10.1017/S0033291703001661

David, D., Cristea, I., & Hofmann, S. G. (2018). Why Cognitive Behavioral Therapy Is the Current Gold Standard of Psychotherapy. *Front Psychiatry, 9*, 4. doi:10.3389/fpsyt.2018.00004

Daviss, W. B., Birmaher, B., Melhem, N. A., Axelson., D. A., Michaels, S. M., & Brent, D. A., (2006). Criterion validity of the Mood and Feelings Questionnaire for depressive episodes in clinic and non-clinic subjects. *Journal of child psychology and psychiatry, and allied disciplines*. *47*(9), 927–934.

De Bolle, M., De Fruyt, F., Quilty, L. C., Rolland, J. P., Decuyper, M., & Bagby, R. M. (2011). Does personality disorder co-morbidity impact treatment outcome for patients with major depression? A multi-level analysis. *J Pers Disord, 25*(1), 1-15. doi:10.1521/pedi.2011.25.1.1

Del Giudice, M. (2014). An Evolutionary Life History Framework for Psychopathology. *Psychological Inquiry, 25*(3-4), 261-300. doi:10.1080/1047840x.2014.884918

Deutz, M. H. F., Shi, Q., Vossen, H. G. M., Huijding, J., Prinzie, P., Dekovic, M., . . . Woltering, S. (2018). Evaluation of the Strengths and Difficulties Questionnaire-Dysregulation Profile (SDQ-DP). *Psychol Assess, 30*(9), 1174-1185. doi:10.1037/pas0000564

Dickinson, D. (2017). "If the Shoe Fits ...": The Hierarchical Structure of Psychopathology and Psychiatric Neuroimaging. *Biol Psychiatry Cogn Neurosci Neuroimaging, 2*(4), 303-304. doi:10.1016/j.bpsc.2017.03.015

DiStefano, C., & Motl, R. W. (2006). Further Investigating Method Effects Associated With Negatively Worded Items on Self-Report Surveys. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(3), 440-464. doi:10.1207/s15328007sem1303_6

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proc Natl Acad Sci U S A, 108*(19), 7716-7720. doi:10.1073/pnas.1018601108

Dueber, D. M. (2017). *Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models*. https://doi.org/10.13023/edp.tool.01

Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An Introductory Guide to Latent Variable Growth Curve Modeling. *Soc Personal Psychol Compass, 3*(6), 979-991. doi:10.1111/j.1751-9004.2009.00224.x

Eaton, N. R. (2015). Latent variable and network models of comorbidity: toward an empirically derived nosology. *Soc Psychiatry Psychiatr Epidemiol, 50*(6), 845-849. doi:10.1007/s00127-015-1012-7

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor Models for Predicting Criteria by General and Specific Factors: Problems of Nonidentifiability and Alternative Solutions. *J Intell, 6*(3). doi:10.3390/jintelligence6030042

Eisenberg, N., Valiente, C., Spinrad, T. L., Liew, J., Zhou, Q., Losoya, S. H., . . . Cumberland, A. (2009). Longitudinal relations of children's effortful control, impulsivity, and negative emotionality to their externalizing, internalizing, and co-occurring behavior problems. *Dev Psychol, 45*(4), 988-1008. doi:10.1037/a0016213

Elliott, M. L., Romer, A., Knodt, A. R., & Hariri, A. R. (2018). A Connectome-wide Functional Signature of Transdiagnostic Risk for Mental Illness. *Biol Psychiatry, 84*(6), 452-459. doi:10.1016/j.biopsych.2018.03.012

Ellis, L. K., & Posner, M. I. (2004). Temperament and self-regulation. In R. F. Baumeister & K. D. Vohs (Eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 357–370). New York: Guilford Press.

Embretson, S, & Reise, S. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychol Methods, 16*(1), 1-16. doi:10.1037/a0022640

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *J Pers, 51*(3), 360-392. doi:10.1111/j.1467-6494.1983.tb00338.x

Erkens, N., Schramm, E., Kriston, L., Hautzinger, M., Harter, M., Schweiger, U., & Klein, J. P. (2018). Association of comorbid personality disorders with clinical characteristics and outcome in a randomized controlled trial comparing two psychotherapies for early-onset persistent depressive disorder. *J Affect Disord, 229*, 262-268. doi:10.1016/j.jad.2017.12.091

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science, 38*(2), 189-200. doi:10.1177/0165551512437638

Etkin, A., Buchel, C., & Gross, J. J. (2015). The neural bases of emotion regulation. *Nat Rev Neurosci, 16*(11), 693-700. doi:10.1038/nrn4044

Euler, F., Jenkel, N., Stadler, C., Schmeck, K., Fegert, J. M., Kolch, M., & Schmid, M. (2015). Variants of girls and boys with conduct disorder: anxiety symptoms and callous-unemotional traits. *J Abnorm Child Psychol, 43*(4), 773-785. doi:10.1007/s10802-014-9946-x

Eysenck, H. J., & Eysenck, M. W. (1985) *Personality and individual differences: a natural science approach*. New York: Plenum.

Eysenck, S. B. G., & Eysenck, H. J. (1963). On the Dual Nature of Extraversion. *British Journal of Social and Clinical Psychology, 2*(1), 46-55. doi:10.1111/j.2044-8260.1963.tb00375.x

Fabozzi, F., Focardi, S., Rachev, S., & Arshanapalli, B. (2014). *The Basics of Financial Econometrics*. Hoboken, NJ: John Wiley & Sons.

Falkenstrom, F., Granstrom, F., & Holmqvist, R. (2013). Therapeutic alliance predicts symptomatic improvement session by session. *J Couns Psychol, 60*(3), 317-328. doi:10.1037/a0032258

Fava, M., Alpert, J. E., Borus, J. S., Nierenberg, A. A., Pava, J. A., & Rosenbaum, J. F. (1996). Patterns of personality disorder comorbidity in early-onset versus late-onset major depression. *Am J Psychiatry, 153*(10), 1308-1312. doi:10.1176/ajp.153.10.1308

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences, 75*, 47-52. doi:10.1016/j.paid.2014.11.017

Ferentinos, P., Yotsidi, V., Porichi, E., Douzenis, A., Papageorgiou, C., & Stalikas, A. (2019). Well-being in Patients with Affective Disorders Compared to Nonclinical Participants: A Multi-Model Evaluation of the Mental Health Continuum-Short Form. *J Clin Psychol*. doi:10.1002/jclp.22780

Fergus, T. A., & Bardeen, J. R. (2014). Emotion regulation and obsessive–compulsive symptoms: A further examination of associations. *Journal of Obsessive-Compulsive and Related Disorders, 3*(3), 243-248. doi:10.1016/j.jocrd.2014.06.001

Fernandez de la Cruz, L., Vidal-Ribas, P., Zahreddine, N., Mathiassen, B., Brondbo, P. H., Simonoff, E., . . . Stringaris, A. (2018). Should Clinicians Split or Lump Psychiatric Symptoms? The Structure of Psychopathology in Two Large Pediatric Clinical Samples from England and Norway. *Child Psychiatry Hum Dev, 49*(4), 607-620. doi:10.1007/s10578-017-0777-1

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., . . . Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med, 10*(11), e1001547. doi:10.1371/journal.pmed.1001547

Finch, H. (2010). Multidimensional Item Response Theory Parameter Estimation With Nonsimple Structure Items. *Applied Psychological Measurement, 35*(1), 67-82. doi:10.1177/0146621610367787

Finkelhor, D., Ormrod, R. K., & Turner, H. A. (2007). Polyvictimization and trauma in a national longitudinal cohort. *Dev Psychopathol, 19*(1), 149-166. doi:10.1017/S0954579407070083

First, M. B., Skodol, A. E., Bender, D. S., Oldham, J. M. (2017). *Structured Clinical Interview for the DSM–5 Alternative Model for Personality Disorders (SCID–AMPD)*. New York: New York State Psychiatric Institute.

First, M. B., Spitzer, R., Gibbon, M., Williams, J., & Benjamin, L. (1994). *Structured Clinical Interview for DSM–IV Axis II personality disorders (SCID II)*. New York, NY: Biometric Research Department.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol Methods, 9*(4), 466-491. doi:10.1037/1082-989X.9.4.466

Fonagy, P., & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy (Chic), 51*(3), 372-380. doi:10.1037/a0036505

Fonagy, P., Butler, S., Cottrell, D., Scott, S., Pilling, S., Eisler, I., . . . Goodyer, I. M. (2018). Multisystemic therapy versus management as usual in the treatment of adolescent antisocial behaviour (START): a pragmatic, randomised controlled, superiority trial. *The Lancet Psychiatry, 5*(2), 119-133. doi:10.1016/s2215-0366(18)30001-4

Fonagy, P, Cottrell, D, Phillips, J., Bevington, D., Glaser, D., & Allison, E. (2014). *What works for whom? A critical review of treatments for children and adolescents* (2nd ed.). New York, NY: Guilford Press.

Fonagy, P., Gergely, G., Jurist, E., & Target, M. (2002). *Affect regulation, mentalization and the development of the self*. New York, NY: Other Press

Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2017). What we have changed our minds about: Part 1. Borderline personality disorder as a limitation of resilience. *Borderline Personal Disord Emot Dysregul, 4*, 11. doi:10.1186/s40479-017-0061-9

Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2017). What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. *Borderline Personal Disord Emot Dysregul, 4*, 9. doi:10.1186/s40479-017-0062-8

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *J Abnorm Psychol, 126*(7), 969-988. doi:10.1037/abn0000276

Forbey, J. D., Lee, T. T., Ben-Porath, Y. S., Arbisi, P. A., & Gartland, D. (2013). Associations between MMPI-2-RF validity scale scores and extra-test measures of personality and psychopathology. *Assessment, 20*(4), 448-461. doi:10.1177/1073191113478154

Fossati, A., Maffei, C., Bagnato, M., Battaglia, M., Donati, D., Donini, M., . . . Prolo, F. (2000). Patterns of covariation of DSM-IV personality disorders in a mixed psychiatric sample. *Comprehensive Psychiatry, 41*(3), 206-215. doi:10.1016/s0010-440x(00)90049-x

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review, 80*(5), 875-908. doi:10.1177/0003122415601618

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Gallop, R., Amsterdam, J. D., & Hollon, S. D. (2008). Antidepressant medications v. cognitive therapy in people with depression with or without personality disorder. *Br J Psychiatry, 192*(2), 124-129. doi:10.1192/bjp.bp.107.037234

Fowler, J. C., Clapp, J. D., Madan, A., Allen, J. G., Frueh, B. C., Fonagy, P., & Oldham, J. M. (2018). A naturalistic longitudinal study of extended inpatient treatment for adults with borderline personality disorder: An examination of treatment response, remission and deterioration. *J Affect Disord, 235*, 323-331. doi:10.1016/j.jad.2017.12.054

Fowler, J. C., Clapp, J. D., Madan, A., Allen, J. G., Frueh, B. C., & Oldham, J. M. (2017). An Open Effectiveness Trial of a Multimodal Inpatient Treatment for Depression and Anxiety Among Adults With Serious Mental Illness. *Psychiatry, 80*(1), 42-54. doi:10.1080/00332747.2016.1196072

Fowler, J. C., Madan, A., Allen, J. G., Patriquin, M., Sharp, C., Oldham, J. M., & Frueh, B. C. (2018). Clinical utility of the DSM-5 alternative model for borderline personality disorder: Differential diagnostic accuracy of the BFI, SCID-II-PQ, and PID-5. *Compr Psychiatry, 80*, 97-103. doi:10.1016/j.comppsych.2017.09.003

Fowler, J. C., & Oldham, J. M. (2013). Co-Occurring Disorders and Treatment Complexity Within Personality Disorders. *Focus, 11*(2), 123-128. doi:10.1176/appi.focus.11.2.123

French, L. R. M., Turner, K. M., Dawson, S., & Moran, P. (2017). Psychological treatment of depression and anxiety in patients with co-morbid personality disorder: A scoping study of trial evidence. *Personal Ment Health, 11*(2), 101-117. doi:10.1002/pmh.1372

Friborg, O., Martinsen, E. W., Martinussen, M., Kaiser, S., Overgard, K. T., & Rosenvinge, J. H. (2014). Comorbidity of personality disorders in mood disorders: a meta-analytic review of 122 studies from 1988 to 2010. *J Affect Disord, 152-154*, 1-11. doi:10.1016/j.jad.2013.08.023

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Can callous-unemotional traits enhance the understanding, diagnosis, and treatment of serious conduct problems in children and adolescents? A comprehensive review. *Psychol Bull, 140*(1), 1-57. doi:10.1037/a0033076

Frick, P. J., Stickle, T. R., Dandreaux, D. M., Farrell, J. M., & Kimonis, E. R. (2005). Callous–Unemotional Traits in Predicting the Severity and Stability of Conduct Problems and Delinquency. *Journal of Abnormal Child Psychology, 33*(4), 471-487. doi:10.1007/s10648-005-5728-9

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philos Trans R Soc Lond B Biol Sci, 358*(1431), 459-473. doi:10.1098/rstb.2002.1218

Furukawa, T. A., Cipriani, A., Atkinson, L. Z., Leucht, S., Ogawa, Y., Takeshima, N., . . . Salanti, G. (2016). Placebo response rates in antidepressant trials: a systematic review of published and unpublished double-blind randomised controlled studies. *The Lancet Psychiatry, 3*(11), 1059-1066. doi:10.1016/s2215-0366(16)30307-8

Garcia-Perez, M. A. (2017). An Analysis of (Dis)Ordered Categories, Thresholds, and Crossings in Difference and Divide-by-Total IRT Models for Ordered Responses. *Span J Psychol, 20*, E10. doi:10.1017/sjp.2017.11

Geeraerts, S. B., Deutz, M. H., Dekovic, M., Bunte, T., Schoemaker, K., Espy, K. A., . . . Matthys, W. (2015). The Child Behavior Checklist Dysregulation Profile in Preschool Children: A Broad Dysregulation Syndrome. *J Am Acad Child Adolesc Psychiatry, 54*(7), 595-602 e592. doi:10.1016/j.jaac.2015.04.012

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436. doi:10.1007/bf02295430

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: an illustration of the bi-factor item response theory model. *J Psychiatr Res, 43*(4), 401-410. doi:10.1016/j.jpsychires.2008.04.013

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science, 50*(1), 21-43.

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence, 42*, 89-97. doi:10.1016/j.intell.2013.11.004

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence, 55*, 57-68. doi:10.1016/j.intell.2016.01.006

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012-1014. doi:10.1038/nature07634

Gnambs, T., & Schroeders, U. (2017). Cognitive Abilities Explain Wording Effects in the Rosenberg Self-Esteem Scale. *Assessment*, 1073191117746503. doi:10.1177/1073191117746503

Goldberg, L. R. (2006). Doing it all Bass-Ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality, 40*(4), 347-358. doi:10.1016/j.jrp.2006.01.001

Gomez, R., Stavropoulos, V., Vance, A., & Griffiths, M. D. (2018). Re-evaluation of the Latent Structure of Common Childhood Disorders: Is There a General Psychopathology Factor (P-Factor)? *International Journal of Mental Health and Addiction, 17*(2), 258-278. doi:10.1007/s11469-018-0017-3

Goodboy, A. K., & Kline, R. B. (2016). Statistical and Practical Concerns With Published Communication Research Featuring Structural Equation Modeling. *Communication Research Reports, 34*(1), 68-77. doi:10.1080/08824096.2016.1214121

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry, 38*(5), 581-586. doi: 10.1111/j.1469-7610.1997.tb01545.x

Goodman. R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*(11), 1337-1345.

Goodman, A. (2010). Substance use and common child mental health problems: examining longitudinal associations in a British sample. *Addiction, 105*(8), 1484-1496. doi:10.1111/j.1360-0443.2010.02981.x

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: Description and Initial Validation of an Integrated Assessment of Child and Adolescent Psychopathology. *Journal of Child Psychology and Psychiatry, 41*(5), 645-655. doi: 10.1111/j.1469-7610.2000.tb02345.x

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum

Gorwood, P., Rouillon, F., Even, C., Falissard, B., Corruble, E., & Moran, P. (2010). Treatment response in major depression: effects of personality dysfunction and prior depression. *Br J Psychiatry, 196*(2), 139-142. doi:10.1192/bjp.bp.109.067058

Grant, B. F. (1995). Comorbidity between DSM-IV drug use disorders and major depression: Results of a national survey of adults. *Journal of Substance Abuse, 7*(4), 481-497. doi:10.1016/0899-3289(95)90017-9

Grant, B. F., & Harford, T. C. (1995). Comorbidity between DSM-IV alcohol use disorders and major depression: results of a national survey. *Drug and Alcohol Dependence, 39*(3), 197-206. doi:10.1016/0376-8716(95)01160-4

Grant, B. F., Hasin, D. S., Stinson, F. S., Dawson, D. A., Patricia Chou, S., June Ruan, W., & Huang, B. (2005). Co-occurrence of 12-month mood and anxiety disorders and personality disorders in the US: results from the national epidemiologic survey on alcohol and related conditions. *J Psychiatr Res, 39*(1), 1-9. doi:10.1016/j.jpsychires.2004.05.004

Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind perception in psychopathology. *Proc Natl Acad Sci U S A, 108*(2), 477-479. doi:10.1073/pnas.1015493108

Greene, A. L., & Eaton, N. R. (2017). The temporal stability of the bifactor model of comorbidity: An examination of moderated continuity pathways. *Compr Psychiatry, 72*, 74-82. doi:10.1016/j.comppsych.2016.09.010

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., . . . Kotov, R. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *J Abnorm Psychol*. doi:10.1037/abn0000434

Greenleaf, E. A. (1992). Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles. *Journal of Marketing Research*, *29*(2), 176. doi:10.2307/3172568

Griffith, J. W., Zinbarg, R. E., Craske, M. G., Mineka, S., Rose, R. D., Waters, A. M., & Sutton, J. M. (2010). Neuroticism as a common dimension in the internalizing disorders. *Psychol Med, 40*(7), 1125-1136. doi:10.1017/S0033291709991449

Grilo, C. M., Sanislow, C. A., Shea, M. T., Skodol, A. E., Stout, R. L., Gunderson, J. G., . . . McGlashan, T. H. (2005). Two-year prospective naturalistic study of remission from major depressive disorder as a function of personality disorder comorbidity. *J Consult Clin Psychol, 73*(1), 78-85. doi:10.1037/0022-006X.73.1.78

Grilo, C. M., Stout, R. L., Markowitz, J. C., Sanislow, C. A., Ansell, E. B., Skodol, A. E., . . . McGlashan, T. H. (2010). Personality disorders predict relapse after remission from an episode of major depressive disorder: a 6-year prospective study. *J Clin Psychiatry, 71*(12), 1629-1635. doi:10.4088/JCP.08m04200gre

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika*, *6*(2), 67-77.

Gunderson, J. G., Morey, L. C., Stout, R. L., Skodol, A. E., Shea, M. T., McGlashan, T. H., . . . Bender, D. S. (2004). Major depressive disorder and borderline personality disorder revisited: longitudinal interactions. *J Clin Psychiatry, 65*(8), 1049-1056. doi:10.4088/jcp.v65n0804

Gustafsson, J. E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73-95). London: Erlbaum

Gustafsson, J. E., & Balke, G. (1993). General and Specific Abilities as Predictors of School Achievement. *Multivariate Behav Res, 28*(4), 407-434. doi:10.1207/s15327906mbr2804_2

Gustafsson, J. E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. 97-121. doi:10.1037/12074-005

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *The British Journal of Statistical Psychology*, *8*, 65-81.

Hair, J. F., Tatham, R. L., Anderson, R. E., & Black, W. (1998). *Multivariate data analysis*. (5th ed.) Prentice-Hall: London.

Haltigan, J. D., Aitken, M., Skilling, T., Henderson, J., Hawke, L., Battaglia, M., . . . Andrade, B. F. (2018). "P" and "DP:" Examining Symptom-Level Bifactor Models of Psychopathology and Dysregulation in Clinically Referred

Children and Adolescents. *J Am Acad Child Adolesc Psychiatry, 57*(6), 384-396. doi:10.1016/j.jaac.2018.03.010

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry, 23*, 56-62.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future – A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Hankin, B. L., Davis, E. P., Snyder, H., Young, J. F., Glynn, L. M., & Sandman, C. A. (2017). Temperament factors and dimensional, latent bifactor models of child psychopathology: Transdiagnostic and specific associations in two youth samples. *Psychiatry Res, 252*, 139-146. doi:10.1016/j.psychres.2017.02.061

Harden, K. P., Engelhardt, L. E., Mann, F. D., Patterson, M. W., Grotzinger, A. D., Savicki, S. L., . . . Tucker-Drob, E. M. (2019). Genetic Associations Between Executive Functions and a General Factor of Psychopathology. *J Am Acad Child Adolesc Psychiatry*. doi:10.1016/j.jaac.2019.05.006

Hardy, G. E., Barkham, M., Shapiro, D. A., Stiles, W. B., Rees, A., & Reynolds, S. (1995). Impact of Cluster C personality disorders on outcomes of contrasting brief psychotherapies for depression. *Journal of Consulting and Clinical Psychology, 63*(6), 997-1004. doi:10.1037/0022-006x.63.6.997

Harley, R., Petersen, T., Scalia, M., Papakostas, G. I., Farabaugh, A., & Fava, M. (2006). Problem-solving ability and comorbid personality disorders in depressed outpatients. *Depress Anxiety, 23*(8), 496-501. doi:10.1002/da.20194

Harman, H. H. (1960). *Modern Factor Analysis*. Chicago: University of Chicago Press.

Harms, C., Jackel, L., & Montag, C. (2017). Reliability and completion speed in online questionnaires under consideration of personality. *Personality and Individual Differences, 111*, 281-290. doi:10.1016/j.paid.2017.02.015

Hayes, A. M., Laurenceau, J. P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: the study of nonlinear and discontinuous patterns of change in psychotherapy. *Clin Psychol Rev, 27*(6), 715-723. doi:10.1016/j.cpr.2007.01.008

Hayes, J. F., Maughan, D. L., & Grant-Peterkin, H. (2016). Interconnected or disconnected? Promotion of mental health and prevention of mental disorder in the digital age. *Br J Psychiatry, 208*(3), 205-207. doi:10.1192/bjp.bp.114.161067

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Hengartner, M. P., Ajdacic-Gross, V., Rodgers, S., Muller, M., & Rossler, W. (2014). The joint structure of normal and pathological personality: further evidence for a dimensional model. *Compr Psychiatry, 55*(3), 667-674. doi:10.1016/j.comppsych.2013.10.011

Henggeler, S. W. (2012). Multisystemic Therapy: Clinical Foundations and Research Outcomes. *Psychosocial Intervention, 21*(2), 181-193. doi:10.5093/in2012a12

Henggeler, S. W., Rowland, M. D., Randall, J., Ward, D. M., Pickrel, S. G., Cunningham, P. B., . . . Santos, A. B. (1999). Home-based multisystemic therapy as an alternative to the hospitalization of youths in psychiatric crisis: clinical outcomes. *J Am Acad Child Adolesc Psychiatry, 38*(11), 1331-1339. doi:10.1097/00004583-199911000-00006

Henggeler, S. W., & Schaeffer, C. M. (2016). Multisystemic Therapy((R)) : Clinical Overview, Outcomes, and Implementation Research. *Fam Process, 55*(3), 514-528. doi:10.1111/famp.12232

Hewson, C., Vogel, C., & Laurent, D. (2016). *Internet research methods* (2nd ed.). London, England: Sage.

