Three Recommendations Based on a Comparison of the Reliability and Validity of the Predominant Models Used in Research on the Empirical Structure of Psychopathology

Miriam K. Forbes, Ashley L. Greene, Holly F. Levin-Aspenson, Ashley L. Watts, Michael Hallquist, Benjamin B. Lahey, Kristian E. Markon, Christopher J. Patrick, Jennifer L. Tackett, Irwin D. Waldman, Aidan G. C. Wright, Avshalom Caspi, Masha Ivanova, Roman Kotov, Douglas B. Samuel, Nicholas R. Eaton, Robert F. Krueger

Funded by:

- Macquarie University
- National Institutes of Health

Three recommendations based on a comparison of the reliability and validity of the predominant models used in research on the empirical structure of psychopathology

Miriam K. Forbes[1], Ashley L. Greene[2], Holly F. Levin-Aspenson[3], Ashley L. Watts[4], Michael Hallquist[5], Benjamin B. Lahey[6], Kristian E. Markon[7], Christopher J. Patrick[8], Jennifer L. Tackett[9], Irwin D. Waldman[10], Aidan G. C. Wright[11], Avshalom Caspi[12], Masha Ivanova[13], Roman Kotov[14], Douglas B. Samuel[15], Nicholas R. Eaton[2]*, Robert F. Krueger[16]*

*joint senior authors

[1] Centre for Emotional Health, Department of Psychology, Macquarie University, Sydney, Australia
[2] Department of Psychology, Stony Brook University, Stony Brook, NY, USA.
[3] Department of Psychology, University of Notre Dame, and Department of Psychiatry and Human Behavior, Brown University Warren Alpert Medical School
[4] Department of Psychological Sciences, University of Missouri, Columbia, MO, USA.
[5] Department of Psychology and Institute for Computational and Data Sciences, Penn State University, PA USA and Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill
[6] Department of Public Health Sciences, University of Chicago, Illinois
[7] Department of Psychology, University of Iowa, Iowa City, IA, USA
[8] Department of Psychology, Florida State University, Florida, USA
[9] Department of Psychology, Northwestern University, IL, USA
[10] Department of Psychology, Emory University, Georgia, USA
[11] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
[12] Department of Psychology & Neuroscience, Duke University, Department of Psychiatry & Behavioral Sciences, Duke University, Center for Genomic and Computational Biology, Duke University, Institute of Psychiatry, Psychology, & Neuroscience, King's College London, PROMENTA Center, University of Oslo, Norway
[13] Department of Psychiatry, University of Vermont, Burlington, Vermont, USA
[14] Department of Psychiatry, Stony Brook University, Stony Brook, NY, USA.
[15] Department of Psychological Sciences, Purdue University, Indiana, USA
[16] Department of Psychology, University of Minnesota, MN, USA

Correspondence to: Miriam K. Forbes, PhD
Room 701 Building 4 First Walk
Centre for Emotional Health,
Department of Psychology,
Macquarie University,
Sydney Australia 2109
Email: miri.forbes@mq.edu.au

Acknowledgement: This project was conceived as a HiTOP Higher-Order Workgroup project. We would like to thank all members of the Workgroup who contributed to the development of the ideas and analyses in this project.

Author note: Preliminary results from this research were disseminated at the HiTOP meeting in 2018. Preprint drafts, analytic plans, and supplementary materials for this manuscript are posted on the Open Science Framework (https://osf.io/76hpu/). The research protocol of the National Epidemiologic Survey on Alcohol and Related Conditions, including informed consent, received full ethical review and approval from the United States Census Bureau and Office of Management and Budget.

Funding: M.K. Forbes is supported by a Macquarie University Research Fellowship. R.F. Krueger is partly supported by the US National Institutes of Health, NIH (R01AG053217, U19AG051426).

Word Count (main text): 10,155
Figures: 5
Tables: 4

Abstract

The present study compared the primary models used in research on the structure of psychopathology (i.e., correlated factor, higher-order, and bifactor models) in terms of structural validity (model fit and factor reliability), longitudinal measurement invariance, concurrent and prospective predictive validity in relation to important outcomes, and longitudinal consistency in individuals' factor score profiles. Two simpler operationalizations of a general factor of psychopathology were also examined—a single-factor model and a count of diagnoses. Models were estimated based on structured clinical interview diagnoses in two longitudinal waves of nationally representative data from the United States (n = 43,093 and n = 34,653). Models that included narrower factors (fear, distress, and externalizing) were needed to capture the observed multidimensionality of the data. In the correlated factor and higher-order models these narrower factors were reliable, largely invariant over time, had consistent associations with indicators of adaptive functioning, and had moderate stability within individuals over time. By contrast, the fear and distress specific factors in the bifactor model did not show good reliability or validity throughout the analyses. Notably, the general factor of psychopathology (p-factor) performed similarly well across tests of reliability and validity regardless of whether the higher-order or bifactor model was used; the simplest (single-factor) model was also comparable across most tests, with the exception of model fit. Given the limitations of categorical diagnoses, it will be important to repeat these analyses using dimensional measures. We conclude that when aiming to understand the structure and correlates of psychopathology it is important to: 1) look beyond model fit indices to choose between different models; 2) examine the reliability of latent variables directly; and 3) be cautious when isolating and interpreting the unique effects of specific psychopathology factors, regardless of which model is used.