Hinton, K. E., Lahey, B. B., Villalta-Gil, V., Meyer, F. A. C., Burgess, L. L., Chodes, L. K., . . . Zald, D. H. (2019). White matter microstructure correlates of general and specific second-order factors of psychopathology. *Neuroimage Clin, 22*, 101705. doi:10.1016/j.nicl.2019.101705

Hoffman, D. L., Kopalle, P. K., & Novak, T. P. (2010). The "Right" Consumers for Better Concepts: Identifying Consumers High in Emergent Nature to Develop New Product Concepts. *Journal of Marketing Research*, *47*(5), 854–865. doi:10.1509/jmkr.47.5.854

Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T., & Fang, A. (2012). The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cognit Ther Res, 36*(5), 427-440. doi:10.1007/s10608-012-9476-1

Hofmann, S. G., & Barlow, D. H. (2014). Evidence-based psychological interventions and the common factors approach: the beginnings of a rapprochement? *Psychotherapy (Chic), 51*(4), 510-513. doi:10.1037/a0037045

Hogue, A., Dauber, S., Samuolis, J., & Liddle, H. A. (2006). Treatment techniques and outcomes in multidimensional family therapy for adolescent behavior problems. *J Fam Psychol, 20*(4), 535-543. doi:10.1037/0893-3200.20.4.535

Holzinger, K. J. (1945). Spearman as I knew him. *Psychometrika, 10*(4), 231-235. doi:10.1007/bf02288890

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika, 2*(1), 41-54. doi:10.1007/bf02287965

Hopwood, C. J. (2018). A framework for treating DSM-5 alternative model for personality disorder features. *Personal Ment Health, 12*(2), 107-125. doi:10.1002/pmh.1414

Hopwood, C., Bagby, R. M., Gralnick, T. M., Ro, E., Ruggero, C., Mullins-Sweatt, S., ... & Patrick, C. J. (2018). Integrating psychotherapy with the Hierarchical Taxonomy of Psychopathology (HiTOP). (in press). *Journal of Psychotherapy Integration.* doi:10.1037/int0000156

Hopwood, C. J., Malone, J. C., Ansell, E. B., Sanislow, C. A., Grilo, C. M., McGlashan, T. H., . . . Morey, L. C. (2011). Personality assessment in DSM-5: empirical support for rating severity, style, and traits. *J Pers Disord, 25*(3), 305-320. doi:10.1521/pedi.2011.25.3.305

Hopwood, C. J., Wright, A. G., Krueger, R. F., Schade, N., Markon, K. E., & Morey, L. C. (2013). DSM-5 pathological personality traits and the personality assessment inventory. *Assessment, 20*(3), 269-285. doi:10.1177/1073191113486286

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording Effects in Self-Esteem Scales: Methodological Artifact or Response Style? *Structural Equation Modeling: A Multidisciplinary Journal, 10*(3), 435-455. doi:10.1207/s15328007sem1003_6

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Hudson, J. I., & Pope, H. G., Jr. (1990). Affective spectrum disorder: does antidepressant response identify a family of disorders with a common pathophysiology? *Am J Psychiatry, 147*(5), 552-564. doi:10.1176/ajp.147.5.552

Huey, S. J., Jr., Henggeler, S. W., Rowland, M. D., Halliday-Boykins, C. A., Cunningham, P. B., Pickrel, S. G., & Edwards, J. (2004). Multisystemic therapy effects on attempted suicide by youths presenting psychiatric emergencies. *J Am Acad Child Adolesc Psychiatry, 43*(2), 183-190. doi:10.1097/00004583-200402000-00014

Humphreys, K. L., Gleason, M. M., Drury, S. S., Miron, D., Nelson, C. A., Fox, N. A., & Zeanah, C. H. (2015). Effects of institutional rearing and foster care on psychopathology at age 12 years in Romania: follow-up of an open, randomised controlled trial. *The Lancet Psychiatry, 2*(7), 625-634. doi:10.1016/s2215-0366(15)00095-4

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17*(7), 475-483. doi:10.1037/h0041550

Hung, I. W., & Labroo, A. A. (2011). From Firm Muscles to Firm Willpower: Understanding the Role of Embodied Cognition in Self-Regulation. *Journal of Consumer Research*, *37*(6), 1046–1064. doi:10.1086/657240

Huprich, S. K., Schmitt, T. A., Richard, D. C., Chelminski, I., & Zimmerman, M. A. (2010). Comparing factor analytic models of the DSM-IV personality disorders. *Personal Disord, 1*(1), 22-37. doi:10.1037/a0018245

Hyland, P., Murphy, J., Shevlin, M., Carey, S., Vallieres, F., Murphy, D., & Elklit, A. (2018). Correlates of a general psychopathology factor in a clinical sample of childhood sexual abuse survivors. *J Affect Disord, 232*, 109-115. doi:10.1016/j.jad.2018.02.048

Ignatyev, Y., Baggio, S., & Mundt, A. P. (2019). The Underlying Structure of Comorbid Mental Health and Substance Use Disorders in Prison Populations. *Psychopathology, 52*(1), 2-9. doi:10.1159/000495844

Ilardi, S. S., & Head, W. C. E. (1995). Personality pathology and response to somatic treatments for major depression: a critical review. *Depression*, *2*(4), 200–217

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *Am J Psychiatry, 171*(4), 395-397. doi:10.1176/appi.ajp.2014.14020138

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Turner, L. V., Arnadottir, H. A., Au, A., . . . Zasepa, E. (2015). Syndromes of collateral-reported psychopathology for ages 18-59 in 18 Societies. *Int J Clin Health Psychol, 15*(1), 18-28. doi:10.1016/j.ijchp.2014.07.001

Jahng, S., Trull, T. J., Wood, P. K., Tragesser, S. L., Tomko, R., Grant, J. D., . . . Sher, K. J. (2011). Distinguishing general and specific personality disorder features and implications for substance dependence comorbidity. *J Abnorm Psychol, 120*(3), 656-669. doi:10.1037/a0023539

Jensen, A. R. (1998). *Human evolution, behavior, and intelligence. The g factor: The science of mental ability.* Westport, CT, US: Praeger Publishers/Greenwood Publishing Group.

Jeronimus, B. F., Kotov, R., Riese, H., & Ormel, J. (2016). Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: a meta-analysis on 59 longitudinal/prospective studies with 443 313 participants. *Psychol Med, 46*(14), 2883-2906. doi:10.1017/S0033291716001653

John, L. K., Acquisti, A., & Loewenstein, G. (2011). Strangers on a Plane: Context-Dependent Willingness to Divulge Sensitive Information. *Journal of Consumer Research*, *37*(5), 858–873. doi:10.1086/656423

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). New York: Guilford Press.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory--Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. In O. P. John, R. W. Robbins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). Guilford Press: New York.

Johnson, J. G., Cohen, P., Skodol, A. E., Oldham, J. M., Kasen, S., & Brook, J. S. (1999). Personality disorders in adolescence and risk of major mental disorders and suicidality during adulthood. *Arch Gen Psychiatry, 56*(9), 805-811. doi:10.1001/archpsyc.56.9.805

Johnson, T. R., & Bolt, D. M. (2010). On the Use of Factor-Analytic Multinomial Logit Item Response Models to Account for Individual Differences in Response Style. *Journal of Educational and Behavioral Statistics, 35*(1), 92-114. doi:10.3102/1076998609340529

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: consistent results from three test batteries. *Intelligence, 32*(1), 95-107. doi:10.1016/s0160-2896(03)00062-x

Johnson, W., Nijenhuis, J. t., & Bouchard, T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence, 36*(1), 81-95. doi:10.1016/j.intell.2007.06.001

Jones, E. E., Ghannam, J., Nigg, J. T., & Dyer, J. F. (1993). A paradigm for single-case research: The time series study of a long-term psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 61*(3), 381-394. doi:10.1037/0022-006x.61.3.381

Jones, H. J., Heron, J., Hammerton, G., Stochl, J., Jones, P. B., Cannon, M., . . . Me Research, T. (2018). Investigating the genetic architecture of general and specific psychopathology in adolescence. *Transl Psychiatry, 8*(1), 145. doi:10.1038/s41398-018-0204-9

Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification and 'race'. *Nat Genet, 36*(11 Suppl), S28-33. doi:10.1038/ng1435

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133. doi:10.1007/bf02291393

Jöreskog, K. G. (1999). *How large can a standardized coefficient be*. Stata FAQ. Available at http://www.statmodel.com/download/Joreskog.pdf.

Judd, L. L., Schettler, P. J., Coryell, W., Akiskal, H. S., & Fiedorowicz, J. G. (2013). Overt irritability/anger in unipolar major depressive episodes: past and current characteristics and implications for long-term course. *JAMA Psychiatry, 70*(11), 1171-1180. doi:10.1001/jamapsychiatry.2013.1957

Kasen, S., Cohen, P., Skodol, A. E., Johnson, J. G., & Brook, J. S. (1999). Influence of Child and Adolescent Psychiatric Disorders on Young Adult Personality Disorder. *American Journal of Psychiatry*, *156*(10), 1529–1535. doi:10.1176/ajp.156.10.1529

Keenan, K., Loeber, R., & Green, S. (1999). Conduct disorder in girls: A review of the literature. *Clinical Child and Family Psychology Review, 2*(1), 3-19.

Kelly, B. D., Nur, U. A., Tyrer, P., & Casey, P. (2009). Impact of severity of personality disorder on the outcome of depression. *Eur Psychiatry, 24*(5), 322-326. doi:10.1016/j.eurpsy.2008.12.004

Kessler, R. C. (1994). Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States. *Archives of General Psychiatry*, *51*(1), 8. doi:10.1001/archpsyc.1994.03950010008002

Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & Ustün, T. B. (2007). Age of onset of mental disorders: a review of recent literature. *Current Opinion in Psychiatry*, *20*(4), 359–364. doi:10.1097/yco.0b013e32816ebc8c

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., … Wang, P. S. (2003). The Epidemiology of Major Depressive Disorder. *JAMA*, *289*(23), 3095. doi:10.1001/jama.289.23.3095

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annu Rev Public Health, 34*, 119-138. doi:10.1146/annurev-publhealth-031912-114409

Kessler, R. C., Gruber, M., Hettema, J. M., Hwang, I., Sampson, N., & Yonkers, K. A. (2008). Co-morbid major depression and generalized anxiety disorders in the National Comorbidity Survey follow-up. *Psychol Med, 38*(3), 365-374. doi:10.1017/S0033291707002012

Kessler, R. C., Zhao, S., Blazer, D. G., & Swartz, M. (1997). Prevalence, correlates, and course of minor depression and major depression in the national comorbidity survey. *Journal of Affective Disorders, 45*(1-2), 19-30. doi:10.1016/s0165-0327(97)00056-6

Keyes, K. M., Eaton, N. R., Krueger, R. F., McLaughlin, K. A., Wall, M. M., Grant, B. F., & Hasin, D. S. (2012). Childhood maltreatment and the structure of common psychiatric disorders. *Br J Psychiatry, 200*(2), 107-115. doi:10.1192/bjp.bp.111.093062

Kim, H., & Eaton, N. R. (2015). The hierarchical structure of common mental disorders: Connecting multiple levels of comorbidity, bifactor models, and predictive validity. *J Abnorm Psychol, 124*(4), 1064-1078. doi:10.1037/abn0000113

Kim, H., & Eaton, N. R. (2017). A Hierarchical Integration of Person-Centered Comorbidity Models: Structure, Stability, and Transition Over Time. *Clinical Psychological Science, 5*(4), 595-612. doi:10.1177/2167702617704018

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, *98*(1), 191-207.

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med, 5*(2), e45. doi:10.1371/journal.pmed.0050045

Knowles, E. S., & Nathan, K. T. (1997). Acquiescent Responding in Self-Reports: Cognitive Style or Social Concern? *Journal of Research in Personality, 31*(2), 293-301. doi:10.1006/jrpe.1997.2180

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychol Methods, 23*(3), 505-523. doi:10.1037/met0000146

Kool, S., Schoevers, R., de Maat, S., Van, R., Molenaar, P., Vink, A., & Dekker, J. (2005). Efficacy of pharmacotherapy in depressed patients with and without personality disorders: a systematic review and meta-analysis. *J Affect Disord, 88*(3), 269-278. doi:10.1016/j.jad.2005.05.017

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., . . . Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *J Abnorm Psychol, 126*(4), 454-477. doi:10.1037/abn0000258

Kotov, R., Ruggero, C. J., Krueger, R. F., Watson, D., Yuan, Q., & Zimmerman, M. (2011). New dimensions in the quantitative classification of mental illness. *Arch Gen Psychiatry, 68*(10), 1003-1011. doi:10.1001/archgenpsychiatry.2011.107

Koziol, L. F., Budding, D., Andreasen, N., D'Arrigo, S., Bulgheroni, S., Imamizu, H., . . . Yamazaki, T. (2014). Consensus paper: the cerebellum's role in movement and cognition. *Cerebellum, 13*(1), 151-177. doi:10.1007/s12311-013-0511-x

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med, 16*(9), 606-613. doi:10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Spitzer, R. L., Williams, J. B., & Lowe, B. (2010). The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *Gen Hosp Psychiatry, 32*(4), 345-359. doi:10.1016/j.genhosppsych.2010.03.006

Krueger, R. F. (1999). The Structure of Common Mental Disorders. *Archives of General Psychiatry, 56*(10), 921. doi:10.1001/archpsyc.56.10.921

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychol Med, 42*(9), 1879-1890. doi:10.1017/S0033291711002674

Krueger, R. F., DeYoung, C. G., & Markon, K. E. (2010). Toward scientifically useful quantitative models of psychopathology: The importance of a comparative approach. *Behavioral and Brain Sciences*, *33*(2-3), 163–164. doi:10.1017/s0140525x10000646

Krueger, R. F., Eaton, N. R., Clark, L. A., Watson, D., Markon, K. E., Derringer, J., . . . Livesley, W. J. (2011). Deriving an empirical structure of personality pathology for DSM-5. *J Pers Disord, 25*(2), 170-191. doi:10.1521/pedi.2011.25.2.170

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., ... & Bagby, R. M. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*(3), 282-293.

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: a model-based approach to understanding and classifying psychopathology. *Annu Rev Clin Psychol, 2*, 111-133. doi:10.1146/annurev.clinpsy.2.022305.095213

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: an integrative quantitative model of the adult externalizing spectrum. *J Abnorm Psychol, 116*(4), 645-666. doi:10.1037/0021-843X.116.4.645

Kuo, E. S., Stoep, A. V., & Stewart, D. G. (2005). Using the short mood and feelings questionnaire to detect depression in detained adolescents. *Assessment, 12*(4), 374-383. doi: 10.1177/1073191105279984

Laceulle, O. M., Vollebergh, W. A. M., & Ormel, J. (2015). The Structure of Psychopathology in Adolescence. *Clinical Psychological Science, 3*(6), 850-860. doi:10.1177/2167702614560750

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood? *J Abnorm Psychol, 121*(4), 971-977. doi:10.1037/a0028355

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychol Bull, 143*(2), 142-186. doi:10.1037/bul0000069

Lahey, B. B., Rathouz, P. J., Keenan, K., Stepp, S. D., Loeber, R., & Hipwell, A. E. (2015). Criterion validity of the general factor of psychopathology in a prospective study of girls. *J Child Psychol Psychiatry, 56*(4), 415-422. doi:10.1111/jcpp.12300

Lahey, B. B., Van Hulle, C. A., Singh, A. L., Waldman, I. D., & Rathouz, P. J. (2011). Higher-order genetic and environmental structure of prevalent forms of child and adolescent psychopathology. *Arch Gen Psychiatry, 68*(2), 181-189. doi:10.1001/archgenpsychiatry.2010.192

Lahey, B. B., Zald, D. H., Perkins, S. F., Villalta-Gil, V., Werts, K. B., Van Hulle, C. A., . . . Waldman, I. D. (2017). Measuring the hierarchical general factor model of psychopathology in young adults. *Int J Methods Psychiatr Res, 27*(1). doi:10.1002/mpr.1593

Lamers, F., van Oppen, P., Comijs, H. C., Smit, J. H., Spinhoven, P., van Balkom, A. J., . . . Penninx, B. W. (2011). Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the Netherlands Study of Depression and Anxiety (NESDA). *J Clin Psychiatry, 72*(3), 341-348. doi:10.4088/JCP.10m06176blu

Laukaityte, I., & Wiberg, M. (2016). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics - Theory and Methods, 46*(22), 11341-11357. doi:10.1080/03610926.2016.1267764

Lechner, C. M., & Rammstedt, B. (2015). Cognitive ability, acquiescence, and the structure of personality in a sample of older adults. *Psychol Assess, 27*(4), 1301-1311. doi:10.1037/pas0000151

Letourneau, E. J., Henggeler, S. W., Borduin, C. M., Schewe, P. A., McCart, M. R., Chapman, J. E., & Saldana, L. (2009). Multisystemic therapy for juvenile sexual offenders: 1-year results from a randomized effectiveness trial. *J Fam Psychol, 23*(1), 89-102. doi:10.1037/a0014352

Levinson, C. A., Brosof, L. C., Vanzhula, I., Christian, C., Jones, P., Rodebaugh, T. L., . . . Fernandez, K. C. (2018). Social anxiety and eating disorder comorbidity and underlying vulnerabilities: Using network analysis to conceptualize comorbidity. *Int J Eat Disord, 51*(7), 693-709. doi:10.1002/eat.22890

Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: The role of implicit theories. *Journal of Personality and Social Psychology*, *74*(6), 1421–1436. doi:10.1037/0022-3514.74.6.1421

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766. doi:10.1016/j.jesp.2013.03.013

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behav Res Methods, 48*(3), 936-949. doi:10.3758/s13428-015-0619-7

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thogersen-Ntoumani, C. (2012). Method effects: the problem with negatively versus positively keyed items. *J Pers Assess, 94*(2), 196-204. doi:10.1080/00223891.2011.645936

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198-1202.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., ... & Neaton, J. D. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355-1360.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychol Methods, 22*(3), 486-506. doi:10.1037/met0000075

Livesley, J. (2015). A hypothesis too far? Commentary on personality dysfunction as the cause of recurrent non-cognitive mental disorder. *Personal Ment Health, 9*(1), 14-16. doi:10.1002/pmh.1285

Livesley, W. J. (2011). An empirically-based classification of personality disorder. *J Pers Disord, 25*(3), 397-420. doi:10.1521/pedi.2011.25.3.397

Livesley, W. J. (2012). Disorder in the proposed DSM-5 classification of personality disorders. *Clin Psychol Psychother, 19*(5), 364-368. doi:10.1002/cpp.1808

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Welsley Publishing Company.

Löwe, B., Kroenke, K., Herzog, W., & Gräfe, K. (2004). Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). *Journal of Affective Disorders, 81*(1), 61-66. doi:10.1016/s0165-0327(03)00198-8

Lundervold, A. J., Breivik, K., Posserud, M. B., Stormark, K. M., & Hysing, M. (2013). Symptoms of depression as reported by Norwegian adolescents on the Short Mood and Feelings Questionnaire. *Front Psychol, 4*, 613. doi: 10.3389/fpsyg.2013.00613

Lundh, L.G., Wangby-Lundh, M., & Bjarehed, J. (2008). Self reported emotional and behavioral problems in Swedish 14 to 15-year-old adolescents: A study with the self-report version of the Strengths and Difficulties Questionnaire. *Scandinavian Journal of Psychology*, *49*, 523–532.

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Sci Rep, 8*(1), 7819. doi:10.1038/s41598-018-25953-0

Lynn, M., & Harris, J. (1997). The desire for unique consumer products: A new individual differences scale. *Psychology & Marketing*, *14*(6), 601-616.

Lysaker, P. H., Gumley, A., & Dimaggio, G. (2011). Metacognitive disturbances in persons with severe mental illness: theory, correlates with psychopathology and models of psychotherapy. *Psychol Psychother, 84*(1), 1-8. doi:10.1111/j.2044-8341.2010.02007.x

MacDonald, K. B. (2008). Effortful control, explicit processing, and the regulation of human evolved predispositions. *Psychol Rev, 115*(4), 1012-1031. doi:10.1037/a0013327

Manea, L., Gilbody, S., & McMillan, D. (2015). A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen Hosp Psychiatry, 37*(1), 67-75. doi:10.1016/j.genhosppsych.2014.09.009

Manor, O., Matthews, S., & Power, C. (2000). Dichotomous or categorical response? Analysing self-rated health and lifetime social class. *Int J Epidemiol, 29*(1), 149-157. doi:10.1093/ije/29.1.149

Markon, K. E. (2019). Bifactor and Hierarchical Models: Specification, Inference, and Interpretation. *Annu Rev Clin Psychol, 15*, 51-69. doi:10.1146/annurev-clinpsy-050718-095522

Markon, K. E., & Jonas, K. G. (2016). Structure as cause and representation: Implications of descriptivist inference for structural modeling across multiple levels of analysis. *J Abnorm Psychol, 125*(8), 1146-1157. doi:10.1037/abn0000206

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. *J Pers Soc Psychol, 88*(1), 139-157. doi:10.1037/0022-3514.88.1.139

Marsh, H. W. (2016). Confirmatory Factor Analyses of Multitrait-Multimethod Data: Many Problems and a Few Solutions. *Applied Psychological Measurement, 13*(4), 335-361. doi:10.1177/014662168901300402

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32*(1), 151-170. doi:10.1016/j.cedpsych.2006.10.008

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: traits, ephemeral artifacts, and stable response styles. *Psychol Assess, 22*(2), 366-381. doi:10.1037/a0019225

Martel, M. M., Gremillion, M., Roberts, B., von Eye, A., & Nigg, J. T. (2010). The structure of childhood disruptive behaviors. *Psychol Assess, 22*(4), 816-826. doi:10.1037/a0020975

Martel, M. M., Pan, P. M., Hoffmann, M. S., Gadelha, A., do Rosario, M. C., Mari, J. J., . . . Salum, G. A. (2017). A general psychopathology factor (P factor) in children: Structural model analysis and external validation through familial risk and child global executive function. *J Abnorm Psychol, 126*(1), 137-148. doi:10.1037/abn0000205

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty Factors in Binary Data. *British Journal of Mathematical and Statistical Psychology, 27*(1), 82-99. doi:10.1111/j.2044-8317.1974.tb00530.x

McElroy, E., Belsky, J., Carragher, N., Fearon, P., & Patalay, P. (2018). Developmental stability of general and specific factors of psychopathology from early childhood to adolescence: dynamic mutualism or p-differentiation? *J Child Psychol Psychiatry, 59*(6), 667-675. doi:10.1111/jcpp.12849

McElroy, E., Shevlin, M., & Murphy, J. (2017). Internalizing and externalizing disorders in childhood and adolescence: A latent transition analysis using ALSPAC data. *Compr Psychiatry, 75*, 75-84. doi:10.1016/j.comppsych.2017.03.003

McEvoy, P. M., Nathan, P., & Norton, P. J. (2009). Efficacy of Transdiagnostic Treatments: A Review of Published Outcome Studies and Future Research Directions. *Journal of Cognitive Psychotherapy, 23*(1), 20-33. doi:10.1891/0889-8391.23.1.20

McNally, R. J., Mair, P., Mugno, B. L., & Riemann, B. C. (2017). Co-morbid obsessive-compulsive disorder and depression: a Bayesian network approach. *Psychol Med, 47*(7), 1204-1214. doi:10.1017/S0033291716003287

Meehl, P. E. & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology*, *30*, 525-564.

Meier, M. A., & Meier, M. H. (2018). Clinical implications of a general psychopathology factor: A cognitive–behavioral transdiagnostic group treatment for community mental health. *Journal of Psychotherapy Integration, 28*(3), 253-268. doi:10.1037/int0000095

Melchers, M., Plieger, T., Montag, C., Reuter, M., Spinath, F. M., & Hahn, E. (2018). The heritability of response styles and its impact on heritability estimates of personality: A twin study. *Personality and Individual Differences, 134*, 16-24. doi:10.1016/j.paid.2018.05.023

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychol Methods, 21*(2), 151-163. doi:10.1037/met0000038

Messick, S. (1991). Psychology and Methodology of Response Styles. In R. E. Snow and D. Wiley (Eds.), *Improving Inquiry in Social Science. A Volume in Honor of Lee J. Cronbach* (pp. 177-216). New York: Routledge.

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016). Method Effects on an Adaptation of the Rosenberg Self-Esteem Scale in Greek and the Role of Personality Traits. *J Pers Assess, 98*(2), 178-188. doi:10.1080/00223891.2015.1089248

Middleton, F. A., & Strick, P. L. (2001). Cerebellar Projections to the Prefrontal Cortex of the Primate. *The Journal of Neuroscience, 21*(2), 700-712. doi:10.1523/jneurosci.21-02-00700.2001

Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personal Disord, 8*(1), 26-34. doi:10.1037/per0000191

Miller, M., Iosif, A. M., Young, G. S., Bell, L. J., Schwichtenberg, A. J., Hutman, T., & Ozonoff, S. (2019). The dysregulation profile in preschoolers with and without a family history of autism spectrum disorder. *J Child Psychol Psychiatry, 60*(5), 516-523. doi:10.1111/jcpp.13003

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), 246-268. doi:10.1037/0033-295x.102.2.246

Mislevy R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika. 56*(2), 177-196.

Miyake, A., & Friedman, N. P. (2012). The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Curr Dir Psychol Sci, 21*(1), 8-14. doi:10.1177/0963721411429458

Moberget, T., Alnaes, D., Kaufmann, T., Doan, N. T., Cordova-Palomera, A., Norbom, L. B., . . . Westlye, L. T. (2019). Cerebellar Gray Matter Volume Is Associated With Cognitive Function and Psychopathology in Adolescence. *Biol Psychiatry, 86*(1), 65-75. doi:10.1016/j.biopsych.2019.01.019

Moffitt, T. E., Caspi, A., Taylor, A., Kokaua, J., Milne, B. J., Polanczyk, G., & Poulton, R. (2010). How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychol Med, 40*(6), 899-909. doi:10.1017/S0033291709991036

Moffitt, T. E., Harrington, H., Caspi, A., Kim-Cohen, J., Goldberg, D., Gregory, A. M., & Poulton, R. (2007). Depression and generalized anxiety disorder: cumulative and sequential comorbidity in a birth cohort followed prospectively to age 32 years. *Arch Gen Psychiatry, 64*(6), 651-660. doi:10.1001/archpsyc.64.6.651

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research & Perspective, 2*(4), 201-218. doi:10.1207/s15366359mea0204_1

Molenaar, P. C. M., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 226-242). Thousand Oaks, CA, US: Sage Publications, Inc.

Monroe, S. M., Anderson, S. F., & Harkness, K. L. (2019). Life stress and major depression: The mysteries of recurrences. *Psychol Rev*. doi:10.1037/rev0000157

Montag, C., & Reuter, M. (2008). Does speed in completing an online questionnaire have an influence on its reliability? *Cyberpsychol Behav, 11*(6), 719-721. doi:10.1089/cpb.2007.0258

Moradveisi, L., Huibers, M. J., Renner, F., Arasteh, M., & Arntz, A. (2013). The influence of comorbid personality disorder on the effects of behavioural activation vs. antidepressant medication for major depressive disorder: results from a randomized trial in Iran. *Behav Res Ther, 51*(8), 499-506. doi:10.1016/j.brat.2013.05.006

Morey, L. C. (2017). Development and initial evaluation of a self-report form of the DSM-5 Level of Personality Functioning Scale. *Psychol Assess, 29*(10), 1302-1308. doi:10.1037/pas0000450

Morey, L. C., & Benson, K. T. (2016). Relating DSM-5 section II and section III personality disorder diagnostic classification systems to treatment planning. *Compr Psychiatry, 68*, 48-55. doi:10.1016/j.comppsych.2016.03.010

Morey, L. C., Benson, K. T., Busch, A. J., & Skodol, A. E. (2015). Personality disorders in DSM-5: emerging research on the alternative model. *Curr Psychiatry Rep, 17*(4), 558. doi:10.1007/s11920-015-0558-0

Morey, L. C., Benson, K. T., & Skodol, A. E. (2016). Relating DSM-5 section III personality traits to section II personality disorder diagnoses. *Psychol Med, 46*(3), 647-655. doi:10.1017/S0033291715002226

Morey, L. C., & Hopwood, C. J. (2013). Stability and change in personality disorders. *Annu Rev Clin Psychol, 9*, 499-528. doi:10.1146/annurev-clinpsy-050212-185637

Morey, L. C., Hopwood, C. J., Markowitz, J. C., Gunderson, J. G., Grilo, C. M., McGlashan, T. H., . . . Skodol, A. E. (2012). Comparison of alternative models for personality disorders, II: 6-, 8- and 10-year follow-up. *Psychol Med, 42*(8), 1705-1713. doi:10.1017/S0033291711002601

Morey, L. C., Skodol, A. E., & Oldham, J. M. (2014). Clinician judgments of clinical utility: A comparison of DSM-IV-TR personality disorders and the alternative model for DSM-5 personality disorders. *J Abnorm Psychol, 123*(2), 398-405. doi:10.1037/a0036481

Morgan, G., Hodge, K., Wells, K., & Watkins, M. (2015). Are Fit Indices Biased in Favor of Bi-Factor Models in Cognitive Ability Research?: A Comparison of Fit in Correlated Factors, Higher-Order, and Bi-Factor Models via Monte Carlo Simulations. *Journal of Intelligence, 3*(1), 2-20. doi:10.3390/jintelligence3010002

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2015). A Bifactor Exploratory Structural Equation Modeling Framework for the Identification of Distinct Sources of Construct-Relevant Psychometric Multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(1), 116-139. doi:10.1080/10705511.2014.961800

Mottus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., . . . Johnson, W. (2012). The effect of response style on self-reported Conscientiousness across

20 countries. *Pers Soc Psychol Bull, 38*(11), 1423-1436. doi:10.1177/0146167212451275

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling: A Multidisciplinary Journal, 4*(3), 193-211. doi:10.1080/10705519709540071

Mulder, R., Murray, G., & Rucklidge, J. (2017). Common versus specific factors in psychotherapy: opening the black box. *The Lancet Psychiatry, 4*(12), 953-962. doi:10.1016/s2215-0366(17)30100-1

Mulder, R. T. (2002). Personality pathology and treatment outcome in major depression: a review. *Am J Psychiatry, 159*(3), 359-371. doi:10.1176/appi.ajp.159.3.359

Muris, P., Meesters, C., & van den Berg, F. (2003). The Strengths and Difficulties Questionnaire (SDQ): Further evidence for its reliability and validity in a community sample of Dutch children and adolescents. *European Child and Adolescent Psychiatry*, *12*(1), 1–8.

Murphy, J. M., Horton, N. J., Laird, N. M., Monson, R. R., Sobol, A. M., & Leighton, A. H. (2004). Anxiety and depression: a 40-year perspective on relationships regarding prevalence, distribution, and comorbidity. *Acta Psychiatr Scand, 109*(5), 355-375. doi:10.1111/j.1600-0447.2003.00286.x

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The Development of the General Factor of Psychopathology 'p Factor' Through Childhood and Adolescence. *J Abnorm Child Psychol, 44*(8), 1573-1586. doi:10.1007/s10802-016-0132-1

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence, 41*(5), 407-422. doi:10.1016/j.intell.2013.06.004

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality, 41*(6), 1213-1233. doi:10.1016/j.jrp.2007.02.003

Muthén, B. O. (1991). Multilevel Factor Analysis of Class and Student Achievement Components. *Journal of Educational Measurement, 28*(4), 338-354. doi:10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological methods & research*, *22*(3), 376-398.

Muthén L. K, & Muthén B. O. (2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, California: Muthén & Muthén.

Nagelkerke, N. J. D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika, 78*(3), 691. doi:10.1093/biomet/78.3.691

Nakagawa, S., Schielzeth, H., & O'Hara, R. B. (2013). A general and simple method for obtainingR2from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133-142. doi:10.1111/j.2041-210x.2012.00261.x

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the Prevalence of Problematic Respondent Behaviors among MTurk, Campus, and Community Participants. *PLoS One, 11*(6), e0157732. doi:10.1371/journal.pone.0157732

Nelson, C. A., Fox, N. A, & Zeanah, C. H. (2014, December 2). *Forgotten Children. What Romania Can Tell Us About Institutional Care.* Retrieved from: https://www.foreignaffairs.com/articles/romania/2014-12-02/forgotten-children

Nesse, R. M., & Stein, D. J. (2012). Towards a genuinely medical model for psychiatric nosology. *BMC Med, 10*, 5. doi:10.1186/1741-7015-10-5

Neumann, A., Pappa, I., Lahey, B. B., Verhulst, F. C., Medina-Gomez, C., Jaddoe, V. W., . . . Tiemeier, H. (2016). Single Nucleotide Polymorphism Heritability of a General Psychopathology Factor in Children. *J Am Acad Child Adolesc Psychiatry, 55*(12), 1038-1045 e1034. doi:10.1016/j.jaac.2016.09.498

Newby, J. M., McKinnon, A., Kuyken, W., Gilbody, S., & Dalgleish, T. (2015). Systematic review and meta-analysis of transdiagnostic psychological treatments for anxiety and depressive disorders in adulthood. *Clin Psychol Rev, 40*, 91-110. doi:10.1016/j.cpr.2015.06.002

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev, 75*(1), 14-49. doi:10.1128/MMBR.00028-10

Newton-Howes, G., Tyrer, P., & Johnson, T. (2006). Personality disorder and the outcome of depression: meta-analysis of published studies. *Br J Psychiatry, 188*, 13-20. doi:10.1192/bjp.188.1.13

Newton-Howes, G., Tyrer, P., Johnson, T., Mulder, R., Kool, S., Dekker, J., & Schoevers, R. (2014). Influence of personality on the outcome of treatment in depression: systematic review and meta-analysis. *J Pers Disord, 28*(4), 577-593. doi:10.1521/pedi_2013_27_070

Niarchou, M., Moore, T. M., Tang, S. X., Calkins, M. E., McDonald-McGuinn, D. M., Zackai, E. H., . . . Gur, R. E. (2017). The dimensional structure of psychopathology in 22q11.2 Deletion Syndrome. *J Psychiatr Res, 92*, 124-131. doi:10.1016/j.jpsychires.2017.04.006

Noordhof, A., Krueger, R. F., Ormel, J., Oldehinkel, A. J., & Hartman, C. A. (2015). Integrating autism-related symptoms into the dimensional internalizing and externalizing model of psychopathology. The TRAILS Study. *J Abnorm Child Psychol, 43*(3), 577-587. doi:10.1007/s10802-014-9923-4

Norton, P. J., & Paulus, D. J. (2016). Toward a Unified Treatment for Emotional Disorders: Update on the Science and Practice. *Behav Ther, 47*(6), 854-868. doi:10.1016/j.beth.2015.07.002

Ochsner, K. N., Silvers, J. A., & Buhle, J. T. (2012). Functional imaging studies of emotion regulation: a synthetic review and evolving model of the cognitive control of emotion. *Ann N Y Acad Sci, 1251*, E1-24. doi:10.1111/j.1749-6632.2012.06751.x

Ogden, T., & Hagen, K. A. (2006). Multisystemic Treatment of Serious Behaviour Problems in Youth: Sustainability of Effectiveness Two Years after Intake. *Child and Adolescent Mental Health, 11*(3), 142-149. doi:10.1111/j.1475-3588.2006.00396.x

Oldham, J. M. (2015). The alternative DSM-5 model for personality disorders. *World Psychiatry, 14*(2), 234-236. doi:10.1002/wps.20232

Oldham, J. M. (2018). DSM models of personality disorders. *Curr Opin Psychol, 21*, 86-88. doi:10.1016/j.copsyc.2017.09.010

Oldham, J. M., Skodol, A. E., Kellman, H. D., Hyler, S. E., Rosnick, L., & Davies, M. (1992). Diagnosis of DSM-III-R personality disorders by two structured interviews: patterns of comorbidity. *Am J Psychiatry, 149*(2), 213-220. doi:10.1176/ajp.149.2.213

Olino, T. M., Bufferd, S. J., Dougherty, L. R., Dyson, M. W., Carlson, G. A., & Klein, D. N. (2018). The Development of Latent Dimensions of Psychopathology across Early Childhood: Stability of Dimensions and Moderators of Change. *J Abnorm Child Psychol, 46*(7), 1373-1383. doi:10.1007/s10802-018-0398-6

Olino, T. M., Dougherty, L. R., Bufferd, S. J., Carlson, G. A., & Klein, D. N. (2014). Testing models of psychopathology in preschool-aged children using a structured interview-based assessment. *J Abnorm Child Psychol, 42*(7), 1201-1211. doi:10.1007/s10802-014-9865-x

Olkin, I., & Sampson, A. R. (2001). Multivariate Analysis: Overview. *International Encyclopedia of the Social & Behavioral Sciences*, 10240–10247. doi:10.1016/b0-08-043076-7/00472-1

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1-18. doi:10.1348/014466506x96931

Papakostas, G. I., & Fava, M. (2009). Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. *Eur Neuropsychopharmacol, 19*(1), 34-40. doi:10.1016/j.euroneuro.2008.08.009

Parasuraman, A. (2000). Technology Readiness Index (TRI). *Journal of Service Research*, *2*(4), 307–320. doi:10.1177/109467050024001

Pardini, D. A., & Fite, P. J. (2010). Symptoms of Conduct Disorder, Oppositional Defiant Disorder, Attention-Deficit/Hyperactivity Disorder, and Callous-Unemotional Traits as Unique Predictors of Psychosocial Maladjustment in Boys. *Journal of the American Academy of Child & Adolescent Psychiatry, 49*(11), 1134-1144. doi:10.1097/00004583-201011000-00007

Patalay, P., Fonagy, P., Deighton, J., Belsky, J., Vostanis, P., & Wolpert, M. (2015). A general psychopathology factor in early adolescence. *Br J Psychiatry, 207*(1), 15-22. doi:10.1192/bjp.bp.114.149591

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological* *attitudes, Vol. 1. Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA, US: Academic Press.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153-163. doi:10.1016/j.jesp.2017.01.006

Perry, J. C. (1988). A Prospective Study of Life Stress, Defenses, Psychotic Symptoms, and Depression in Borderline and Antisocial Personality Disorders and Bipolar Type II Affective Disorder. *Journal of Personality Disorders, 2*(1), 49-59. doi:10.1521/pedi.1988.2.1.49

Pettersson, E., Lahey, B. B., Larsson, H., & Lichtenstein, P. (2018). Criterion Validity and Utility of the General Factor of Psychopathology in Childhood: Predictive Associations With Independently Measured Severe Adverse Mental Health Outcomes in Adolescence. *J Am Acad Child Adolesc Psychiatry, 57*(6), 372-383. doi:10.1016/j.jaac.2017.12.016

Pettersson, E., Larsson, H., & Lichtenstein, P. (2016). Common psychiatric disorders share the same genetic origin: a multivariate sibling study of the Swedish population. *Mol Psychiatry, 21*(5), 717-721. doi:10.1038/mp.2015.116

Pezzoli, P., Antfolk, J., & Santtila, P. (2017). Phenotypic factor analysis of psychopathology reveals a new body-related transdiagnostic factor. *PLoS One, 12*(5), e0177674. doi:10.1371/journal.pone.0177674

Phillips, D. L., & Clancy, K. J. (1970). Response biases in field studies of mental illness. *American Sociological Review*, 503-515.

Phillips, J. R., Hewedi, D. H., Eissa, A. M., & Moustafa, A. A. (2015). The cerebellum and psychiatric disorders. *Front Public Health, 3*, 66. doi:10.3389/fpubh.2015.00066

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control, C., & Davis, O. S. (2013). Common DNA markers can account for more than half of the genetic influence on cognitive abilities. *Psychol Sci, 24*(4), 562-568. doi:10.1177/0956797612457952

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J Appl Psychol, 88*(5), 879-903. doi:10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annu Rev Psychol, 63*, 539-569. doi:10.1146/annurev-psych-120710-100452

Poldrack, R. A. (2010). Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed? *Perspect Psychol Sci, 5*(6), 753-761. doi:10.1177/1745691610388777

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychol Methods, 17*(1), 1-14. doi:10.1037/a0026804

Preti, A., Carta, M. G., & Petretto, D. R. (2019). Factor structure models of the SCL-90-R: Replicability across community samples of adolescents. *Psychiatry Res, 272*, 491-498. doi:10.1016/j.psychres.2018.12.146

Quilty, L. C., Ayearst, L., Chmielewski, M., Pollock, B. G., & Bagby, R. M. (2013). The psychometric properties of the personality inventory for DSM-5 in an APA DSM-5 field trial sample. *Assessment, 20*(3), 362-369. doi:10.1177/1073191113486183

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale Method Effects. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(1), 99-117. doi:10.1207/s15328007sem1301_5

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111. doi:10.2307/271063

Raines, A. M., Boffa, J. W., Allan, N. P., Short, N. A., & Schmidt, N. B. (2015). Hoarding and eating pathology: the mediating role of emotion regulation. *Compr Psychiatry, 57*, 29-35. doi:10.1016/j.comppsych.2014.11.005

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*(1), 203-212. doi:10.1016/j.jrp.2006.02.001

Reed, G. M. (2018). Progress in developing a classification of personality disorders for ICD-11. *World Psychiatry, 17*(2), 227-229. doi:10.1002/wps.20533

Reise, S., Moore, T., & Maydeu-Olivares, A. (2011). Target Rotations and Assessing the Impact of Model Violations on the Parameters of Unidimensional Item Response Theory Models. *Educational and Psychological Measurement, 71*(4), 684-711. doi:10.1177/0013164410378690

Reise, S. P. (2012). Invited Paper: The Rediscovery of Bifactor Measurement Models. *Multivariate Behav Res, 47*(5), 667-696. doi:10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *J Pers Assess, 95*(2), 129-140. doi:10.1080/00223891.2012.725437

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the Bifactor Model a Better Model or Is It Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-

Esteem Scale. *Multivariate Behav Res, 51*(6), 818-838. doi:10.1080/00273171.2016.1243461

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J Pers Assess, 92*(6), 544-559. doi:10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling. *Educational and Psychological Measurement, 73*(1), 5-26. doi:10.1177/0013164412449831

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *J Pers Assess, 84*(2), 126-136. doi:10.1207/s15327752jpa8402_02

Remster, B. (2013). Self-Control and the Depression–Delinquency Link. *Deviant Behavior, 35*(1), 66-84. doi:10.1080/01639625.2013.822226

Rick, S. I., Small, D. A., & Finkel, E. J. (2011). Fatal (Fiscal) Attraction: Spendthrifts and Tightwads in Marriage. *Journal of Marketing Research*, *48*(2), 228–237. doi:10.1509/jmkr.48.2.228

Riem, M. M. E., van Hoof, M. J., Garrett, A. S., Rombouts, S., van der Wee, N. J. A., van, I. M. H., & Vermeiren, R. (2019). General psychopathology factor and unresolved-disorganized attachment uniquely correlated to white matter integrity using diffusion tensor imaging. *Behav Brain Res, 359*, 1-8. doi:10.1016/j.bbr.2018.10.014

Rigdon, E. (2015, August 17). Re: Bias in Weighted Least Squares Estimates [Google Groups]. Retrieved from https://groups.google.com/forum/#!topic/lavaan/fMoIk9Dl8gw

Rindskopf, D., & Rose, T. (1988). Some Theory and Applications of Confirmatory Second-Order Factor Analysis. *Multivariate Behav Res, 23*(1), 51-67. doi:10.1207/s15327906mbr2301_3

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2016). Evaluating the Impact of Careless Responding on Aggregated-Scores: To Filter Unmotivated Examinees or Not? *International Journal of Testing, 17*(1), 74-104. doi:10.1080/15305058.2016.1231193

Riosa, P. B., McArthur, B. A., & Preyde, M. (2011). Effectiveness of psychosocial intervention for children and adolescents with comorbid problems: a systematic review. *Child and Adolescent Mental Health, 16*(4), 177-185. doi:10.1111/j.1475-3588.2011.00609.x

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*(1), 3-25. doi:10.1037/0033-2909.126.1.3

Roberts, R. E., Forthofer, R. N., & Fabrega Jr, H. (1976). The Langner items and acquiescence. *Social Science & Medicine (1967), 10*(2), 69-75.

Robins, L. N. & Regier, D. A. (1991). *Psychiatric Disorders in America Disorders in America*. New York: Free Press.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *J Pers Assess, 98*(3), 223-237. doi:10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychol Methods, 21*(2), 137-150. doi:10.1037/met0000045

Rodriguez-Seijas, C., Eaton, N. R., Stohl, M., Mauro, P. M., & Hasin, D. S. (2017). Mental disorder comorbidity and treatment utilization. *Compr Psychiatry, 79*, 89-97. doi:10.1016/j.comppsych.2017.02.003

Rogers, J. H., Widiger, T. A., & Krupp, A. (1995). Aspects of depression associated with borderline personality disorder. *Am J Psychiatry, 152*(2), 268-270. doi:10.1176/ajp.152.2.268

Rohde, P., Lewinsohn, P. M., & Seeley, J. R. (1991). Comorbidity of unipolar depression: II. Comorbidity with other mental disorders in adolescents and adults. *Journal of Abnormal Psychology, 100*(2), 214-222.

Roiser, J. (2015). What has neuroscience ever done for us? *Psychologist*, *28*(4), 284-287.