Keywords: Quantitative psychopathology, latent variable models, bifactor model, hierarchical model, validity

General Scientific Summary: This study used an applied example to compare the reliability and validity of the most widely used latent variable models in research on the empirical structure of psychopathology, from four perspectives. The results suggest that it is important to 1) look beyond model fit indices to choose between different models; 2) examine the reliability of latent variables directly; and 3) be cautious when isolating and interpreting the unique effects of specific psychopathology factors, regardless of which model is used. For research focused on a general factor of psychopathology (*p*-factor), various general factor models performed similarly well.

Three recommendations based on a comparison of the reliability and validity of the predominant models used in research on the empirical structure of psychopathology

Many well-known limitations of traditional psychiatric classification approaches, such as the *Diagnostic and Statistical Manual of Mental Disorders (DSM-5;* American Psychiatric Association, 2013), are a direct consequence of categorizing dimensional phenomena (Kotov et al., 2017). Most notably, there is extensive diagnostic comorbidity (e.g., Hasin & Kilcoyne, 2012; Kessler et al., 2005), pervasive heterogeneity within disorders (e.g., Widiger & Trull, 2007), and arbitrary cutoffs that demarcate the presence or absence of a mental disorder, each of which contribute to relatively poor diagnostic reliability (Regier et al., 2013). Data-driven classification (i.e., the Hierarchical Taxonomy of Psychopathology [HiTOP]; Kotov et al., 2017) overcomes many of these limitations by reformulating categorical conceptualizations of psychopathology as a hierarchy of latent dimensions that account for systematic patterns of covariation among disorders and also demonstrate greater reliability, validity, and clinical utility than categorical diagnoses (e.g., Markon, Chmielewski, & Miller, 2011; Rodriguez-Seijas, Eaton, & Krueger, 2015; Ruggero et al., 2019; Verheul, 2005).

Three models tend to be the predominant focus of investigations on latent structures of psychopathology and maladaptive personality (see Figure 1): the correlated factor model (Model 1), the higher-order model (Model 2), and the bifactor model (Model 3); some research has also examined a single-factor model (Model 4). To date, no study has reported on an in-depth comparison of these models to our knowledge. Indeed, we are aware of only one study that compared Models 1-4 (Carragher et al., 2016), and it did so only on the basis of model fit. As such, the present study comprehensively examined each of these models in an applied example, in turn characterizing their reliability and external validity, with the aim of improving our understanding of each model's relative strengths and weaknesses.

[Figure 1]

Popular Structural Models of Psychopathology

The correlated factor, higher-order, and bifactor models are statistically closely related; for example, when three first-order factors are estimated, the correlated factor and higher-order models are equivalent, and may be nested within the bifactor model (Mansolf & Reise, 2017; Morin et al., 2016; Reise, 2012). However, each model represents a substantively different characterization of the latent structure of common mental disorders.

The correlated factor model (Model 1 in Figure 1) is the original structural model used in child and adult psychopathology research (e.g., Achenbach, 1966; Krueger et al., 1998). Two or more distinct but related latent variables summarize the shared variance among their indicators, accounting for the shared processes among the disorders within each spectrum (Wright, 2020). Correlated factor models tend to identify at least two transdiagnostic spectra of psychopathology—internalizing and externalizing (Achenbach, 1966)—and internalizing is sometimes split into lower-order *fear* and *distress* factors (e.g., Greene & Eaton, 2015; Krueger & Markon, 2006).

There are sizable correlations between the spectra underlying common mental disorders (typically ranging from .4 to .7, depending on sample composition; Eaton et al., 2010), which could be modeled as a *general factor of psychopathology* (also often referred to as the *p-factor*; Caspi et al., 2014; Lahey et al., 2012). This general factor of psychopathology has been a topic of considerable interest in recent psychopathology research (e.g., Caspi & Moffitt, 2018; Lahey et al., 2016; Patalay et al., 2015; Stochl et al., 2015) and can take a number of different forms in a statistical model. For example, to directly model the correlations among a set of latent variables in the correlated factor model, a higher-order general factor that captures factor intercorrelations can be modeled atop fear, distress, and externalizing spectra[1] (e.g., Blanco et al., 2015; Carragher et al., 2016; Chen et al., 2006). In

[1] Note that when there are three first-order factors, the fit of the correlated factor and higher-order models cannot be compared, as the general factor in the higher-order model is a just-identified reparameterization of the three correlations among the first-order factors.

this higher-order factor model (Model 2 in Figure 1), the first-order factors still represent the shared variance among the diagnostic indicators within each spectrum, and the higher-order general factor represents the shared variance among the first-order factors (i.e., the higher-order factor has only indirect relationships with the diagnostic indicators).