Romer, A. L., Knodt, A. R., Houts, R., Brigidi, B. D., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2018). Structural alterations within cerebellar circuitry are associated with general liability for common mental disorders. *Mol Psychiatry, 23*(4), 1084-1090. doi:10.1038/mp.2017.57

Rose, P., & Segrist, D. J. (2012). Difficulty Identifying Feelings, Distress Tolerance and Compulsive Buying: Analyzing the Associations to Inform Therapeutic Strategies. *International Journal of Mental Health and Addiction, 10*(6), 927-935. doi:10.1007/s11469-012-9389-y

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rosenstrom, T., Gjerde, L. C., Krueger, R. F., Aggen, S. H., Czajkowski, N. O., Gillespie, N. A., . . . Ystrom, E. (2018). Joint factorial structure of psychopathology and personality. *Psychol Med*, 1-10. doi:10.1017/S0033291718002982

Rothbart, M. K., Ellis, L. K., Rueda, M. R., & Posner, M. I. (2003). Developing mechanisms of temperamental effortful control. *Journal of Personality*, *71*, 1113–1143. doi:10.1111/1467-6494.7106009

Rowland, M. D., Halliday-Boykins, C. A., Henggeler, S. W., Cunningham, P. B., Lee, T. G., Kruesi, M. J. P., & Shapiro, S. B. (2016). A Randomized Trial of Multisystemic Therapy With Hawaii's Felix Class Youths. *Journal of Emotional and Behavioral Disorders, 13*(1), 13-23. doi:10.1177/10634266050130010201

Rozeboom, W. W. (1982). The determinacy of common factors in large item domains. *Psychometrika, 47*(3), 281-295. doi:10.1007/bf02294160

Ruzzano, L., Borsboom, D., & Geurts, H. M. (2015). Repetitive behaviors in autism and obsessive-compulsive disorder: new perspectives from a network analysis. *J Autism Dev Disord, 45*(1), 192-202. doi:10.1007/s10803-014-2204-9

Rydell, A.-M., Berlin, L., & Bohlin, G. (2003). Emotionality, emotion regulation, and adaptation among 5- to 8-year-old children. *Emotion, 3*(1), 30-47. doi:10.1037/1528-3542.3.1.30

Rytila-Manninen, M., Frojd, S., Haravuori, H., Lindberg, N., Marttunen, M., Kettunen, K., & Therman, S. (2016). Psychometric properties of the Symptom Checklist-90 in adolescent psychiatric inpatients and age- and gender-matched community youth. *Child Adolesc Psychiatry Ment Health, 10*, 23. doi:10.1186/s13034-016-0111-x

Sallis, H., Szekely, E., Neumann, A., Jolicoeur-Martineau, A., van, I. M., Hillegers, M., . . . Evans, J. (2019). General psychopathology, internalising and externalising in children and functional outcomes in late adolescence. *J Child Psychol Psychiatry*. doi:10.1111/jcpp.13067

Sasso, K. E., & Strunk, D. R. (2013). Thin slice ratings of client characteristics in intake assessments: predicting symptom change and dropout in cognitive therapy for depression. *Behav Res Ther, 51*(8), 443-450. doi:10.1016/j.brat.2013.04.007

Sato, J. R., Salum, G. A., Gadelha, A., Crossley, N., Vieira, G., Manfro, G. G., . . . Bressan, R. A. (2016). Default mode network maturation and psychopathology in children and adolescents. *J Child Psychol Psychiatry, 57*(1), 55-64. doi:10.1111/jcpp.12444

Savalei, V., & Falk, C. F. (2014). Recovering Substantive Factor Loadings in the Presence of Acquiescence Bias: A Comparison of Three Approaches. *Multivariate Behav Res, 49*(5), 407-424. doi:10.1080/00273171.2014.931800

Sawyer, S. M., Afifi, R. A., Bearinger, L. H., Blakemore, S.-J., Dick, B., Ezeh, A. C., & Patton, G. C. (2012). Adolescence: a foundation for future health. *The Lancet, 379*(9826), 1630-1640. doi:10.1016/s0140-6736(12)60072-5

Sayyareh, A., Obeidi, R., & Bar-Hen, A. (2010). Empiricial Comparison between Some Model Selection Criteria. *Communications in Statistics - Simulation and Computation, 40*(1), 72-86. doi:10.1080/03610918.2010.530367

Schaefer, J. D., Moffitt, T. E., Arseneault, L., Danese, A., Fisher, H. L., Houts, R., . . . Caspi, A. (2018). Adolescent Victimization and Early-Adult Psychopathology: Approaching Causal Inference Using a Longitudinal Twin Study to Rule Out Noncausal Explanations. *Clin Psychol Sci, 6*(3), 352-371. doi:10.1177/2167702617741381

Schalke, D., Brunner, M., Geiser, C., Preckel, F., Keller, U., Spengler, M., & Martin, R. (2013). Stability and change in intelligence from age 12 to age 52: results from the Luxembourg MAGRIP study. *Dev Psychol, 49*(8), 1529-1543. doi:10.1037/a0030623

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*(1), 53-61. doi:10.1007/bf02289209

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: exploring the universal and culture-specific features of global self-esteem. *J Pers Soc Psychol, 89*(4), 623-642. doi:10.1037/0022-3514.89.4.623

Sellbom, M., Smid, W., de Saeger, H., Smit, N., & Kamphuis, J. H. (2014). Mapping the Personality Psychopathology Five domains onto DSM-IV personality disorders in Dutch clinical and forensic samples: implications for DSM-5. *J Pers Assess, 96*(2), 185-191. doi:10.1080/00223891.2013.825625

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychol Assess*. doi:10.1037/pas0000623

Selzam, S., Coleman, J. R. I., Caspi, A., Moffitt, T. E., & Plomin, R. (2018). A polygenic p factor for major psychiatric disorders. *Transl Psychiatry, 8*(1), 205. doi:10.1038/s41398-018-0217-4

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science, 1*(2), 213-220. doi:10.1177/2167702612469015

Sharma, P. (2009). Measuring personal cultural orientations: scale development and validation. *Journal of the Academy of Marketing Science*, *38*(6), 787–806. doi:10.1007/s11747-009-0184-7

Sharp, C., Goodyer, I. M., & Croudace, T. J. (2006). The Short Mood and Feelings Questionnaire (SMFQ): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *J Abnorm Child Psychol, 34*(3), 379-391. doi: 10.1007/s10802-006-9027-x

Sharp, C., Wright, A. G., Fowler, J. C., Frueh, B. C., Allen, J. G., Oldham, J., & Clark, L. A. (2015). The structure of personality pathology: Both general ('g') and specific ('s') factors? *J Abnorm Psychol, 124*(2), 387-398. doi:10.1037/abn0000033

Shea, M. T., Pilkonis, P. A., Beckham, E., Collins, J. F., Elkin, I., Sotsky, S. M., & Docherty, J. P. (1990). Personality disorders and treatment outcome in the NIMH Treatment of Depression Collaborative Research Program. *Am J Psychiatry, 147*(6), 711-718. doi:10.1176/ajp.147.6.711

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Med Res Methodol, 6*, 57. doi:10.1186/1471-2288-6-57

Simms, L. J., Gros, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depress Anxiety, 25*(7), E34-46. doi:10.1002/da.20432

Simon, H. A. (1973). The organization of complex systems. In H. H. Pattee (Ed.), *Hierarchy Theory: The challenge of complex systems* (pp. 1-27). New York, NY: George Braziller.

Skodol, A. E. (2012). Personality disorders in DSM-5. *Annu Rev Clin Psychol, 8*, 317-344. doi:10.1146/annurev-clinpsy-032511-143131

Skodol, A. E., Clark, L. A., Bender, D. S., Krueger, R. F., Morey, L. C., Verheul, R., . . . Oldham, J. M. (2011). Proposed changes in personality and personality disorder assessment and diagnosis for DSM-5 Part I: Description and rationale. *Personal Disord, 2*(1), 4-22. doi:10.1037/a0021891

Skodol, A. E., Grilo, C. M., Keyes, K. M., Geier, T., Grant, B. F., & Hasin, D. S. (2011). Relationship of personality disorders to the course of major depressive disorder in a nationally representative sample. *Am J Psychiatry, 168*(3), 257-264. doi:10.1176/appi.ajp.2010.10050695

Skodol, A. E., Morey, L. C., Bender, D. S., & Oldham, J. M. (2013). The ironic fate of the personality disorders in DSM-5. *Personal Disord, 4*(4), 342-349. doi:10.1037/per0000029

Skodol, A. E., Morey, L. C., Bender, D. S., & Oldham, J. M. (2015). The Alternative DSM-5 Model for Personality Disorders: A Clinical Application. *Am J Psychiatry, 172*(7), 606-613. doi:10.1176/appi.ajp.2015.14101220

Snyder, H. R., Gulley, L. D., Bijttebier, P., Hartman, C. A., Oldehinkel, A. J., Mezulis, A., . . . Hankin, B. L. (2015). Adolescent emotionality and effortful control: Core latent constructs and links to psychopathology and functioning. *J Pers Soc Psychol, 109*(6), 1132-1149. doi:10.1037/pspp0000047

Snyder, H. R., Hankin, B. L., Sandman, C. A., Head, K., & Davis, E. P. (2017). Distinct patterns of reduced prefrontal and limbic grey matter volume in childhood general and internalizing psychopathology. *Clin Psychol Sci, 5*(6), 1001-1013. doi:10.1177/2167702617714563

Snyder, H. R., Young, J. F., & Hankin, B. L. (2017). Strong Homotypic Continuity in Common Psychopathology-, Internalizing-, and Externalizing-Specific Factors Over Time in Adolescents. *Clin Psychol Sci, 5*(1), 98-110. doi:10.1177/2167702616651076

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality, 43*(1), 84-90. doi:10.1016/j.jrp.2008.10.002

Spearman, C. (1904). “General Intelligence,” Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201. doi:10.2307/1412107

Spearman, C. (1927). *The abilities of man*. Oxford, England: Macmillan.

Spinhoven, P., Elzinga, B. M., Hovens, J. G., Roelofs, K., van Oppen, P., Zitman, F. G., & Penninx, B. W. (2011). Positive and negative life events and personality traits in predicting course of depression and anxiety. *Acta Psychiatr Scand, 124*(6), 462-473. doi:10.1111/j.1600-0447.2011.01753.x

St Clair, M. C., Neufeld, S., Jones, P. B., Fonagy, P., Bullmore, E. T., Dolan, R. J., . . . Goodyer, I. M. (2017). Characterising the latent structure and organisation of self-reported thoughts, feelings and behaviours in adolescents and young adults. *PLoS One, 12*(4), e0175381. doi:10.1371/journal.pone.0175381

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245-251. doi:10.1037/0033-2909.87.2.245

Steinley, D., Hoffman, M., Brusco, M. J., & Sher, K. J. (2017). A method for making inferences in network analysis: Comment on Forbes, Wright, Markon, and Krueger (2017). *J Abnorm Psychol, 126*(7), 1000-1010. doi:10.1037/abn0000308

Stevens J. P. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing Samples in Cognitive Science. *Trends Cogn Sci, 21*(10), 736-748. doi:10.1016/j.tics.2017.06.007

Stochl, J., Khandaker, G. M., Lewis, G., Perez, J., Goodyer, I. M., Zammit, S., . . . Jones, P. B. (2015). Mood, anxiety and psychotic phenomena measure a common psychopathological factor. *Psychol Med, 45*(7), 1483-1493. doi:10.1017/S003329171400261X

Stone, A. A., Schneider, S., Junghaenel, D. U., & Broderick, J. E. (2019). Response styles confound the age gradient of four health and well-being outcomes. *Soc Sci Res, 78*, 215-225. doi:10.1016/j.ssresearch.2018.12.004

Strandholm, T., Karlsson, L., Kiviruusu, O., Pelkonen, M., & Marttunen, M. (2013). Treatment Characteristics and Outcome of Depression Among Depressed Adolescent Outpatients With and Without Comorbid Axis II Disorders. *J Pers Disord*. doi:10.1521/pedi_2013_27_073

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess, 80*(1), 99-103. doi:10.1207/S15327752JPA8001_18

Strickland, C. M., Hopwood, C. J., Bornovalova, M. A., Rojas, E. C., Krueger, R. F., & Patrick, C. J. (2019). Categorical and Dimensional Conceptions of Personality Pathology in DSM-5: Toward a Model-Based Synthesis. *J Pers Disord, 33*(2), 185-213. doi:10.1521/pedi_2018_32_339

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Multivariate applications series. Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183-206). New York, NY, US: Routledge/Taylor & Francis Group.

Sundell, K., Hansson, K., Lofholm, C. A., Olsson, T., Gustle, L. H., & Kadesjo, C. (2008). The transportability of multisystemic therapy to Sweden: short-term results from a randomized trial of conduct-disordered youths. *J Fam Psychol, 22*(4), 550-560. doi:10.1037/a0012790

Swendsen, J. (2000). The comorbidity of depression and substance use disorders. *Clinical Psychology Review, 20*(2), 173-189. doi:10.1016/s0272-7358(99)00026-4

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Boston, MA: Pearson.

Tackett, J. L., Lahey, B. B., van Hulle, C., Waldman, I., Krueger, R. F., & Rathouz, P. J. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *J Abnorm Psychol, 122*(4), 1142-1153. doi:10.1037/a0034151

Tallis, R. (2016). *Aping Mankind: Neuromania, Darwinitis and the Misrepresentation of Humanity*. London: Routledge.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613-625. doi:10.1007/bf02289858

Thelen, S. T., Yoo, B., & Magnini, V. P. (2010). An examination of consumer sentiment toward offshored services. *Journal of the Academy of Marketing Science*, *39*(2), 270–289. doi:10.1007/s11747-010-0192-7

Thomas, K. M., Yalch, M. M., Krueger, R. F., Wright, A. G., Markon, K. E., & Hopwood, C. J. (2013). The convergent structure of DSM-5 personality trait facets and five-factor model trait domains. *Assessment, 20*(3), 308-311. doi:10.1177/1073191112457589

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: demonstration of a bifactor model for the Brief Symptom Inventory. *Psychol Assess, 24*(1), 101-113. doi:10.1037/a0024712

Thomson, G. H. (1939). *The factorial analysis of human ability*. Oxford, England: Houghton Mifflin.

Thurstone, L. L. (1940). Current issues in factor analysis. *Psychological Bulletin*, *37*(4), 189.

Thurstone, L. L. (1944). *A factorial study of perception*. Chicago, IL, US: University of Chicago Press.

Thurstone, L. L. (1944). Second-order factors. *Psychometrika, 9*(2), 71–100. doi:10.1007/bf02288715

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of The Vectors of Mind*. Chicago, IL, US: University of Chicago Press.

Tian, K. T., Bearden, W. O., & Hunter, G. L. (2001). Consumers' need for uniqueness: Scale development and validation. *Journal of consumer research*, *28*(1), 50-66. doi:10.1086/321947

Tomás, J. M., Oliver, A., Galiana, L., Sancho, P., & Lila, M. (2013). Explaining Method Effects Associated With Negatively Worded Items in Trait and State Global and Domain-Specific Self-Esteem Scales. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(2), 299-313. doi:10.1080/10705511.2013.769394

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychol Bull, 133*(5), 859-883. doi:10.1037/0033-2909.133.5.859

Trapnell, P. D., & Campbell, J. D. (1999). Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology*, *76*(2), 284–304. doi:10.1037/0022-3514.76.2.284

Trull, T. J., Verges, A., Wood, P. K., & Sher, K. J. (2013). The structure of DSM-IV-TR personality disorder diagnoses in NESARC: a reanalysis. *J Pers Disord, 27*(6), 727-734. doi:10.1521/pedi_2013_27_107

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, *358*(3), 252-260.

Tyrer, P. (2015). Personality dysfunction is the cause of recurrent non-cognitive mental disorder: a testable hypothesis. *Personal Ment Health, 9*(1), 1-7. doi:10.1002/pmh.1255

Tyrer, P., Reed, G. M., & Crawford, M. J. (2015). Classification, assessment, prevalence, and effect of personality disorder. *The Lancet, 385*(9969), 717-726. doi:10.1016/s0140-6736(14)61995-4

Tyrer, P., Tyrer, H., Yang, M., & Guo, B. (2016). Long-term impact of temporary and persistent personality disorder on anxiety and depressive disorders. *Personal Ment Health, 10*(2), 76-83. doi:10.1002/pmh.1324

Unger, T., Hoffmann, S., Kohler, S., Mackert, A., & Fydrich, T. (2013). Personality disorders and outcome of inpatient treatment for depression: a 1-year prospective follow-up study. *J Pers Disord, 27*(5), 636-651. doi:10.1521/pedi_2012_26_052

Urban, R., Arrindell, W. A., Demetrovics, Z., Unoka, Z., & Timman, R. (2016). Cross-cultural confirmation of bi-factor models of a symptom distress measure: Symptom Checklist-90-Revised in clinical samples. *Psychiatry Res, 239*, 265-274. doi:10.1016/j.psychres.2016.03.039

Urban, R., Kun, B., Farkas, J., Paksi, B., Kokonyei, G., Unoka, Z., . . . Demetrovics, Z. (2014). Bifactor structural model of symptom checklists: SCL-90-R and Brief Symptom Inventory (BSI) in a non-clinical community sample. *Psychiatry Res, 216*(1), 146-154. doi:10.1016/j.psychres.2014.01.027

Vainik, U., Mõttus, R., Allik, J., Esko, T., & Realo, A. (2015). Are Trait-Outcome Associations Caused by Scales or Particular Items? Example Analysis of Personality Facets and BMI. *European Journal of Personality, 29*(6), 622-634. doi:10.1002/per.2009 van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology, 27*(6), 759-773. doi:10.1177/0959354317737185

van Bronswijk, S. C., Lemmens, L., Viechtbauer, W., Huibers, M. J. H., Arntz, A., & Peeters, F. (2018). The impact of personality disorder pathology on the effectiveness of Cognitive Therapy and Interpersonal Psychotherapy for Major Depressive Disorder. *J Affect Disord, 225*, 530-538. doi:10.1016/j.jad.2017.08.043

van den Hout, M., Brouwers, C., & Oomen, J. (2006). Clinically diagnosed axis II co-morbidity and the short term outcome of CBT for axis I disorders. *Clinical Psychology & Psychotherapy, 13*(1), 56-63. doi:10.1002/cpp.477

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*(3), 315-327. doi:10.1016/j.jrp.2010.03.003

van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychol Rev, 113*(4), 842-861. doi:10.1037/0033-295X.113.4.842

van der Stouwe, T., Asscher, J. J., Stams, G. J., Dekovic, M., & van der Laan, P. H. (2014). The effectiveness of Multisystemic Therapy (MST): a meta-analysis. *Clin Psychol Rev, 34*(6), 468-481. doi:10.1016/j.cpr.2014.06.006

van Hoof, M. J., Riem, M. M. E., Garrett, A. S., van der Wee, N. J. A., van, I. M. H., & Vermeiren, R. (2019). Unresolved-disorganized attachment adjusted for a general psychopathology factor associated with atypical amygdala restingstate functional connectivity. *Eur J Psychotraumatol, 10*(1), 1583525. doi:10.1080/20008198.2019.1583525

van Os, J., & Reininghaus, U. (2016). Psychosis as a transdiagnostic and extended phenotype in the general population. *World Psychiatry, 15*(2), 118-124. doi:10.1002/wps.20310

van Straten, A., Geraedts, A., Verdonck-de Leeuw, I., Andersson, G., & Cuijpers, P. (2010). Psychological treatment of depressive symptoms in patients with medical disorders: a meta-analysis. *J Psychosom Res, 69*(1), 23-32. doi:10.1016/j.jpsychores.2010.01.019

Van Vaerenbergh, Y., & Thomas, T. D. (2012). Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies. *International Journal of Public Opinion Research, 25*(2), 195-217. doi:10.1093/ijpor/eds021

Van Wert, M., Mishna, F., Trocme, N., & Fallon, B. (2017). Which maltreated children are at greatest risk of aggressive and criminal behavior? An examination of maltreatment dimensions and cumulative risk. *Child Abuse Negl, 69*, 49-61. doi:10.1016/j.chiabu.2017.04.013

Viinamaki, H., Haatainen, K., Honkalampi, K., Tanskanen, A., Koivumaa-Honkanen, H., Antikainen, R., . . . Hintikka, J. (2006). Which factors are important predictors of non-recovery from major depression? A 2-year prospective observational study. *Nord J Psychiatry, 60*(5), 410-416. doi:10.1080/08039480600937801

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series*. *2*, 9-36.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.

Wade, M., Fox, N. A., Zeanah, C. H., & Nelson, C. A. (2018). Effect of Foster Care Intervention on Trajectories of General and Specific Psychopathology Among Children With Histories of Institutional Rearing. *JAMA Psychiatry, 75*(11), 1137. doi:10.1001/jamapsychiatry.2018.2556

Wade, M., Fox, N. A., Zeanah, C. H., Nelson, C. A., & Drury, S. S. (2019). Telomere Length and Psychopathology: Specificity and Direction of Effects Within the Bucharest Early Intervention Project. *J Am Acad Child Adolesc Psychiatry*. doi:10.1016/j.jaac.2019.02.013

Waldman, I. D., Poore, H. E., van Hulle, C., Rathouz, P. J., & Lahey, B. B. (2016). External validity of a hierarchical dimensional model of child and adolescent psychopathology: Tests using confirmatory factor analyses and multivariate behavior genetic analyses. *J Abnorm Psychol, 125*(8), 1053-1066. doi:10.1037/abn0000183

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry, 14*(3), 270-277. doi:10.1002/wps.20238

Wampold, B. & Imel, Z. E. (2015). *The great psychotherapy debate. The evidence for what makes psychotherapy work*. New York: Routledge.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H.-n. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, "all must have prizes.". *Psychological Bulletin, 122*(3), 203-215. doi:10.1037/0033-2909.122.3.203

Warne, R. T., & Burningham, C. (2019). Spearman's g found in 31 non-Western nations: Strong evidence that g is a universal phenomenon. *Psychol Bull, 145*(3), 237-272. doi:10.1037/bul0000184

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier Tests of the Validity of the Bifactor Model of Psychopathology. *Clinical Psychological Science*, 216770261985503. doi:10.1177/2167702619855035

Waugh, M. H., Hopwood, C. J., Krueger, R. F., Morey, L. C., Pincus, A. L., & Wright, A. G. C. (2017). Psychological Assessment with the DSM-5 Alternative Model for Personality Disorders: Tradition and Innovation. *Prof Psychol Res Pr, 48*(2), 79-89. doi:10.1037/pro0000071

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The Individual Consistency of Acquiescence and Extreme Response Style in Self-Report Questionnaires. *Applied Psychological Measurement, 34*(2), 105-121. doi:10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychol Methods, 15*(1), 96-110. doi:10.1037/a0018721

Weijters, B., Schillewaert, N., & Geuens, M. (2007). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409-422. doi:10.1007/s11747-007-0077-6

Weissman, D. G., Bitran, D., Miller, A. B., Schaefer, J. D., Sheridan, M. A., & McLaughlin, K. A. (2019). Difficulties with emotion regulation as a transdiagnostic mechanism linking child maltreatment with the emergence of psychopathology. *Dev Psychopathol, 31*(3), 899-915. doi:10.1017/S0954579419000348

Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., . . . Fordwood, S. R. (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *Am Psychol, 72*(2), 79-117. doi:10.1037/a0040360

Weisz, J. R., McCarty, C. A., & Valeri, S. M. (2006). Effects of psychotherapy for depression in children and adolescents: a meta-analysis. *Psychol Bull, 132*(1), 132-149. doi:10.1037/0033-2909.132.1.132

Westen, D., Moses, M. J., Silk, K. R., Lohr, N. E., Cohen, R., & Segal, H. (1992). Quality of Depressive Experience in Borderline Personality Disorder and Major Depression: When Depression is Not Just Depression. *Journal of Personality Disorders, 6*(4), 382-393. doi:10.1521/pedi.1992.6.4.382

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349-363). New York, NY, US: Oxford University Press.

Wicherts, J. M. (2017). Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence, 60*, 26-38. doi:10.1016/j.intell.2016.11.002

Widaman, K. F. (2016). Hierarchically Nested Covariance Structure Models for Multitrait-Multimethod Data. *Applied Psychological Measurement, 9*(1), 1-26. doi:10.1177/014662168500900101

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological methods*, *8*(1), 16.

Widiger, T. A. (2011). Personality and psychopathology. *World Psychiatry, 10*(2), 103-106. doi:10.1002/j.2051-5545.2011.tb00024.x

Widiger, T. A., & Costa, P. T., Jr. (2013). *Personality disorders and the five-factor model of personality* (3rd ed.). Washington, DC: American Psychological Association.

Widiger, T. A., Livesley, W. J., & Clark, L. A. (2009). An integrative dimensional classification of personality disorder. *Psychol Assess, 21*(3), 243-255. doi:10.1037/a0016606

Widiger, T. A., & Oltmanns, J. R. (2017). The General Factor of Psychopathology and Personality. *Clin Psychol Sci, 5*(1), 182-183. doi:10.1177/2167702616657042

Wiggins, C. W., Wygant, D. B., Hoelzle, J. B., & Gervais, R. O. (2012). The More You Say the Less It Means: Overreporting and Attenuated Criterion Validity in a Forensic Disability Sample. *Psychological Injury and Law, 5*(3-4), 162-173. doi:10.1007/s12207-012-9137-4

Williams, L. J., & Anderson, S. E. (1994). An alternative approach to method effects by using latent-variable models: Applications in organizational behavior research. *Journal of Applied Psychology, 79*(3), 323-331. doi:10.1037/0021-9010.79.3.323

Williams, P., Tarnopolsky, A., & Hand, D. (1980). Case definition and case identification in psychiatric epidemiology: review and assessment. *Psychological medicine*, *10*(1), 101-114.

Williams, T. F., Scalco, M. D., & Simms, L. J. (2018). The construct validity of general and specific dimensions of personality pathology. *Psychol Med, 48*(5), 834-848. doi:10.1017/S0033291717002227

Winiarski, D. A., Schechter, J. C., Brennan, P. A., Foster, S. L., Cunningham, P. B., & Whitmore, E. A. (2017). Adolescent Physiological and Behavioral Patterns of Emotion Dysregulation Predict Multisystemic Therapy Response. *J Emot Behav Disord, 25*(3), 131-142. doi:10.1177/1063426616638315

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychol Methods, 12*(1), 58-79. doi:10.1037/1082-989X.12.1.58

Wittchen, H. U., Nelson, C. B., & Lachner, G. (1998). Prevalence of mental disorders and psychosocial impairments in adolescents and young adults. *Psychol Med, 28*(1), 109-126. doi:10.1017/s0033291797005928

Wolff, H. G., & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. *Behavior Research Methods, 37*(1), 48-58. doi:10.3758/bf03206397

Wood, A., Kroll, L., Moore, A., & Harrington, R. (1995). Properties of the Mood and Feelings Questionnaire in Adolescent Psychiatric Outpatients: A Research Note. *Journal of Child Psychology and Psychiatry, 36*(2), 327-334. doi: 10.1111/j.1469-7610.1995.tb01828.x

Woods, W. C., Edershile, E. A., Wright, A. G. C., & Lenzenweger, M. F. (2019). Illuminating ipsative change in personality disorder and normal personality: A multimethod examination from a prospective longitudinal perspective. *Personal Disord, 10*(1), 80-86. doi:10.1037/per0000302

World Health Organization. (2018). *International statistical classification of diseases and related health problems* (11th Revision). Retrieved from https://icd.who.int/browse11/l-m/en

Wright, A. G. (in press). Latent variable models in clinical psychology. In Wright, A. G., & Hallquist, M. N. (Eds.). *Cambridge handbook of research methods in clinical psychology*. New York, NY: Cambridge University Press.

Wright, A. G., Hopwood, C. J., Skodol, A. E., & Morey, L. C. (2016). Longitudinal validation of general and specific structural features of personality pathology. *J Abnorm Psychol, 125*(8), 1120-1134. doi:10.1037/abn0000165

Wright, A. G., Thomas, K. M., Hopwood, C. J., Markon, K. E., Pincus, A. L., & Krueger, R. F. (2012). The hierarchical structure of DSM-5 pathological personality traits. *J Abnorm Psychol, 121*(4), 951-957. doi:10.1037/a0027669

Wright, A. G., & Simms, L. J. (2014). On the structure of personality disorder traits: conjoint analyses of the CAT-PD, PID-5, and NEO-PI-3 trait models. *Personal Disord, 5*(1), 43-54. doi:10.1037/per0000037

Wu, J. (2013). Hierarchy theory: an overview. In M. J. Torres, S. Pickett, C. Palmer, J. J. Armesto, & J. B. Callicott (Eds.), *Linking ecology and ethics for a changing world* (pp. 281-301). Dordrecht: Springer.

Wusten, C., Schlier, B., Jaya, E. S., Genetic, R., Outcome of Psychosis, I., Fonseca-Pedrero, E., . . . Lincoln, T. M. (2018). Psychotic Experiences and Related Distress: A Cross-national Comparison and Network Analysis Based on 7141 Participants From 13 Countries. *Schizophr Bull, 44*(6), 1185-1194. doi:10.1093/schbul/sby087

Yang-Wallentin, F., Joreskog, K., & Luo, H. (2010). Confirmatory Factor Analysis of Ordinal Variables With Misspecified Models. *Structural Equation Modeling: A Multidisciplinary Journal, 17*(3), 392-423. doi:10.1080/10705511.2010.489003

Yao, S., Zhang, C., Zhu, X., Jing, X., McWhinnie, C. M., & Abela, J. R. Z. (2009). Measuring Adolescent Psychopathology: Psychometric Properties of the Self-Report Strengths and Difficulties Questionnaire in a sample of Chinese adolescents. *Journal of Adolescent Health*, *45*, 55–62. doi: 10.9734/indj/2017/37760

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*(2), 113-128. doi:10.1007/bf02294531

Zald, D. H., & Lahey, B. B. (2017). Implications of the Hierarchical Structure of Psychopathology for Psychiatric Neuroimaging. *Biol Psychiatry Cogn Neurosci Neuroimaging, 2*(4), 310-317. doi:10.1016/j.bpsc.2017.02.003

Zeanah, C. H., Egger, H. L., Smyke, A. T., Nelson, C. A., Fox, N. A., Marshall, P. J., & Guthrie, D. (2009). Institutional rearing and psychiatric disorders in Romanian preschool children. *Am J Psychiatry, 166*(7), 777-785. doi:10.1176/appi.ajp.2009.08091438

Zelazo, P. D., & Cunningham, W. A. (2007). Executive Function: Mechanisms Underlying Emotion Regulation. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 135-158). New York, NY: Guilford Press.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2016). Estimating Generalizability to a Latent Variable Common to All of a Scale's Indicators: A Comparison of Estimators for ωh. *Applied Psychological Measurement, 30*(2), 121-144. doi:10.1177/0146621605278814

Appendices

Appendix A. Study Characteristics and Model-Based Reliability Estimates for Bifactor Studies of Psychopathology Published Between April 2009-June 2019.

Author	Method	Sample	Items	Factor	ECV(s)	ω(s)	ω_H(s)	Rel. ω	H	FD	PUC
Urban et al.	A. SCL-90	D. 2710	83	*p*	.84	.99	.97	.98	.99	.99	.89
(2014)	B. Questionnaire	E. 40	12	Somatic	.06	.95	.37	.39	.80	.93	
	C. Item	F. Self	10	O-C	.01	.92	.03	.04	.27	.64	
		G. Population	9	IS	.01	.91	.05	.05	.26	.65	
			13	Depressive	.02	.94	.09	.10	.53	.82	
			10	Anxious	.02	.94	.08	.09	.48	.81	
			6	Hostile	.02	.90	.16	.18	.56	.87	
			7	Phobic	.02	.93	.17	.18	.56	.89	
			6	Paranoia	.01	.87	.09	.11	.33	.69	
			10	Psychoticism	.01	.93	.02	.02	.25	.65	
Miller et al.	A. CBCL, ADOS	D. 415	32	*p* (DP)	.80	.97	.93	.96	.97	.98	.58
(2019)	B. Questionnaire	E. 3	8	Anx/depressed	.09	.89	.33	.37	.70	.87	
	C. Item	F. Caregiver	19	Aggressive	.06	.97	.02	.02	.58	.88	
		G. Community	5	Attention	.06	.84	.31	.38	.60	.88	
Neumann et al.	A. CBCL, SRS,										
(2016)	CPRS, TRF	D. 1954	28	*p*	.76	.81	.73	.90	.85	.91	.74
	B. Questionnaire	E. 7	10	Internalizing	.07	.55	.06	.11	.32	.57	
	C. Subscale	F. Caregiver	11	Externalizing	.16	.69	.30	.43	.50	.70	
		G. Population									

Geeraerts et al. (2015)	A. CBCL	D. 247	32	*p* (DP)	.76	.98	.90	.92	.98	.99	.58
	B. Questionnaire	E. 5	8	Anx/depressed	.10	.91	.41	.46	.79	.94	
	C. Item	F. Caregiver	19	Aggressive	.09	.98	.09	.10	.74	.92	
		G. Clinical	5	Attention	.05	.90	.24	.27	.61	.94	
Laceulle, Vollebergh, & Ormel (2015)	A. YSR, RCADS, CAPE	D. 2230	12	*p*	.76	.98	.89	.91	.97	.98	.73
	B. Questionnaire	E. 11-19	6	Internalizing	.10	.95	.17	.18	.55	.89	
	C. Item	F. Self	3	Externalizing	.14	.94	.51	.55	.79	.96	
		G. Community									
Preti, Carta, & Petretto (2019)	A. SCL-90	D. 817	83	*p*	.76	.97	.95	.97	.97	.98	.89
	B. Questionnaire	E. 18	12	Somatic	.02	.84	.07	.09	.36	.63	
	C. Item	F. Self	10	O-C	.03	.81	.16	.20	.47	.73	
		G. Community	9	IS	.04	.86	.27	.31	.61	.83	
			13	Depressive	.04	.89	.15	.16	.56	.80	
			10	Anxious	.01	.86	.02	.02	.24	.58	
			6	Hostile	.04	.83	.32	.38	.63	.83	
			7	Phobic	.03	.72	.27	.38	.49	.74	
			6	Paranoia	.02	.78	.20	.25	.42	.71	
			10	Psychoticism	.02	.80	.17	.21	.46	.72	
St Clair et al. (2017)	A. Multiple	D. 2228	106	*p*	.76	.97	.92	.94	.99		.87
	B. Questionnaire	E. 19	13	Self-confidence	.04	.93	.31	.33	.72		
	C. Item	F. Self	9	Antisocial	.06	.90	.56	.62	.83		
		G. Community	7	Worry	.02	.94	.18	.19	.55		
			17	Aberrant thgts	.07	.91	.48	.53	.85		
			30	Mood	.06	.79	.00	.01	.79		
Urban, Arrindell,	A. SCL-90	D. 972	83	*p*	.75	.99	.96	.97	.98	.99	.89
	B. Questionnaire	E. 35	12	Somatic	**.05**	.94	**.33**	.36	.76	.91	

Demetrovics,											
Unoka, &											
Timman (2016)	C. Item	F. Self	10	O-C	**.02**	.91	**.06**	.06	.52	.81	
		G. Clinical	9	IS	**.03**	.91	**.26**	.29	.63	.86	
			13	Depressive	**.01**	.94	**.02**	.02	.35	.75	
			10	Anxious	**.02**	.92	**.02**	.02	.46	.81	
			6	Hostile	**.04**	.90	**.45**	.50	.75	.91	
			7	Phobic	**.04**	.92	**.34**	.37	.76	.94	
			6	Paranoia	**.02**	.83	**.27**	.32	.53	.81	
			10	Psychoticism	**.02**	.85	**.09**	.10	.46	.75	
Preti, Carta, &	A. SCL-90	D. 507	83	*p*	.74	.97	.94	.97	.97	.98	.89
Petretto (2019)	B. Questionnaire	E. 17	12	Somatic	.06	.85	.39	.46	.69	.84	
	C. Item	F. Self	10	O-C	.02	.80	.10	.13	.38	.67	
		G. Community	9	IS	.03	.84	.18	.22	.53	.79	
			13	Depressive	.03	.89	.10	.11	.54	.77	
			10	Anxious	.03	.84	.17	.20	.52	.76	
			6	Hostile	.04	.76	.38	.50	.59	.80	
			7	Phobic	.02	.75	.11	.15	.40	.68	
			6	Paranoia	.02	.76	.15	.19	.37	.65	
			10	Psychoticism	.01	.81	.01	.01	.31	.61	
Hankin et al.	A. CBCL	D. 554	8	*p*	.72	.94	.82	.88	.91	.94	.79
(2017)	B. Questionnaire	E. 8	3	Internalizing	.15	.85	.33	.39	.58	.87	
	C. Subscale	F. Caregiver	3	Externalizing	.13	.89	.29	.33	.48	.77	
		G. Clinical									
McElroy,	A. CBCL	D. 1253	68	*p*	.71	.97	.86	.89	.97	.98	.60
Belsky,	B. Questionnaire	E. 11	31	Internalizing	.15	.93	.37	.40	.84	.92	
Carragher,	C. Item	F. Caregiver	30	Externalizing	.10	.96	.11	.12	.78	.90	

Fearon, &		G. Community	7	Attention	.04	.86	.26	.30	.62	.85	
Patalay (2017)											
McElroy,	A. CBCL	D. 1253	70	*p*	.71	.97	.86	.89	.97	.98	.60
Belsky,	B. Questionnaire	E. 14	31	Internalizing	.15	.93	.37	.40	.84	.92	
Carragher,	C. Item	F. Caregiver	31	Externalizing	.10	.96	.11	.12	.78	.90	
Fearon, &		G. Community	8	Attention	.04	.86	.26	.30	.62	.85	
Patalay (2017)											
McElroy,	A. CBCL	D. 1253	67	*p*	.71	.96	.86	.89	.96	.97	.61
Belsky,	B. Questionnaire	E. 5	30	Internalizing	.15	.91	.36	.39	.81	.89	
Carragher,	C. Item	F. Caregiver	29	Externalizing	.10	.94	.09	.10	.82	.94	
Fearon, &		G. Community	8	Attention	.05	.83	.21	.25	.65	.90	
Patalay (2017)											
Deutz et al.	A. SDQ	D. 768	15	*p* (DP)	.70	.94	.85	.90	.92	.95	.71
(2018)	B. Questionnaire	E. 14	5	Emotional	.11	.82	.32	.40	.55	.78	
	C. Item	F. Caregiver	5	Conduct	.08	.88	.13	.14	.52	.82	
		G. Community	5	Hyp-inattention	.11	.90	.21	.23	.68	.95	
McElroy,	A. CBCL	D. 1253	66	*p*	.69	.97	.85	.88	.97	.98	.61
Belsky,	B. Questionnaire	E. 10	31	Internalizing	.16	.92	.39	.42	.84	.91	
Carragher,	C. Item	F. Caregiver	27	Externalizing	.10	.95	.10	.11	.76	.90	
Fearon, &		G. Community	8	Attention	.05	.87	.27	.32	.62	.85	
Patalay (2017)											
Calkins et al.	A. GOASSESS	D. 9498	15	*p*	.69	.90	.81	.90	.88	.93	.70
(2015)	B. Interview	E. 14	5	Anxious-misery	.05	.78	.08	.10	.27	.57	
	C. Subscale	F. Multiple	6	Fear	.06	.77	.08	.10	.34	.63	
		G. Community	4	Behavioral	.20	.82	.49	.60	.69	.86	

Tackett et al. (2013)	A. CAPS	D. 1569	11	*p*	.69	.92	.77	.84	.95	.98	.71
	B. Interview	E. 14	5	Internalizing	.20	.79	.43	.55	.66	.84	
	C. Subscale	F. Multiple	4	Externalizing	.12	.85	.26	.30	.50	.82	
		G. Community									
McElroy, Belsky, Carragher, Fearon, & Patalay (2017)	A. CBCL	D. 1253	65	*p*	.68	.97	.84	.87	.96	.97	.61
	B. Questionnaire	E. 9	31	Internalizing	.16	.93	.35	.37	.83	.90	
	C. Item	F. Caregiver	26	Externalizing	.10	.95	.15	.16	.76	.87	
		G. Community	8	Attention	.06	.86	.34	.39	.66	.84	
Wade, Fox, Zeanah, & Nelson (2018)	A. MHBQ	D. 220	8	*p*	.68	.95	.78	.82	.91	.93	.54
	B. Questionnaire	E. 16	3	Internalizing	.16	.87	.41	.47	.61	.82	
	C. Subscale	F. Caregiver	5	Externalizing	.16	.95	.21	.22	.62	.94	
		G. Clinical									
Arrindell et al. (2017)	A. SCL-90	D. 2593	83	*p*	.67	.99	.94	.95	.98	.99	.89
	B. Questionnaire	E. 37	12	Somatic	.07	.92	.48	.53	.82	.93	
	C. Item	F. Self	10	O-C	.03	.91	.23	.25	.66	.86	
		G. Clinical	9	IS	.03	.91	.24	.26	.65	.88	
			13	Depressive	.01	.93	.03	.03	.43	.79	
			10	Anxious	.04	.94	.23	.24	.69	.89	
			6	Hostile	.05	.91	.55	.61	.81	.94	
			7	Phobic	.04	.92	.40	.43	.73	.91	
			6	Paranoia	.02	.85	.31	.37	.58	.83	
			10	Psychoticism	.03	.89	.25	.28	.67	.86	
Lahey et al. (2012)	A. AUDADIS-IV	D. 43093	11	*p*	.66	.93	.77	.83	.91	.94	.71
	B. Interview	E. 35	3	Distress	.04	.90	.10	.11	.25	.58	

	C. Subscale	F. Self	3	Fear	.06	.78	.16	.21	.32	.62	
		G. Population	5	Externalizing	.24	.87	.49	.56	.72	.87	
McElroy, Belsky, Carragher, Fearon, & Patalay (2017)	A. CBCL	D. 1253	66	*p*	.66	.96	.82	.85	.96	.97	.61
	B. Questionnaire	E. 6	31	Internalizing	.19	.91	.48	.53	.84	.91	
	C. Item	F. Caregiver	27	Externalizing	.09	.94	.08	.08	.70	.86	
		G. Community	8	Attention	.06	.85	.32	.37	.64	.83	
Liu, Mustanski, Dick, Bolland, & Kertes (2017)	A. YSR	D. 592	12	*p*	.65	.92	.77	.83	.90	.94	.55
	B. Questionnaire	E. 16	6	Internalizing	.11	.88	.05	.06	.65	.94	
	C. Subscale	F. Self	6	Externalizing	.23	.87	.44	.50	.69	.85	
		G. Community									
Caspi et al. (2014)	A. DIS	D. 1037	11	*p*	.65	.95	.77	.81	.96	.98	.76
	B. Interview	E. 18-38	3	Internalizing	.07	.91	.22	.24	.41	.84	
	C. Subscale	F. Self	5	Externalizing	.28	.91	.59	.65	.82	.95	
		G. Population									
McElroy, Belsky, Carragher, Fearon, & Patalay (2017)	A. CBCL	D. 1253	66	*p*	.64	.97	.81	.84	.96	.97	.61
	B. Questionnaire	E. 8	31	Internalizing	.19	.92	.47	.51	.86	.92	
	C. Item	F. Caregiver	27	Externalizing	.11	.95	.14	.15	.79	.91	
		G. Community	8	Attention	.05	.86	.32	.37	.65	.84	
Pettersson, Lahey, Larsson, & Lichtenstein, (2018)	A. ATAC	D. 8403	43	*p*	.64	.98	.86	.88	.97	.96	.77
	B. Questionnaire	E. 9 or 12	12	Anxiety	.08	.90	.30	.34	.73	.86	
	C. Item	F. Caregiver	10	Conduct	.06	.93	.22	.24	.69	.85	
		G. Community	11	Inattention	.12	.96	.39	.41	.83	.91	
			10	Impulsivity	.09	.95	.35	.37	.77	.89	
	A. CBCL, CSBQ	D. 2230	12	*p*	.64	.92	.76	.83	.88	.90	.48

Noordhof, Krueger, Ormel, Oldehinkel, & Hartman (2015)	B. Questionnaire	E. 14	4	Internalizing	.09	.76	.26	.34	.44	.72	
	C. Subscale	F. Caregiver	3	Externalizing	.09	.89	.25	.28	.44	.75	
		G. Community	5	Attention	.06	.84	.12	.14	.34	.66	
Haltigan et al. (2018)	A. CBCL	D. 2934	78	*p*	.63	.98	.83	.85	.97	.97	.69
	B. Questionnaire	E. 13	30	Internalizing	.12	.95	.26	.28	.83	.90	
	C. Item	F. Caregiver	29	Externalizing	.19	.95	.48	.51	.90	.94	
		G. Clinical	12	Thought prob.	.04	.90	.10	.11	.66	.87	
			7	Attention	.02	.82	.14	.17	.54	.84	
Martel et al. (2017)	A. FHS	D. 8012	11	*p*	.63	.91	.79	.87	.88	.93	.68
	B. Questionnaire	E. 36	3	Internalizing	.09	.83	.09	.11	.47	.76	
	C. Subscale	F. Self	5	Externalizing	.21	.82	.53	.64	.70	.85	
		G. Community	3	Thought dis.	.07	.78	.02	.02	.45	.82	
Constantinou, Allison, & Fonagy (2019)	A. ASR	D. 1200	99	*p*	.62	.98	.83	.85	.98	.99	.73
	B. Questionnaire	E. 37	27	Internalizing	.05	.97	.09	.09	.74	.90	
	C. Item	F. Self	35	Externalizing	.2	.95	.64	.67	.93	.97	
		G. Community	25	Cognitive	.08	.94	.32	.34	.83	.92	
			12	Somatic	.05	.92	.37	.40	.76	.91	
Hyland et al. (2018)	A. MCMI	D. 420	9	*p*	.62	.93	.81	.87	.95	.99	.72
	B. Questionnaire	E. 36	4	Internalizing	.10	.96	.04	.04	.47	.95	
	C. Subscale	F. Self	2	Externalizing	.16	.74	.64	.87	.67	.84	
		G. Clinical	3	Thought dis.	.13	.80	.38	.47	.63	.86	
McElroy, Belsky, Carragher,	A. CBCL	D. 1253	60	*p*	.61	.96	.75	.78	.95	.95	.54
	B. Questionnaire	E. 3	36	Internalizing	.24	.93	.43	.46	.87	.91	
	C. Item	F. Caregiver	19	Externalizing	.11	.93	.23	.25	.75	.84	

Fearon, &		G. Community	5	Attention	.04	.79	.19	.24	.63	.88	
Patalay (2017)											
Wade, Fox,	A. MHBQ	D. 220	8	*p*	.61	.95	.73	.77	.98	.99	.54
Zeanah, &	B. Questionnaire	E. 8	3	Internalizing	.23	.85	.62	.73	.83	.93	
Nelson (2018)	C. Subscale	F. Caregiver	5	Externalizing	.16	.96	.20	.20	.57	.93	
		G. Clinical									
Stochl et al.	A. MFQ, PLIKS-										
(2015)	Q, DISC-IV,										
	SCAN 2	D. 1074	25	*p*	.61	.95	.71	.75	.93	.96	.52
	B. Questionnaire	E. 17	13	Anx/depressed	.27	.94	.39	.41	.84	.91	
	C. Item	F. Self	12	Psychotic exp.	.12	.92	.24	.26	.98	.99	
		G. Community									
Brodbeck et al.	A. BSI	D. 1024	53	*p*	.61	.98	.90	.92	.97	.97	.86
(2014)	B. Questionnaire	E. 40	8	Depression	.05	.26	.23	.25	.67	.90	
	C. Item	F. Self	10	Phobia	.09	.44	.42	.46	.81	.92	
		G. Clinical	3	Aggression	.04	.55	.42	.52	.68	.91	
			4	Suicidal	.03	.36	.28	.32	.58	.86	
			3	Nervous tension	.02	.37	.29	.35	.49	.81	
			7	Somatic	.06	.46	.42	.48	.70	.88	
			6	Info. processing	.04	.29	.26	.29	.58	.85	
			12	IS	.06	.32	.28	.30	.71	.86	
Rytilä-	A. SCL-90	D. 201	90	*p*	.60	.99	.92	.93	.99	.99	.91
Manninen et al.	B. Questionnaire	E. 15	12	Somatic	.07	.92	.53	.58	.84	.94	
(2016)	C. Item	F. Self	10	O-C	.05	.94	.43	.46	.80	.93	
		G. Clinical	9	IS	.04	.92	.36	.39	.75	.92	
			13	Depressive	.03	.96	.19	.20	.66	.88	

			10	Anxious	.05	.94	.40	.43	.79	.93	
			6	Hostile	.04	.89	.49	.55	.77	.92	
			7	Phobic	.06	.92	.61	.66	.86	.95	
			6	Paranoia	.03	.85	.46	.54	.69	.89	
			10	Psychoticism	.04	.91	.36	.39	.73	.89	
McElroy, Belsky, Carragher, Fearon, & Patalay (2017)	A. CBCL	D. 1253	60	*p*	.59	.96	.74	.77	.95	.96	.54
	B. Questionnaire	E. 2	36	Internalizing	.28	.94	.47	.50	.90	.94	
	C. Item	F. Caregiver	19	Externalizing	.09	.93	.18	.20	.72	.84	
		G. Community	5	Attention	.04	.76	.18	.24	.59	.84	
Deutz et al. (2018)	A. SDQ	D. 768	15	*p* (DP)	.58	.93	.76	.82	.91	.95	.71
	B. Questionnaire	E. 7	5	Emotional	.20	.82	.59	.71	.73	.87	
	C. Item	F. Caregiver	5	Conduct	.13	.84	.34	.41	.62	.83	
		G. Community	5	Hyp-inattention	.08	.90	.07	.07	.55	.91	
Schaefer et al. (2018)	A. DIS + others	D. 2066	11	*p*	.57	.88	.70	.79	.84	.89	.69
	B. Interview	E. 18	5	Internalizing	.26	.79	.49	.62	.71	.85	
	C. Subscale	F. Self	4	Externalizing	.07	.76	.13	.17	.31	.59	
		G. Population	2	Thought dis.	.10	.84	.30	.35	.48	.78	
Deutz et al. (2018)	A. SDQ	D. 768	15	*p* (DP)	.57	.92	.74	.80	.89	.91	.71
	B. Questionnaire	E. 10	5	Emotional	.17	.79	.49	.62	.68	.83	
	C. Item	F. Caregiver	5	Conduct	.11	.88	.20	.23	.60	.85	
		G. Community	5	Hyp-inattention	.16	.87	.39	.44	.65	.81	
Martel et al. (2017)	A. DAWBA	D. 2512	15	*p*	.56	.92	.65	.70	.88	.89	.45
	B. Interview	E. 10	11	Internalizing	.29	.89	.40	.45	.75	.83	
	C. Subscale	F. Self	3	Externalizing	.15	.89	.43	.48	.69	.89	

		G. Community									
Harden et al.	A. CBCL, CPRS,										
(2019)	BFI	D. 1913	10	*p*	.55	.92	.70	.76	.85	.90	.64
	B. Questionnaire	E. 13	3	Internalizing	.15	.80	.43	.54	.62	.82	
	C. Subscale	F. Self	5	Externalizing	.20	.88	.37	.43	.63	.85	
		G. Community	3	Attention	.10	.88	.29	.33	.45	.79	
Urban,	A. SCL-90	D. 1902	83	*p*	.55	.98	.90	.92	.97	.98	.89
Arrindell,	B. Questionnaire	E. 30	12	Somatic	.09	.89	.65	.74	.84	.92	
Demetrovics,	C. Item	F. Self	10	O-C	.05	.88	.32	.36	.72	.87	
Unoka, &		G. Clinical	9	IS	.04	.90	.27	.30	.67	.88	
Timman (2016)			13	Depressive	.02	.93	.04	.05	.54	.84	
			10	Anxious	.05	.91	.38	.42	.75	.89	
			6	Hostile	.06	.90	.65	.72	.85	.94	
			7	Phobic	.07	.91	.60	.66	.83	.93	
			6	Paranoia	.03	.83	.42	.51	.65	.85	
			10	Psychoticism	.04	.85	.28	.33	.72	.87	
Black,	A. M&MS,										
Panayiotou, &	CORS	D. 1982	19	*p*	.55	.92	.74	.80	.88	.88	.67
Humphrey	B. Questionnaire	E. 11	9	Internalizing	.12	.87	.21	.24	.58	.70	
(2019)	C. Item	F. Self	6	Externalizing	.20	.87	.50	.57	.74	.85	
		G. Community	4	Wellbeing	.13	.76	.44	.58	.81	.97	
Jones et al.	A. MFQ, PLIKS-										
(2018)	Q, DAWBA	D. 3650	51	*p*	.54	.97	.79	.81	.96	.95	.75
	B. Questionnaire	E. 16	13	Depression	.07	.95	.15	.15	.69	.82	
	C. Subscale	F. Self	17	Anxiety	.19	.92	.66	.72	.88	.93	
		G. Community	10	Psychosis (pos)	.12	.91	.57	.62	.83	.91	

			11	Psychosis (neg)	.08	.92	.33	.35	.73	.86	
Snyder, Young, & Hankin (2017)	A. CDI, MASC, CBCL, EAT-QR, SNAP-IV	D. 519	9	*p*	.53	.92	.70	.76	.85	.90	.75
	B. Questionnaire	E. 15	4	Internalizing	.29	.84	.62	.74	.77	.89	
	C. Subscale	F. Multiple	3	Externalizing	.19	.91	.42	.46	.62	.82	
		G. Community									
Gomez, Stavropoulos, Vance, & Griffiths (2019)	A. ADISC-IV	D. 866	13	*p*	.52	.87	.68	.78	.85	.93	.38
	B. Interview	E. >12	10	Internalizing	.21	.86	.14	.16	.62	.83	
	C. Subscale	F. Caregiver	3	Externalizing	.27	.81	.75	.93	.84	.92	
		G. Clinical									
Wade, Fox, Zeanah, & Nelson (2018)	A. MHBQ	D. 220	8	*p*	.52	.96	.63	.66	.94	.98	.54
	B. Questionnaire	E. 12	3	Internalizing	.12	.90	.31	.34	.56	.76	
	C. Subscale	F. Caregiver	5	Externalizing	.35	.96	.53	.55	.83	.96	
		G. Clinical									
Snyder, Young, & Hankin (2017)	A. CDI, MASC, CBCL, EAT-QR, SNAP-IV	D. 571	9	*p*	.52	.90	.69	.76	.84	.90	.75
	B. Questionnaire	E. 14	4	Internalizing	.25	.81	.56	.69	.69	.84	
	C. Subscale	F. Multiple	3	Externalizing	.23	.92	.47	.51	.68	.85	
		G. Community									
Ignatyev, Baggio, & Mundt (2018)	A. MINI, SCID-II	D. 427	10	*p*	.51	.78	.52	.67	.80	.88	.53
	B. Interview	E. 21	6	Internalizing	.29	.67	.42	.62	.58	.76	
	C. Subscale	F. Self	4	Externalizing	.20	.74	.24	.32	.58	.77	
		G. Clinical									

Lahey et al. (2015)	A. CSI, SCARED	D. 2450	10	*p*	.51	.89	.60	.67	.82	.85	.44
	B. Questionnaire	E. 5-11	6	Internalizing	.21	.86	.29	.34	.63	.76	
	C. Subscale	F. Caregiver	5	Externalizing	.28	.86	.45	.53	.73	.84	
		G. Community									
Constantinou et al. (2019)	A. SDQ, MFQ	D. 683	20	*p*	.50	.91	.79	.87	.87	.92	.67
	B. Questionnaire	E. 14-16	5	Mood	.13	.83	.43	.52	.64	.81	
	C. Item	F. Self	8	Anxiety	.14	.78	.01	.01	.63	.83	
		G. Clinical	6	Antisocial	.10	.74	.20	.27	.57	.77	
			5	Attention	.12	.82	.35	.43	.66	.84	
Lahey et al. (2017)	A. DISC	D. 499	11	*p*	.49	.81	.61	.75	.77	.86	.77
	B. Interview	E. 26	4	Internalizing	.34	.77	.59	.77	.73	.85	
	C. Subscale	F. Self	4	Externalizing	.17	.65	.42	.65	.53	.74	
		G. Community									
Castellanos-Ryan et al. (2016)	A. DAWBA	D. 2144	12	*p*	.48	.75	.51	.69	.75	.88	.55
	B. Interview	E. 16	6	Internalizing	.30	.63	.54	.87	.63	.80	
	C. Subscale	F. Multiple	6	Externalizing	.21	.74	.05	.07	.52	.78	
		G. Community									
Stochl et al. (2015)	A. MFQ, PLIKS-Q, DISC-IV, SCAN 2	D. 6617	25	*p*	.48	.95	.64	.67	.89	.84	.52
	B. Questionnaire	E. 13	13	Anx-depressed	.24	.92	.38	.42	.82	.84	
	C. Item	F. Self	12	Psychotic exp.	.28	.93	.53	.57	.85	.86	
		G. Community									
	A. DAWBA	D. 2144	12	*p*	.47	.81	.57	.70	.75	.87	.55
	B. Interview	E. 14	6	Internalizing	.31	.72	.54	.74	.66	.81	

Castellanos-											
Ryan et al.											
(2016)	C. Subscale	F. Multiple	6	Externalizing	.22	.75	.10	.13	.60	.81	
		G. Community									
Olino et al.	A. PAPA	D. 545	9	*p*	.44	.79	.49	.62	.81	.90	.56
(2018)	B. Interview	E. 3	5	Internalizing	.21	.69	.35	.51	.56	.74	
	C. Subscale	F. Caregiver	4	Externalizing	.35	.78	.55	.71	.70	.85	
		G. Community									
Carragher et al.	A. SDQ, BSI,										
(2016)	RAPI, DISC	D. 2175	44	*p*	.42	.97	.70	.72	.94	.96	.65
	B. Questionnaire	E. 13	20	Internalizing	.23	.96	.44	.45	.93	.96	
	C. Item	F. Self	15	Externalizing	.19	.94	.49	.52	.96	.98	
		G. Community	9	Thought dis.	.15	.94	.66	.70	.89	.95	
Patalay et al.	A. SDQ, M&MS	D. 23447	25	*p*	.42	.94	.57	.61	.88	.88	.51
(2015)	B. Questionnaire	E. 12	14	Internalizing	.24	.91	.42	.46	.82	.87	
	C. Item	F. Self	11	Externalizing	.34	.92	.66	.73	.88	.92	
		G. Community									
Niarchou et al.	A. K-SADS	D. 331	60	*p*	.42	.98	.71	.72	.97	.98	.75
(2017)	B. Interview	E. 17	17	Mood	.15	.96	.54	.56	.93	.97	
	C. Item	F. Self	9	Anxiety	.10	.94	.60	.64	.94	.98	
		G. Clinical	15	Psychosis	.11	.96	.39	.41	.90	.97	
			19	ADHD	.23	.97	.69	.71	.97	.98	
Afzali,	A. SDQ, BSI	D. 3826	36	*p*	.42	.95	.64	.67	.91	.90	.61
Sunderland,	B. Questionnaire	E. 13	20	Internalizing	.29	.94	.47	.49	.90	.91	
	C. Item	F. Self	7	Externalizing	.08	.82	.33	.40	.75	.87	

Carragher, & Conrod (2017)		G. Community	9	Thought dis.	.21	.92	.69	.75	.87	.92	
Pezzoli, Antfolk, & Santtila (2017)	A. Multiple	D. 13024	9	*p*	.42	.86	.58	.67	.77	.85	.72
	B. Questionnaire	E. 35	2	Internalizing	.10	.82	.30	.36	.40	.65	
	C. Subscale	F. Self	4	Externalizing	.29	.76	.58	.77	.84	.92	
		G. Population	3	Body	.19	.77	.37	.48	.76	.92	
Gomez, Stavropoulos, Vance, & Griffiths (2019)	A. ADISC-IV	D. 1233	13	*p*	.41	.88	.47	.54	.87	.92	.38
	B. Interview	E. <12	10	Internalizing	.35	.88	.40	.46	.76	.87	
	C. Subscale	F. Caregiver	3	Externalizing	.24	.78	.73	.93	.79	.91	
		G. Clinical									
Conway, Mansolf, & Reise (2019)	A. Diagnostic Screener	D. 25002	15	*p*	.41	.87	.58	.67	.78	.81	.57
	B. Questionnaire	E. 22	9	Internalizing	.24	.81	.40	.49	.69	.79	
	C. Subscale	F. Self	4	Externalizing	.15	.70	.40	.56	.63	.79	
		G. Community	3	Eating problems	.20	.88	.54	.62	.73	.87	
Olino et al. (2018)	A. PAPA	D. 545	9	*p*	.41	.79	.45	.58	.75	.86	.56
	B. Questionnaire	E. 6	5	Internalizing	.31	.75	.52	.69	.65	.80	
	C. Subscale	F. Caregiver	4	Externalizing	.28	.73	.47	.64	.65	.81	
		G. Community									
Murray, Eisner, & Ribeaud (2016)	A. SBQ	D. 1572	40	*p*	.40	.95	.70	.73	.93	1.00	.69
	B. Questionnaire	E. 10	8	Internalizing	.16	.85	.83	.97	.87	.94	
	C. Item	F. Teacher	9	ADHD	.10	.97	.32	.33	.75	1.00	
		G. Community	18	Aggression	.20	.92	.54	.58	.84	.94	
			8	Prosociality	.14	.84	.79	.95	.83	.92	
	A. SBQ	D. 1572	40	*p*	.36	.95	.67	.71	.89	.85	.69

Murray, Eisner, & Ribeaud (2016)	B. Questionnaire	E. 12	8	Internalizing	.16	.88	.73	.84	.87	.93	
	C. Item	F. Teacher	9	ADHD	.15	.91	.52	.57	.83	.88	
		G. Community	18	Aggression	.17	.92	.44	.48	.81	.81	
			8	Prosociality	.15	.86	.74	.87	.84	.91	
Murray, Eisner, & Ribeaud (2016)	A. SBQ	D. 1572	40	*p*	.36	.95	.66	.70	.90	.86	.69
	B. Questionnaire	E. 9	8	Internalizing	.16	.86	.81	.95	.88	.94	
	C. Item	F. Teacher	9	ADHD	.12	.92	.47	.51	.78	.85	
		G. Community	18	Aggression	.19	.93	.44	.47	.83	.83	
			8	Prosociality	.16	.88	.74	.84	.86	.92	
Murray, Eisner, & Ribeaud (2016)	A. SBQ	D. 1572	40	*p*	.35	.95	.67	.70	.90	.87	.69
	B. Questionnaire	E. 7	8	Internalizing	.15	.85	.80	.94	.87	.94	
	C. Item	F. Teacher	9	ADHD	.14	.93	.49	.53	.83	.88	
		G. Community	18	Aggression	.20	.93	.50	.54	.84	.86	
			8	Prosociality	.16	.89	.74	.82	.86	.92	
Murray, Eisner, & Ribeaud (2016)	A. SBQ	D. 1572	40	*p*	.35	.95	.65	.69	.89	.85	.69
	B. Questionnaire	E. 11	8	Internalizing	.17	.87	.77	.88	.86	.93	
	C. Item	F. Teacher	9	ADHD	.12	.91	.43	.47	.78	.83	
		G. Community	18	Aggression	.21	.92	.50	.55	.84	.85	
			8	Prosociality	.15	.85	.78	.92	.84	.91	
Romer et al. 2017	A. MINI + others	D. 1246	13	*p*	.34	.87	.50	.57	.74	.82	.74
	B. Interview	E. 20	5	Internalizing	.34	.87	.60	.69	.85	.92	
	C. Subscale	F. Self	5	Externalizing	.32	.83	.66	.80	.81	.90	
		G. Community									
	A. SBQ	D. 1572	40	*p*	.33	.95	.64	.67	.89	.84	.69

Murray, Eisner,	B. Questionnaire	E. 8	8	Internalizing	.16	.85	.84	.98	.88	.94	
& Ribeaud	C. Item	F. Teacher	9	ADHD	.14	.93	.49	.53	.84	.88	
(2016)		G. Community	18	Aggression	.19	.93	.44	.47	.83	.82	
			8	Prosociality	.17	.88	.82	.93	.87	.93	
Murray, Eisner,	A. SBQ	D. 1572	40	*p*	.32	.94	.66	.70	.89	.86	.69
& Ribeaud	B. Questionnaire	E. 13	8	Internalizing	.18	.87	.87	1.00	.90	.95	
(2016)	C. Item	F. Teacher	9	ADHD	.13	.93	.46	.49	.79	.83	
		G. Community	18	Aggression	.19	.92	.54	.59	.85	.87	
			8	Prosociality	.18	.89	.82	.92	.88	.94	
Murray, Eisner,	A. SBQ	D. 1572	40	*p*	.31	.94	.62	.66	.88	.84	.69
& Ribeaud	B. Questionnaire	E. 15	8	Internalizing	.19	.87	.86	.99	.89	.94	
(2016)	C. Item	F. Teacher	9	ADHD	.12	.92	.41	.45	.76	.79	
		G. Community	18	Aggression	.21	.92	.55	.59	.84	.86	
			8	Prosociality	.18	.86	.85	1.00	.86	.93	
Gibbons, Rush,	A. PDSQ	D. 3791	139	*p*	.30	.98	.83	.84	.97		.92
& Immekus	B. Questionnaire	E. 40	26	Depression	.09	.93	.66	.71	.94		
(2009)	C. Item	F. Self	7	Dysthymia	.04	.92	.71	.77	.88		
		G. Clinical	6	GAD	.02	.87	.39	.44	.69		
			11	Agoraphobia	.05	.95	.61	.64	.89		
			14	Panic Disorder	.06	.96	.49	.51	.90		
			15	Social phobia	.07	.96	.58	.61	.91		
			7	PTSD	.04	.92	.67	.73	.88		
			8	OCD	.03	.91	.49	.54	.80		
			5	Somatoform	.02	.81	.58	.71	.75		
			5	Hypochondria	.03	.94	.65	.69	.87		
			7	Alcohol	.06	.95	.91	.96	.96		
			6	Drug	.06	.97	.90	.93	.96		

10	Bulimia	.08	.96	.84	.88	.96
6	Mania	.03	.88	.74	.85	.88
6	Psychosis	.02	.86	.44	.51	.68

Note. Studies have been ordered by ECV values from highest to lowest. A = assessment measure; B = method (Questionnaire vs. Interview); C = indicator type (Item vs. Subscale); D = sample size; E = average sample age (years); F = respondent (Self; Caregiver; Teacher; Multiple); G = sample type (Clinical, Community, Population).

DP = Dysregulation profile; IS = Interpersonal sensitivity.

Measures: ADISC-IV = Anxiety Disorders Interview Schedule for Children for the DSM-IV; ADOS = Autism Diagnostic Observation Scale; AUDADIS-IV = Alcohol Use Disorder and Associated Disabilities Interview Schedule–DSM–IV Version; ASR = Adult Self Report; ATAC = Autism–Tics, AD/HD, and Other Comorbidities; BFI = Big Five Inventory; BSI = Brief Symptom Inventory; CAPE = Community Assessment of Psychic Experiences; CAPS = Child and Adolescent Psychopathology Scale; CBCL = Child Behavior Checklist; CDI = Children's Depression Inventory; CORS = Child Outcome Rating Scale; CPRS = Conner's Parent Rating Scale; CSBQ = Child Social Behaviour Questionnaire; CSI = Child Symptom Inventory; DIS = Diagnostic Interview Schedule; DAWBA = Development and Well-being Assessment; DISC-IV = Diagnostic Interview Schedule for Children-IV; EATQ-R = Aggression scale of the Early Adolescent Temperament Questionnaire-Revised; FHS = Family History Screen; GOASSESS = National Institute of Mental Health Grand Opportunity Assessment; K-SADS = Kiddie Schedule for Affective Disorders and Schizophrenia for School-Age; M&MS = Me & My School Questionnaire; MASC = Manifest Anxiety Scale for Children; MCMI = Millon Clinical Multiaxial Inventory-III; MHBQ = MacArthur Health and Behavior Questionnaire; MINI = Mini-International Neuropsychiatric Interview; PAPA = Preschool Age Psychiatric Assessment; PLIKS-Q = Psychosis-Like Symptom Questionnaire; RAPI = Rutgers Alcohol Problem Index; RCADS = Revised Child Anxiety and Depression Scale; SBQ = Social Behaviour Questionnaire; SCAN = Schedules for Clinical Assessment in Neuropsychiatry; SCARED = Screen for Anxiety-Related Emotional Disorders; SCID-II = Structured Clinical Interview Axis II for DSM-IV; SRS = Social Responsiveness Scale; SDQ = Strengths and Difficulties Questionnaire; SNAP-IV = Swanson, Nolan and Pelham Questionnaire for DSM-IV; TRF = Teacher's Rating Form; SCL-90 = Symptom Checklist-90; YSR = Youth Self-Report.

Appendix B. Standardized Factor Loadings and Model-Based Reliability Indices for Bifactor Models of the ASR Split by Fast and Slow Study Completers

	Bifactor (Fast Completers)					Bifactor (Slow Completers)				
ASR Item	*p*	INT	SOM	EXT	COG	*p*	INT	SOM	EXT	COG
Anxious/Depressed										
12. Lonely	.67	.29				.61	.21			
13. Confused	.70	.19				.71	.10[a]			
14. Cries	.70	-.12[a]				.72	-.10[a]			
22. Worries about future	.73	.12[a]				.67	.05[a]			
31. Fears thinking/doing something bad	.66	.22				.61	.24			
33. Unloved	.69	.47				.63	.39			
34. Others out to get him/her	.65	.34				.60	.27			
35. Worthless	.81	.32				.84	.21			
45. Nervous	.85	.00[a]				.84	-.09[a]			
47. Lacks self-confidence	.78	.15[a]				.79	.06[a]			
50. Fearful	.88	-.01[a]				.87	-.10[a]			
52. Too guilty	.71	.08[a]				.71	.02[a]			
71. Self-conscious	.72	.11[a]				.74	.01[a]			
91. Thinks about suicide	.64	.34				.70	.26			
103. Sad	.85	.20				.85	.15			

107. Can't succeed	.78	.30		.79	.16	
112. Worries	.85	-.13[a]		.86	-.23	
113. Worries about relations with opposite sex	.51	.40		.47	.23	
Withdrawn						
25. Doesn't get along with others	.48	.56		.40	.64	
30. Poor relations with opposite sex	.48	.43		.47	.45	
42. Would rather be alone	.42	.31		.43	.38	
48. Not liked	.59	.56		.56	.55	
60. Enjoys little	.73	.31		.73	.32	
65. Won't talk	.56	.45		.54	.27	
67. No friends	.57	.56		.55	.52	
69. Secretive	.39	.46		.45	.32	
111. Withdrawn	.53	.44		.48	.43	
Somatic Complaints						
51. Dizzy	.69		.40	.57		.43
54. Feels tired	.67		.27	.67		.23
56a. Aches	.49		.53	.47		.52
56b. Headaches	.49		.54	.41		.54
56c. Nausea	.58		.67	.62		.54
56d. Eye problems	.49		.45	.26		.47
56e. Skin problems	.32		.19	.28		.30
56f. Stomach aches	.52		.56	.51		.53
56g. Vomits	.51		.54	.52		.44

56h. Heart pounds	.69	.40		.61	.39	
56i. Numbness	.60	.55		.48	.57	
100. Sleep problems	.44	.26		.49	.34	
Aggressive Behavior						
3. Argues	.30		.48	.27		.43
5. Blames others	.43		.34	.48		.27
16. Mean	.22		.58	.23		.62
28. Gets along badly with family	.38		.37	.37		.22
37. Fights	.35		.61	.36		.64
55. Elation-depression	.72		.33	.68		.35
57. Attacks	.33		.63	.18[a]		.67
68. Screams	.49		.44	.33		.49
81. Behavior changes	.55		.46	.55		.47
86. Stubborn	.57		.38	.51		.40
87. Mood changes	.75		.35	.68		.40
95. Temper	.42		.42	.36		.50
97. Threatens	.28		.75	.31		.57
116. Upset	.78		.11[a]	.77		.05[a]
118. Impatient	.52		.39	.40		.44
Rule-Breaking Behavior						
6. Uses drugs	.14[a]		.43	.18		.38
20. Damages own things	.43		.62	.43		.43
23. Breaks rules	.13[a]		.58	.04[a]		.62
26. Lacks guilt	-.05[a]		.53	-.10[a]		.41
39. Bad companions	.27		.77	.19		.59

41. Impulsive	.29	.64		.24	.61	
43. Lies, cheats	.30	.63		.25	.53	
76. Irresponsible	.43	.65		.40	.61	
82. Steals	.23	.79		.21	.50	
90. Gets drunk	.13[a]	.49		.08[a]	.36	
92. Trouble with the law	.14[a]	.64		.21	.66	
114. Fails to pay debts	.41	.31		.36	.35	
117. Can’t manage money	.40	.37		.31	.40	
122. Can’t keep a job	.55	.27		.53	.21	
Intrusive						
7. Brags	.00[a]	.54		-.02[a]	.60	
19. Demands attention	.24	.57		.13	.59	
74. Shows off	.04[a]	.68		.05[a]	.58	
93. Talks too much	.10[a]	.42		.04[a]	.42	
94. Teases	.03[a]	.61		.04[a]	.47	
104. Loud	.08[a]	.61		.06[a]	.63	
Thought Problems						
9. Can’t get mind off thoughts	.64		.18	.62		.13
18. Harms self	.65		.28	.65		.12[a]
36. Gets hurt	.43		.38	.42		.36
40. Hears things	.51		.41	.36		.59
46. Twitches	.55		.37	.52		.20
63. Prefers older people	.32		.18	.29		.05[a]
66. Repeats acts	.40		.39	.31		.21
70. Sees things	.37		.56	.49		.45

84. Strange behaviour	.35				.50	.32				.44
85. Strange ideas	.52				.38	.43				.34
Attention Problems										
1. Forgetful	.35				.46	.34				.45
8. Can't concentrate	.52				.46	.53				.42
11. Dependent	.54				.26	.51				.18
17. Daydreams	.34				.30	.36				.31
53. Can't plan	.71				.23	.63				.27
59. Fails to finish	.54				.53	.51				.49
61. Poor work performance	.62				.45	.60				.42
64. Can't prioritize	.54				.55	.51				.53
78. Trouble with decisions	.68				.27	.63				.25
101. Avoids work	.44				.36	.40				.38
102. Lacks energy	.69				.26	.70				.16
105. Disorganized	.36				.61	.38				.73
108. Loses things	.39				.56	.41				.52
119. Poor at details	.40				.49	.42				.46
121. Tends to be late	.30				.43	.25				.44
Mean	.48	.27	.45	.51	.39	.46	.21	.44	.47	.36
Standard Deviation	.21	.20	.15	.16	.12	.22	.22	.11	.14	.17
ECV/ECV_s	.58	.06	.06	.21	.09	.60	.06	.06	.20	.09
ω/ω_s	.98	.97	.93	.96	.94	.98	.96	.91	.94	.93
ω_H/ω_{Hs}	.81	.15	.38	.68	.37	.81	.09	.41	.69	.35
Relative Omega	.83	.16	.41	.71	.40	.83	.10	.45	.73	.37

H index	.98	.79	.79	.94	.85	.98	.76	.77	.93	.84
FD	.98	.91	.92	.97	.93	.99	.91	.90	.96	.93
Inter-factor Correlations		INT	SOM	EXT	COG		INT	SOM	EXT	COG
	INT	—				INT	—			
	SOM	.10[a]	—			SOM	-.01[a]	—		
	EXT	.45	.27	—		EXT	.31	.18	—	
	COG	.55	.39	.67	—	COG	.39	.17	.63	—

Note. COG = Cognitive; ECV/ECV$_s$ = Explained Common Variance/Explained Common Variance-Subsale; Ext = Externalizing; FD = Factor Determinacy; Int = Internalizing; ω/ω_s = Omega/Omega-subsale; Som = Somatic Problems; ω_H/ω_{Hs} = Omega hierarchical/Omega hierarchical-subscale.

[a] Estimates that are not significant ($p > .05$).

Appendix C. Supplementary Tables, Figures, and Methods for Chapter 5

Table C1

Standardized Within-Person Factor Loadings for the Mood and Feelings Questionnaire

	Factor	
Scale/Item	Self-Attitudes	Mood
1. I felt miserable or unhappy.	0.35	**0.36**
2. I didn't enjoy anything at all.	0.36	**0.34**
3. I felt so tired I just sat around and did nothing.	0.02	**0.61**
4. I was very restless.	-0.01	**0.68**
5. I felt I was no good anymore.	0.69	**0.34**
6. I cried a lot.	0.65	0.14
7. I found it hard to think properly or concentrate.	0.38	0.26
8. I hated myself.	0.85	0.02
9. I was a bad person.	0.72	0.00
10. I felt lonely.	0.78	0.03
11. I thought nobody really loved me.	0.84	-0.02
12. I thought I could never be as good as other kids.	0.83	-0.08
13. I did everything wrong.	0.81	-0.05

Note: Top five items loading ≥ .32 on the mood factor are in bold and were used in the primary model.

Table C2

Within-level Standardized Factor Loadings for the Exploratory Bifactor Model (Bi-Geomin Orthogonal Rotation)

	Factor				
Scale/Item	*P*	Anxiety	Antisocial	Attention	Mood
SDQ					
3. I get a lot of headaches	0.55	0.26	-0.07	0.01	0.09
8. I worry a lot	0.65	0.44	-0.30	-0.14	-0.03
13. I am often unhappy	0.74	0.17	0.00	-0.23	0.08
16. I am nervous in new situations	0.54	0.33	**-0.37**	0.06	-0.08
24. I have many fears	0.54	0.39	-0.24	-0.13	-0.07
5. I get very angry	0.58	-0.16	0.25	0.16	0.04
7. I [do not] usually do as I am told	0.30	**-0.50**	0.32	-0.04	0.02
12. I fight a lot	0.38	0.02	0.60	0.11	-0.03
18. I often get accused of lying or cheating	0.46	-0.03	0.33	0.09	0.01
22. I take things that are not mine	0.32	-0.01	0.48	-0.04	-0.07
2. I am restless	0.48	-0.04	0.05	0.66	0.05
10. I am constantly fidgeting	0.52	-0.06	0.04	0.62	-0.02
15. I am easily distracted	0.56	-0.25	-0.04	0.44	-0.10
21. I [do not] think before I do things	0.35	**-0.52**	0.23	0.14	-0.01
25. I [do not] finish the work I am doing	0.28	**-0.58**	-0.03	0.09	-0.03
MFQ					
1. I felt miserable/unhappy	0.60	0.08	-0.05	-0.20	0.39
2. I didn't enjoy anything	0.46	-0.04	0.03	-0.18	0.54
3. I felt so tired I just sat around and did nothing	0.37	0.05	-0.07	0.06	0.54
4. I was very restless	0.45	0.02	0.07	0.23	0.51

5. I felt I was no good anymore	0.66	-0.04	0.00	-0.21	0.46

Note: Items in bold reflect cross-loadings meeting the threshold of .32. Model fit: CFI = .95, TLI = .91, RMSEA = .06, SRMR = .04.

MFQ = Mood and Feelings Questionnaire; SDQ = Strengths and Difficulties Questionnaire.

Table C3

Standardized Between-Person Regression Coefficients for the Clinical and Demographic Covariates on the General and Specific Factor Random Effects

	Estimate				
Variable	*B*	95% LL	95% UL	z	*p*
Random Intercept					
p factor					
Treatment Arm (MST vs. MAU)	0.03	-0.14	0.2	0.34	0.734
Conduct Disorder Onset (Early vs. Late)	0.04	-0.13	0.21	0.43	0.670
Region 2 vs. 1	0.14	-0.13	0.42	1.03	0.304
Region 3 vs. 1	0.09	-0.1	0.27	0.89	0.372
Age (Baseline)	-0.05	-0.11	0.02	-1.47	0.142
Sex (Boys vs. Girls)	**0.54**	**0.36**	**0.71**	**6.18**	**< .001**
SES	0.07	-0.03	0.17	1.37	0.170
FSIQ	0	-0.01	0	-0.47	0.640
Ethnicity (White British vs. Other)	-0.21	-0.44	0.02	-1.83	0.068†
Anxiety					
Treatment Arm (MST vs. MAU)	0	-0.19	0.19	0.01	0.995
Conduct Disorder Onset (Early vs. Late)	-0.01	-0.2	0.18	-0.06	0.952
Region 2 vs. 1	-0.16	-0.48	0.17	-0.93	0.351
Region 3 vs. 1	-0.03	-0.25	0.19	-0.27	0.786
Age (Baseline)	0.01	-0.06	0.08	0.19	0.850
Sex (Boys vs. Girls)	0.06	-0.15	0.26	0.53	0.595
SES	-0.01	-0.12	0.11	-0.08	0.936
FSIQ	0	-0.01	0.01	0.24	0.811
Ethnicity (White British vs. Other)	0.18	-0.08	0.45	1.37	0.169
Mood					

Treatment Arm (MST vs. MAU)	0.03	-0.15	0.22	0.36	0.718
Conduct Disorder Onset (Early vs. Late)	0	-0.18	0.19	0.02	0.981
Region 2 vs. 1	-0.16	-0.48	0.17	-0.94	0.345
Region 3 vs. 1	0.07	-0.15	0.3	0.63	0.527
Age (Baseline)	0	-0.07	0.06	-0.11	0.909
Sex (Boys vs. Girls)	0.01	-0.18	0.21	0.14	0.892
SES	-0.02	-0.13	0.09	-0.39	0.699
FSIQ	0	-0.01	0.01	0.24	0.809
Ethnicity (White British vs. Other)	0.1	-0.17	0.38	0.73	0.469
Antisocial					
Treatment Arm (MST vs. MAU)	0.01	-0.19	0.21	0.07	0.947
Conduct Disorder Onset (Early vs. Late)	0.01	-0.19	0.2	0.06	0.952
Region 2 vs. 1	0.21	-0.14	0.55	1.18	0.237
Region 3 vs. 1	0.1	-0.13	0.34	0.88	0.380
Age (Baseline)	-0.02	-0.1	0.06	-0.46	0.649
Sex (Boys vs. Girls)	-0.08	-0.28	0.13	-0.71	0.477
SES	-0.01	-0.13	0.11	-0.16	0.876
FSIQ	0	-0.01	0.01	0.24	0.807
Ethnicity (White British vs. Other)	-0.02	-0.29	0.25	-0.12	0.904
Attention					
Treatment Arm (MST vs. MAU)	0.05	-0.14	0.24	0.48	0.631
Conduct Disorder Onset (Early vs. Late)	0.02	-0.17	0.21	0.21	0.836
Region 2 vs. 1	0.22	-0.1	0.53	1.36	0.175
Region 3 vs. 1	0.07	-0.14	0.28	0.64	0.520
Age (Baseline)	-0.04	-0.11	0.03	-1.09	0.277
Sex (Boys vs. Girls)	-0.03	-0.22	0.17	-0.26	0.795
SES	0.01	-0.1	0.13	0.19	0.852
FSIQ	-0.01	-0.01	0	-1.61	0.109
Ethnicity (White British vs. Other)	-0.09	-0.35	0.17	-0.66	0.512

Random Linear Slope					
p factor					
Treatment Arm (MST vs. MAU)	0.12	-0.13	0.37	0.92	0.358
Conduct Disorder Onset (Early vs. Late)	-0.04	-0.28	0.2	-0.33	0.743
Region 2 vs. 1	-0.09	-0.51	0.33	-0.41	0.681
Region 3 vs. 1	0.1	-0.18	0.38	0.71	0.478
Age (Baseline)	0.08	-0.01	0.18	1.68	0.093†
Sex (Boys vs. Girls)	**-0.27**	**-0.53**	**-0.02**	**-2.08**	**0.038**
SES	-0.13	-0.27	0.02	-1.68	0.093†
FSIQ	0	-0.01	0.01	0.1	0.920
Ethnicity (White British vs. Other)	0.13	-0.23	0.48	0.68	0.495
Anxiety					
Treatment Arm (MST vs. MAU)	-0.02	-0.31	0.27	-0.14	0.890
Conduct Disorder Onset (Early vs. Late)	-0.01	-0.3	0.28	-0.05	0.959
Region 2 vs. 1	0.01	-0.49	0.5	0.03	0.977
Region 3 vs. 1	0.04	-0.31	0.39	0.22	0.823
Age (Baseline)	0.02	-0.1	0.13	0.29	0.774
Sex (Boys vs. Girls)	0.08	-0.23	0.39	0.49	0.625
SES	-0.04	-0.22	0.14	-0.4	0.692
FSIQ	0	-0.01	0.01	0.34	0.731
Ethnicity (White British vs. Other)	-0.06	-0.46	0.35	-0.29	0.774
Antisocial					
Treatment Arm (MST vs. MAU)	-0.03	-0.31	0.26	-0.18	0.858
Conduct Disorder Onset (Early vs. Late)	0.04	-0.25	0.34	0.29	0.771
Region 2 vs. 1	0.38	-0.12	0.89	1.48	0.138
Region 3 vs. 1	0.19	-0.15	0.53	1.1	0.270
Age (Baseline)	0.02	-0.09	0.13	0.28	0.778
Sex (Boys vs. Girls)	0.01	-0.3	0.31	0.03	0.973
SES	-0.04	-0.21	0.13	-0.44	0.657
FSIQ	0	-0.01	0.01	0.55	0.585

Ethnicity (White British vs. Other)	-0.01	-0.42	0.41	-0.03	0.979
Attention					
Treatment Arm (MST vs. MAU)	-0.01	-0.32	0.3	-0.07	0.947
Conduct Disorder Onset (Early vs. Late)	0	-0.3	0.31	0.02	0.987
Region 2 vs. 1	-0.19	-0.72	0.34	-0.7	0.481
Region 3 vs. 1	-0.15	-0.51	0.21	-0.8	0.425
Age (Baseline)	-0.01	-0.14	0.11	-0.21	0.835
Sex (Boys vs. Girls)	-0.06	-0.39	0.27	-0.36	0.717
SES	0.04	-0.15	0.22	0.38	0.701
FSIQ	0	-0.01	0.01	-0.07	0.946
Ethnicity (White British vs. Other)	0.07	-0.37	0.5	0.3	0.765
Mood					
Treatment Arm (MST vs. MAU)	-0.1	-0.37	0.18	-0.68	0.496
Conduct Disorder Onset (Early vs. Late)	-0.07	-0.34	0.2	-0.52	0.606
Region 2 vs. 1	-0.15	-0.61	0.31	-0.64	0.522
Region 3 vs. 1	-0.05	-0.36	0.25	-0.34	0.731
Age (Baseline)	0	-0.1	0.1	0.01	0.994
Sex (Boys vs. Girls)	-0.02	-0.3	0.26	-0.13	0.896
SES	-0.04	-0.21	0.12	-0.53	0.595
FSIQ	0	-0.01	0.01	0.11	0.910
Ethnicity (White British vs. Other)	0.11	-0.28	0.49	0.54	0.589
Random Quadratic Slope					
p factor					
Treatment Arm (MST vs. MAU)	-0.04	-0.12	0.05	-0.86	0.387
Conduct Disorder Onset (Early vs. Late)	0.01	-0.07	0.09	0.2	0.841
Region 2 vs. 1	0.03	-0.11	0.17	0.41	0.686
Region 3 vs. 1	-0.03	-0.12	0.07	-0.54	0.589
Age (Baseline)	-0.03	-0.06	0	-1.78	0.074†
Sex (Boys vs. Girls)	0.07	-0.02	0.16	1.58	0.115

SES	0.04	-0.01	0.09	1.43	0.152
FSIQ	0	0	0	-0.15	0.885
Ethnicity (White British vs. Other)	-0.02	-0.14	0.09	-0.38	0.706
Anxiety					
Treatment Arm (MST vs. MAU)	0.01	-0.08	0.11	0.25	0.802
Conduct Disorder Onset (Early vs. Late)	0	-0.09	0.1	0.08	0.934
Region 2 vs. 1	0.02	-0.14	0.18	0.2	0.839
Region 3 vs. 1	0	-0.11	0.11	-0.03	0.975
Age (Baseline)	-0.01	-0.04	0.03	-0.26	0.798
Sex (Boys vs. Girls)	-0.01	-0.11	0.09	-0.19	0.851
SES	0.01	-0.05	0.07	0.31	0.755
FSIQ	0	0	0	-0.48	0.631
Ethnicity (White British vs. Other)	0	-0.13	0.13	0.01	0.995
Antisocial					
Treatment Arm (MST vs. MAU)	0.01	-0.09	0.1	0.12	0.906
Conduct Disorder Onset (Early vs. Late)	-0.02	-0.12	0.07	-0.45	0.654
Region 2 vs. 1	-0.12	-0.28	0.04	-1.44	0.150
Region 3 vs. 1	-0.06	-0.17	0.05	-1.12	0.262
Age (Baseline)	0	-0.04	0.03	-0.14	0.889
Sex (Boys vs. Girls)	0.02	-0.08	0.11	0.3	0.761
SES	0.02	-0.04	0.07	0.63	0.530
FSIQ	0	0	0	-0.73	0.463
Ethnicity (White British vs. Other)	0	-0.13	0.13	0.01	0.996
Attention					
Treatment Arm (MST vs. MAU)	0	-0.1	0.1	0.06	0.949
Conduct Disorder Onset (Early vs. Late)	0.01	-0.09	0.1	0.11	0.909
Region 2 vs. 1	0.05	-0.13	0.22	0.54	0.590
Region 3 vs. 1	0.04	-0.08	0.16	0.7	0.486
Age (Baseline)	0	-0.04	0.04	0.21	0.836
Sex (Boys vs. Girls)	0.02	-0.08	0.13	0.41	0.681

SES	-0.01	-0.07	0.05	-0.39	0.698
FSIQ	0	0	0	-0.02	0.984
Ethnicity (White British vs. Other)	-0.02	-0.15	0.12	-0.21	0.836
Mood					
Treatment Arm (MST vs. MAU)	0.03	-0.06	0.11	0.62	0.539
Conduct Disorder Onset (Early vs. Late)	0.03	-0.06	0.11	0.57	0.566
Region 2 vs. 1	0.03	-0.12	0.18	0.4	0.689
Region 3 vs. 1	0.01	-0.09	0.11	0.14	0.889
Age (Baseline)	0	-0.03	0.03	0.14	0.886
Sex (Boys vs. Girls)	0.01	-0.09	0.1	0.13	0.893
SES	0.01	-0.04	0.07	0.48	0.633
FSIQ	0	0	0	0.38	0.708
Ethnicity (White British vs. Other)	-0.04	-0.16	0.08	-0.68	0.495

Note. LL = Lower Limit; UL = Upper Limit; Significant coefficients are in bold (p < .05).

†*Marginal result (p < .1)*

Table C4

Within-level Standardized Factor Loadings for a Confirmatory Bifactor Model Without Cross-loadings

	Factor				
Scale/Item	*p*	Anxiety	Antisocial	Attention	Mood
SDQ					
3. I get a lot of headaches	0.49***	0.34***			
8. I worry a lot	0.46***	0.63***			
13. I am often unhappy	0.62***	0.40***			
16. I am nervous in new situations	0.42***	0.38***			
24. I have many fears	0.34***	0.59***			
5. I get very angry	0.67***		0.22***		
7. I [do not] usually do as I am told	0.35***		0.29***		
12. I fight a lot	0.37***		0.57***		
18. I often get accused of lying or cheating	0.48***		0.35***		
22. I take things that are not mine	0.27***		0.55***		
2. I am restless	0.47***			0.64***	
10. I am constantly fidgeting	0.51***			0.63***	
15. I am easily distracted	0.55***			0.48***	
21. I [do not] think before I do things	0.39***			0.27***	
25. I [do not] finish the work I am doing	0.28***			0.28***	
MFQ					
1. I felt miserable/unhappy	0.54***				0.47***
2. I didn't enjoy anything	0.45***				0.60***
3. I felt so tired I just sat around and did nothing	0.41***				0.47***

4. I was very restless	0.52***				0.35***
5. I felt I was no good anymore	0.66***				0.50***
M	0.46	0.47	0.40	0.46	0.48
SD	0.11	0.13	0.16	0.18	0.09
ECV/ECV$_s$	0.51	0.13	0.10	0.13	0.13
ω/ω_s	0.91	0.80	0.73	0.78	0.83
ω_H/ω_{Hs}	0.73	0.40	0.34	0.41	0.39
Relative Omega	0.81	0.50	0.46	0.52	0.46
H index	0.86	0.63	0.54	0.64	0.62
FD	0.88	0.79	0.73	0.81	0.77

Note: ECV = Explained Common Variance; *M* = mean; MFQ = Mood and Feelings Questionnaire; SD = standard deviation; SDQ = Strengths and Difficulties Questionnaire; ω = Omega; ω_h = Omega hierarchical.

$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

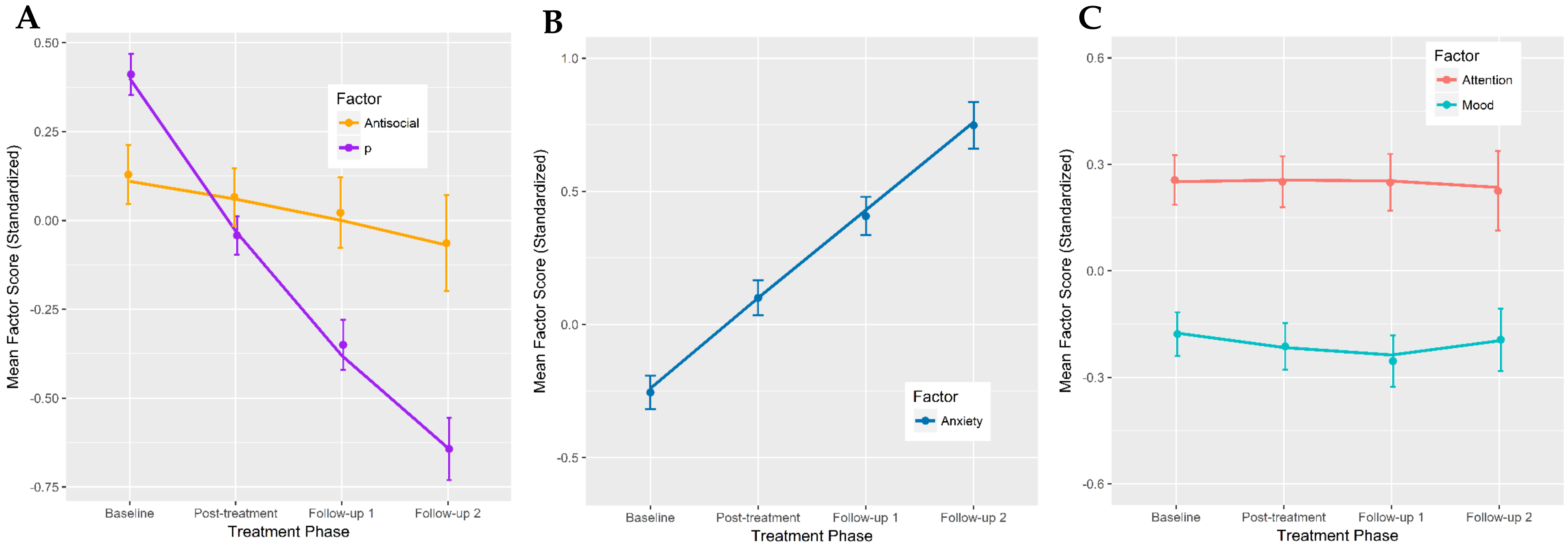

Figure C1. Average predicted trajectories (curves) and observed means (data points with error bars) for (A) general psychopathology and specific antisocial factors, (B) specific anxiety factor, and (C) specific mood and attention factors, estimated without cross-loadings. The 0 point reflects the factor mean. Error bars indicate 95% CIs.

Supplement C1. Data Quality Checks Background

Missing Data. It is rare to find an applied longitudinal study that is free from missing data. Some patients drop-out after being randomized but before baseline measures are taken, perhaps because they lose interest, fear the commitment, or find an alternative treatment. Others drop-out during the study because they find the trial does not meet their expectations, their presentation was hard to manage from the outset, or because they feel they have benefitted and no longer require the intervention.

Researchers might try to over-sample for participants to reduce the threat of missingness on statistical power and parameter estimates. But the consequences of missing data are more widespread: participants who drop-out often differ in clinical characteristics (e.g., illness severity, willingness for treatment) and demographic characteristics (Enders, 2011). Therefore, analyses on compliers will reflect a biased selection of the population that ultimately threatens the reliability and validity of the findings (Little, Jorgensen, Lang, & Moore, 2014). For example, participants with complete data may be less severe and hence respond better to treatment, giving a false impression that a given intervention was effective.

There are three main missing data mechanisms (Allison, 2001). When observations are 'Missing Completely at Random' (MCAR), the missing data patterns are not explainable by other variables, including the variable with missing data. In other words, there is no systematic reason why the observations are missing; the likelihood of not observing some data is equal to the likelihood of observing it. This might occur when a participant misses certain items on a questionnaire or when data is accidentally lost. 'Missing at Random' (MAR) occurs when the missing observations can be explained by known or unknown variables excluding the variable with missing data. For instance, there might be some questions about mental health during the menstrual cycle that men cannot answer; in this case, the MAR mechanisms is known (e.g., sex). Finally, 'Missing Not at Random' (MNAR) occurs when the likelihood of missingness depends on the variable itself. For example, the likelihood of missing data on a depression scale may be increased in the highest scoring patients.

One way of testing the extent to which the missing data are MCAR vs. MAR is by comparing the parameter estimates between a growth model that assumes the data are MCAR, i.e. a complete case analysis, with a model that assumes the data are missing at random, i.e. a model that includes predictors of missingness other than the outcome variable itself (Little et al., 2012). Substantial differences between estimates (e.g., differences > 10% of the MCAR estimate) suggest that the unobserved data is influenced by observed or unobserved variables, but estimates might not be biased.

The assumption of MAR vs. MNAR cannot be formally tested as this would require the very outcome data that is missing to compare with the available data (Allison, 2001). Hence, we can only test the extent to which MAR vs. MNAR is plausible. One way of inferring whether missingness depends on the unobserved values is by estimating the unobserved values with pattern mixture models or selection models. In pattern mixture models, a model is estimated for each pattern of drop-out (e.g., dummy codes are created for participants who drop out after the baseline phase, middle phase, or final phase) and estimates from each model are 'mixed' to form a weighted model that takes into account different missing data patterns (O'Kelly & Ratitch, 2014). In selection models, one adds survival indicators to the model that code for the time until dropout, rather than stratifying the sample by dropout patterns. The advantage of selection models is that one can explicitly test the influence of past values of the outcome variable and values estimated at the time of dropout on model estimates (Diggle & Kenward, 1994). Significant survival variables suggest that missingness is dependent on unobserved data and thus the missing data are potentially MNAR.

After making plausible hypotheses about the missing data mechanisms and their impact, the next step is to identify the optimal method for handle missing data (Little et al., 2012). The 'gold standard' missing data methods are multiple imputation (MI) and full-information maximum likelihood (FIML; Graham, 2009). With MI, missing values are replaced with predicted values sampled from a random distribution. The model is run with each replica data set separately and results are combined in a way that factors in the uncertainty associated with the true unobserved values. Hence, the aim of MI is to simulate various possible values rather than to 'fill in' the most likely values (Johnson & Young, 2011).

FIML uses all available data for a case, including partially available data, to estimate their likelihood function (i.e. the likelihood of estimating certain parameters for a participant given the available data; Graham, 2009). In other words, FIML uses complete data of other variables, as well as partially available data through its correlations with the complete data, to maximise the likelihood function for each participant. In turn, the most likely parameters are estimated for both observed and unobserved data points without requiring the latter.

Both MI and FIML have similar assumptions (e.g., both assume the missing data mechanisms are MAR and that the data are normally distributed) and yield unbiased parameter estimates and small standard errors (Acock, 2005; Enders, 2011; Johnson & Young, 2011). However, some studies show that FIML outperforms MI with non-normal data, multilevel designs, and small sample sizes typical of clinical studies (Larsen, 2011; Shin, Davison, & Long, 2017). Full-information maximum likelihood is used throughout this thesis to handle missing data.

Measurement Invariance. An advantage of multilevel factor analysis is that fewer parameters are needed to estimate factors expressed at the individual level since the analysis is collapsed over time rather than repeatedly at each time-point (Wright et al., 2016). A disadvantage is that it is not possible to test for measurement invariance, i.e. the extent to which within-person change is driven by changes in measurement properties (e.g. differential item functioning or response biases) rather than the factors (Cheung & Rensvold, 2002).

In conventional measurement invariance tests, factors are repeatedly estimated at each time-point and model parameters are compared when freely estimated or held constant across time-points (Liu et al., 2017). This is not possible in multilevel models because 'time' is an inherent feature of model parameters, e.g., a within-level factor loading reflects the way in which an item is predicted to covary with other items across time. Instead, factor loadings and item intercepts/thresholds are assumed to be invariant. For example, an item intercept is the mean of that item over the within-level (e.g., time) when a given factor equals zero.

Longitudinal measurement invariance was tested using the conventional method (e.g., estimating a separate bifactor model at each time-point individually), even though this would demonstrate properties of the parameters that are not immediately transferable to the multi-level method. A factor loading in one model is not the same as a factor loading in the other. Moreover, full or partial invariance shown using the conventional approach cannot be carried over to the multilevel model, since there are no parameters to hold constant. That said, the results of both single-level and multi-level growth models should ultimately converge (Curran, Obeidat, & Losardo, 2010), and so invariance observed using one approach should roughly translate to the other.

Metric invariance (e.g., equal factor loadings between the adjacent time-points) was tested using Wald chi-square tests via the MODEL CONSTRAINT command in Mplus 8.0. All factor loadings showed metric invariance except for those associated with the mood factor between time 2 (post-treatment) and time 3 (6-months follow-up; $\chi2(4) = 11.54$, $p = .021$).

Scalar invariance (e.g., equal item intercepts or thresholds between adjacent time-points) was tested by comparing individual item thresholds between two adjacent time-points using Wald chi-square tests, while simultaneously testing for differences among all factor loadings (the latter was intended to mimic equality constraints on all factor loadings, which is a prerequisite when testing scalar invariance). Each of the 20 items had two thresholds (threshold A and B) which were compared at three adjacent time-points (time 1 vs. time 2, time 2 vs. time 3, time 3 vs. time 4), resulting in 120 tests. To minimize family-wise error rates, the alpha level was corrected for the number of tests conducted on a single threshold between two adjacent time-points using the Bonferroni method (e.g., α/k, where α is the type I error rate and k is the number of tests). Therefore, $\alpha = .003$ ($\alpha/k = .05/20$) when testing the equivalence of one of the two thresholds for each of the 20 items between two adjacent time-points.

Threshold A was invariant for 80% of items between time 1 and 2, while threshold B was invariant for 60% of items. Between time 2 and 3, threshold A was invariant for 90% of items, while threshold B was invariant for 95% of items. Finally, 100% of items showed invariance in threshold A and B between time 3 and 4. Non-invariance of item thresholds was thus mainly apparent between time 1 (baseline) and 2 (post-treatment), which may be because pre-treatment distributions can deviate from post-treatment distributions (Hedeker & Gibbons, 2006). Three of the nine items (33%) that showed non-invariance in threshold A between time 1 and 2 also showed non-invariance in threshold B (e.g., SDQ items 5 and 12, and MFQ item 5). Therefore, the majority of non-invariance in item intercepts was sporadic rather than systematic.

In all, the conventional measurement invariance analysis demonstrates partial longitudinal measurement invariance, but caution is warranted when extending these findings to the multilevel model.

Supplement C2. Modelling Background

Multilevel Factor Analysis. Multilevel factor analysis is typically used to estimate separate factor structures for the within-person and between-person portions a hierarchically clustered covariance matrix (Muthén, 1994; 1991). For instance, researchers might investigate the factor structure of a new measure of societal well-being and collect data from thousands of individuals across 52 countries. A multilevel factor analysis could be used to test the factor structure at the individual-level (i.e. within-country effects) and the country-level (i.e. between-country effects). The researchers might be interested in differences in the factor structure at different levels, or they might simply focus on one level while correcting the standard errors for nesting between levels.

The START trial data is hierarchically organized into two main levels[30]: repeated observations (within-person level) nested within each individual (between-person level). We can therefore apply multilevel factor analysis to determine the within-person factor structure (i.e. factors that account for the covariation between symptoms within a given individual) and between-person factor structure (i.e. factors that account for the covariation in symptoms between individuals).

A multilevel CFA can be expressed as follows:

$$Y_{ijt} = v_{W_{ijt}} + \Lambda_{\mathrm{W}} \eta_{\mathrm{W}_{ijt}} + \varepsilon_{\mathrm{W}_{ijt}},$$

[30] Technically, the START data is organized into at least four-levels: repeated measures nested within persons nested within site nested within region, but there are an insufficient number of data points and predictors at the higher levels to estimate this.

where Y is a matrix reflecting the observed responses on each item, $j = 1,\ldots,J$, at each time-point, $t = 1,\ldots,T$ across individuals, $i = 1,\ldots,N$, $v_{W_{ijt}}$is a vector of within-person item thresholds; Λ_{W} is a within-person factor loading matrix, $\eta_{\mathrm{W}_{ijt}}$ is a vector of factors which vary randomly across time-points and items within individuals, and $e_{ijt}$is the within-person error. The $\Lambda_{\mathrm{W}}\eta_{\mathrm{W}_{ijt}}$term can be expressed more fully in the context of a bifactor model as:

$$\Lambda_{\mathrm{W}}\eta_{\mathrm{W}_{ijt}} = \lambda_{Wgeneral_j}\theta_{Wgeneral_{it}} + \lambda_{Wspecific1_j}\theta_{Wspecific1_{it}} + \lambda_{Wspecific2_j}\theta_{Wspecific2_{it}} + \lambda_{Wspecific3_j}\theta_{Wspecific3_{it}} + \lambda_{Wspecific4_j}\theta_{Wspecific4_{it}},$$

where $\lambda_{W_j}$are within-person factor loadings for each item and $\theta_{W_{it}}$ are within-person factor vectors which vary across individuals and time-points for the general factor, $general$, and specific factors, $specific1, \ldots, specificK$, where $K = 4$ in the current model.[31]

Factor Score Estimation. Like factor scores, BPVs are observed estimates of latent variables, but unlike factor scores, they also take into account the uncertainty or 'indeterminacy' in estimating factor scores by averaging over a distribution of possible factor scores using multiple imputation (Aitkin & Aitkin, 2005; Mislevy, 1991). Theoretical and simulation studies suggest that BPVs provide less biased estimates of population parameters than factor scores (von Davier, Gonzalez, & Mislevy, 2009; Wu, 2005). In practice, BPVs and factor scores probably produce

[31]The reader might recognise this as a three-level notation, with repeated observations at the lowest level ('time') nested in each item ('item'), nested within individuals ('subject'). However, when implementing the model in Mplus, each item was included as a different within-level variable (see Table 5.2), making it a multi-indicator two-level multilevel factor model. Nonetheless, the models are equivalent.

similar estimates when sample sizes are sufficient (Lüdtke, Robitzsch, & Trautwein, 2018; Marsman, Maris, Bechger, & Glas, 2016). BPVs were used instead of factor scores due to the unreliability of specific factors after accounting for the general factor. It is important to incorporate estimates of imprecision when using factor scores in secondary analyses (Laukaityte & Wiberg, 2017; Skrondal & Laake, 2001).

BPVs were estimated using the same multi-level growth model in the main analysis to minimise bias (Mislevy, 1991). There is little consensus over how many imputations to estimate. While Asparouhov and Muthén (2010) suggest that five imputations are sufficient for secondary analyses, one-hundred imputations were estimated with a thinning rate of 1 (e.g., random estimates were sampled on every iteration).

Multilevel Growth Model. Within-person changes in BPVs and between-person differences in within-person change were analysed using a parallel process multilevel growth model (Hoffman, 2007). The within-person part of the growth model can be written as:

$$y_{it}^{(f)} = \beta_{0_i}^{(f)} + \beta_{1_i}^{(f)} Time_{it} + \beta_{2_i}^{(f)} Time^2{}_{it} + \varepsilon_{it}^{(f)},$$

where $y_{it}^{(f)}$ reflects BPVs for each individual, i = 1 ,…, N at each time-point, t = 0 ,…, T for a given factor, $\beta_{0_i}^{(f)}$ reflects the intercept or baseline factor scores for each individual when t = 0 (for each factor), $\beta_{1_i}^{(f)}$ and $\beta_{2_i}^{(f)}$reflect the linear and quadratic slopes of time on each factor, respectively, which vary randomly across individuals, $Time_{it}$ and $Time^2{}_{it}$ reflect the observed values of time (0, 1, 2, 3) and time-squared (0, 1, 4, 9) for each individual at each time-point, and $\varepsilon_{it}^{(f)}$ reflects the individual- and time-specific residuals.

The between-person part of the growth model can be expressed as:

$$\beta_{0_i}^{(f)} = \gamma_{00}^{(f)} + U_{0i}^{(f)},$$

$$\beta_{1_i}^{(f)} = \gamma_{10}^{(f)} + U_{1i}^{(f)},$$

$$\beta_{2_i}^{(f)} = \gamma_{20}^{(f)} + U_{2i}^{(f)},$$

where $\gamma_{00}^{(f)}$, $\gamma_{10}^{(f)}$, and $\gamma_{20}^{(f)}$ are the overall mean intercept, mean linear slope of time, and mean quadratic slope of time, respectively, across individuals for each factor, and $U_{0i}^{(f)}$, $U_{1i}^{(f)}$, and $U_{2i}^{(f)}$ reflect person-specific deviations from the overall intercept, linear slope of time, and quadratic slope of time, respectively, for each factor.

The between-person portion of a growth model with clinical and demographic covariates would be expressed as:

$$\beta_{0_i}^{(f)} = \gamma_{00}^{(f)} + \gamma_{01}^{(f)} cov_i + \gamma_{02}^{(f)} cov2_i + \gamma_{03}^{(f)} cov3_i + \cdots \gamma_{0K}^{(f)} covK_i + U_{0i}^{(f)},$$

$$\beta_{1_i}^{(f)} = \gamma_{10}^{(f)} + \gamma_{11}^{(f)} cov_i + \gamma_{12}^{(f)} cov2_i + \gamma_{13}^{(f)} cov3_i + \cdots \gamma_{1K}^{(f)} covK_i U_{1i}^{(f)},$$

$$\beta_{2_i}^{(f)} = \gamma_{20}^{(f)} + \gamma_{21}^{(f)} cov_i + \gamma_{22}^{(f)} cov2_i + \gamma_{23}^{(f)} cov3_i + \cdots \gamma_{2K}^{(f)} covK_i + U_{2i}^{(f)},$$

where $\gamma_{0K}^{(f)}$,$\gamma_{1K}^{(f)}$, and $\gamma_{2K}^{(f)}$ are the effect of between-person differences in a given covariate, '$covK$', on the intercept, linear time slope, and quadratic time slope for each factor. In other words, they reflect how individual differences in covariates predict variation in baseline factor scores, linear growth curves, and quadratic growth curves, respectively.

The covariance structure for the random effects across factors was unrestricted in all models. That is, the covariances between the random intercepts, linear slopes, and quadratic slopes were freely estimated for each factor, as well as between factors, forming a 15 x 15 unrestricted covariance matrix:

$$V\begin{bmatrix} U_{0i}^{(1)} \\ U_{1i}^{(1)} \\ U_{2i}^{(1)} \\ U_{0i}^{(2)} \\ U_{1i}^{(2)} \\ U_{2i}^{(2)} \\ \vdots \\ U_{2i}^{(5)} \end{bmatrix} = \begin{bmatrix} \tau_{00}^{(1)} & & & & & & & \\ \tau_{10}^{(1)} & \tau_{11}^{(1)} & & & & & & \\ \tau_{20}^{(1)} & \tau_{21}^{(1)} & \tau_{22}^{(1)} & & & & & \\ \tau_{00}^{(2,1)} & \tau_{01}^{(2,1)} & \tau_{02}^{(2,1)} & \tau_{00}^{(2)} & & & & \\ \tau_{10}^{(2,1)} & \tau_{11}^{(2,1)} & \tau_{12}^{(2,1)} & \tau_{10}^{(2)} & \tau_{11}^{(2)} & & & \\ \tau_{20}^{(2,1)} & \tau_{21}^{(2,1)} & \tau_{22}^{(2,1)} & \tau_{20}^{(2)} & \tau_{21}^{(2)} & \tau_{22}^{(2)} & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ \tau_{20}^{(5,1)} & \tau_{21}^{(5,1)} & \tau_{22}^{(5,1)} & \tau_{20}^{(5,2)} & \tau_{21}^{(5,2)} & \tau_{22}^{(5,2)} & \dots & \tau_{22}^{(5)} \end{bmatrix}$$

Appendix C References

Acock, A. C. (2005). Working With Missing Values. *Journal of Marriage and Family, 67*(4), 1012-1028. doi:10.1111/j.1741-3737.2005.00191.x

Aitkin, M. & Aitkin, I. (2005). Bayesian inference for factor scores. In A. Maydeu-Olivares, & J. J. McArdle. (Eds.). *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald*, (pp. 207-22). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Allison, P. D. (2001). *Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.

Asparouhov, T., & Muthén, B. (2010). *Plausible values for latent variables using Mplus*. Unpublished manuscript, available at http://www. statmodel. com/download/Plausible. pdf.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255. doi:10.1207/s15328007sem0902_5

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve Frequently Asked Questions About Growth Curve Modeling. *J Cogn Dev, 11*(2), 121-136. doi:10.1080/15248371003699969

Diggle, P., & Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Applied Statistics, 43*(1), 49. doi:10.2307/2986113

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychol Methods, 16*(1), 1-16. doi:10.1037/a0022640

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annu Rev Psychol, 60*, 549-576. doi:10.1146/annurev.psych.58.110405.085530

Hedeker, D., & Gibbons, R. D. (2006). *Wiley Series in Probability and Statistics. Longitudinal data analysis.* Hoboken, NJ, US: Wiley-Interscience.

Hoffman, L. (2008). Multilevel Models for Examining Individual Differences in Within-Person Variation and Covariation Over Time. *Multivariate Behavioral Research, 42*(4), 609-629. doi:10.1080/00273170701710072

Johnson, D. R., & Young, R. (2011). Toward Best Practices in Analyzing Datasets with Missing Data: Comparisons and Recommendations. *Journal of Marriage and Family, 73*(5), 926-945. doi:10.1111/j.1741-3737.2011.00861.x

Larsen, R. (2011). Missing Data Imputation versus Full Information Maximum Likelihood with Second-Level Dependencies. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(4), 649-662. doi:10.1080/10705511.2011.607721

Laukaityte, I., & Wiberg, M. (2016). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics - Theory and Methods, 46*(22), 11341-11357. doi:10.1080/03610926.2016.1267764

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., ... & Neaton, J. D. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355-1360.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. (2014). On the joys of missing data. *J Pediatr Psychol, 39*(2), 151-162. doi:10.1093/jpepsy/jst048

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychol Methods, 22*(3), 486-506. doi:10.1037/met0000075

Ludtke, O., Robitzsch, A., & Trautwein, U. (2018). Integrating Covariates into Social Relations Models: A Plausible Values Approach for Handling Measurement

Error in Perceiver and Target Effects. *Multivariate Behav Res, 53*(1), 102-124. doi:10.1080/00273171.2017.1406793

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from Plausible Values? *Psychometrika, 81*(2), 274-289. doi:10.1007/s11336-016-9497-x

Mislevy R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika. 56*(2), 177-196.

Muthen, B. O. (1991). Multilevel Factor Analysis of Class and Student Achievement Components. *Journal of Educational Measurement, 28*(4), 338-354. doi:10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological methods & research*, *22*(3), 376-398.

O'Kelly, M., & Ratitch, B. (2014). Analyses under Missing-not-at-random Assumptions. 257-368. doi:10.1002/9781118762516.ch7

Shin, T., Davison, M. L., & Long, J. D. (2017). Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. *Psychol Methods, 22*(3), 426-449. doi:10.1037/met0000094

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*(4), 563-575. doi:10.1007/bf02296196

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series*. *2*, 9-36.

Wright, A. G., Hopwood, C. J., Skodol, A. E., & Morey, L. C. (2016). Longitudinal validation of general and specific structural features of personality pathology. *J Abnorm Psychol, 125*(8), 1120-1134. doi:10.1037/abn0000165

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128. doi:10.1016/j.stueduc.2005.05.005

Appendix D. Background to Latent Growth Curve Models

The aim of growth curve modelling is to describe how individuals change on processes of interest over a given timeframe (Grimm, Ram, & Estabrook, 2016). Growth models typically summarize where a group of individuals begin and how each of them changes over time with two main parameters: an intercept, g_{0i}[32], which describes a pooled estimate of each person's average on the outcome variable when time, $t = 1 \dots T$, equals 0 (i.e. at baseline or the 0 value for a centred time variable), and a slope, g_{1i}, which reflects a pooled estimate of each person's rate and direction of change on the outcome variable over time (e.g., linear increase, decrease, or flat slope). However, growth models differ in how they parameterize the intercept and slope.

Multilevel approaches to growth modelling were introduced in Chapter 5 and detailed in Appendix C. To recap, multilevel growth models partition the variance in the outcome variable into different levels of hierarchically clustered data. For example, 'time' is explicitly modelled as a variable at level 1 (i.e. the within-person level) and predicts where the outcome variable starts (e.g., when t = 0) and how it changes for each individual across time-points. Interindividual

[32] Notice that I have described the 'pooled' estimates of parameters for each person, i, rather than group means, e.g., β_0, β_1. This is because the parameters vary across individuals using random effects (e.g., $g_{0i} = \gamma_{00} + u_{0i}$ and $g_{1i} = \gamma_{10} + u_{1i}$). Therefore, each parameter has a mean and variance, allowing us to look at intra-individual change and inter-individual differences in change, respectively. This differs from traditional approaches for analysing repeated measures, such as multivariate analysis of variance, which treat each measurement occasion as independent, and hence the variances as noise (Curran et al., 2010). Growth models thus provide a powerful tool for charting individual differences in growth to describe "how and why individuals follow different paths of development" (Ram & Grimm, 2007).

differences in the intercept and slope are modelled as random effects at level 2 (i.e. the between-person level).

In Latent Growth Curve Models (LGCMs), the intercept and slope are modelled as latent variables (Duncan & Duncan, 2009). The model can be summarized as follows:

$$y_{it} = \lambda_{t0}\eta_{0i} + \lambda_{t1}\eta_{1i} + \varepsilon_{ti}$$

$$\eta_{0i} = \alpha_0 + \zeta_{0i}$$

$$\eta_{0i} = \alpha_1 + \zeta_{1i}$$

These equations might remind the reader of the single-factor model described in Chapter 1. This is no coincidence: the outcome variable at each timepoint is regressed or 'loads' onto a latent intercept factor ('$\lambda_{t0}\eta_{t0}$') and a latent slope factor ('$\lambda_{t1}\eta_{t1}$'), which each have a mean (e.g., α_0, α_1) and variance (e.g., ζ_{0i}, ζ_{1i}). The intercept factor loadings are constrained to one (e.g., $\lambda_{t0} = 1, 1, 1, 1$ for a four-wave study) to represent the fact that the intercept factor influences the outcome variable in the same way over time. The slope factor loadings are typically fixed to values that reflect the intervals between waves (e.g., $\lambda_{t0} = 0, 1, 2, 3$). In other words, the slope factor represents the predicted value of the outcome variable when time equals t, where $t = 0, 1, 2, 3, \ldots T - 1$). By contrast, the intercept is the predicted outcome when $t = 0$, i.e. at baseline.

Covariates that are usually constant over time (e.g., a participants' biological sex) and those that change over time (e.g., mood) can also be used to predict variation in the intercept and slope factors, for instance:

$$\eta_{0i} = \alpha_0 + \gamma_0 W_i + \zeta_{0i},$$

$$\eta_{1i} = \alpha_1 + \gamma_1 W_i + \zeta_{1i},$$

where $\gamma_0 W_i$ and $\gamma_1 W_i$ describe the influence of a time-varying or time-invariant covariate on the mean intercept and slope estimates, respectively, pooled across participants.

Latent growth models and multilevel growth models are more similar than different, and their results should ultimately converge (Curran et al., 2010). Many of the reasons why one would choose a latent growth model over a multilevel growth model have become obsolete with technological innovation (Hox & Stoel, 2005). However, one might use a multilevel model for computational efficiency (see Chapter 5) and a latent growth model to integrate development within a broader structural equation model (see Chapter 6).

Appendix D References

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve Frequently Asked Questions About Growth Curve Modeling. *J Cogn Dev, 11*(2), 121-136. doi:10.1080/15248371003699969

Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An Introductory Guide to Latent Variable Growth Curve Modeling. *Soc Personal Psychol Compass, 3*(6), 979-991. doi:10.1111/j.1751-9004.2009.00224.x

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Hox, J., & Stoel, R. D. (2014). Multilevel and SEM Approaches to Growth Curve Modeling. *Wiley StatsRef: Statistics Reference Online*. doi:10.1002/9781118445112.stat06603

Ram, N., & Grimm, K. (2007). Using simple and complex growth models to articulate developmental change: Matching theory to method. *International Journal of Behavioral Development, 31*(4), 303-316. doi:10.1177/0165025407077751