Another way to parameterize the general factor of psychopathology is within the context of a bifactor model (Model 3 in Figure 1), which has been widely adopted in the literature (e.g., Caspi et al., 2014; Caspi & Moffitt, 2018; Lahey et al., 2012, 2018). In contrast to the higher-order model, the bifactor model's general factor directly captures the shared variance among all of the observed diagnostic variables. The bifactor model then partitions the remaining variance among subsets of observed variables into uncorrelated specific factors (e.g., fear, distress, externalizing). These specific factors thus represent what differentiates the disorders within each spectrum from those in other spectra, but are expected to contain more measurement error than the general factor (Demars, 2013; Markon, 2019). Interpretation of the specific factors hinges on a clear understanding of their residual properties, with some uncertainty as to what stable clinical constructs (i.e., observable in individuals) can be reliably attributed to these residual factors after the shared variance among all diagnoses is accounted for by the general factor. The bifactor model's partitioning of variance is also a strength that is often used in the psychological measurement literature to test (1) whether a given dataset is essentially unidimensional (i.e., captured by a general factor alone), and (2) whether specific factors add incremental value beyond the general factor in understanding the structure of the data (Demars, 2013; Markon, 2019).

Burgeoning research suggests that a general factor of psychopathology is a powerful predictor of important outcomes in clinical research (e.g., Caspi & Moffitt, 2018; Forbes et al., 2019; Lahey et al., 2016), but general factors captured in the higher-order and bifactor models are relatively removed from constructs that can be observed or assessed directly in a

clinical context, especially at the level of an individual. Therefore, the present study will also consider two less complex models for comparison: a single-factor model (Model 4 in Figure 1) and a count variable of diagnoses, which are more easily computed in clinical settings (e.g., as a weighted or unweighted count of an individual's diagnosis, respectively).

Model Comparison Criteria

The present study conducts a comprehensive comparison of various approaches to modeling a general factor of psychopathology using categorical diagnoses—as well as the narrower fear, distress, and externalizing factors in Models 1-3—according to four criteria: 1) structural validity (model fit and factor reliability), 2) longitudinal measurement invariance, 3) concurrent and prospective predictive validity in relation to important outcomes, and 4) longitudinal consistency in individuals' factor score profiles.

***Structural validity*.** Much of the existing literature has relied on model fit as a justification for the use of a given model, but traditional model fit indices have considerable shortcomings. First, they provide no justification for a model's substantive operationalization of psychopathology constructs. Second, simulation studies show that traditional fit indices tend to favor the highly flexible bifactor model over other less complex models (e.g., correlated factor, higher-order; Greene et al., 2019; Morgan et al., 2015; Murray & Johnson, 2013), because it can better accommodate various types of unmodeled complexity (e.g., negligible cross-loadings, correlated residuals, and random noise or error; Bonifay & Cai, 2017; Reise et al., 2016). Consequently, even with apparent close fit on traditional model fit indices, the bifactor model could generate idiosyncratic parameter estimates that may not be robust across samples or time (Eid et al., 2017; Levin-Aspenson et al., in press).

This issue poses considerable challenges for selecting the "best" model on the basis of fit, and so it is important to consider other features in adjudicating between structural models (Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Hopwood & Donnellan, 2010; McCrae

et al., 1996; Rodriguez et al., 2016b; Schmitt et al., 2018; Sellbom & Tellegen, 2019; Waldman et al., 2017; Watts et al., 2019). Other work has called for greater consideration of model interpretability (e.g., the strength, sign, and precision of factor loadings; Eid et al., 2017; Waldman et al., 2017; Watts et al., 2019) and the reporting of model-based reliability indices (Rodriguez et al., 2016a, 2016b; Watts et al., 2019). Psychopathology research has begun making important inroads into this issue—for example, comparing a subset of models (e.g., bifactor, correlated factors model) on the basis of model-based reliability indices recommended by Rodriguez et al. (2016b; see Conway et al., 2019; Gomez et al., 2018; Watts et al., 2019). These recent developments in the literature have highlighted diverging levels of interpretability and factor reliability among structural models of psychopathology (e.g., occasionally weak or negative specific factor loadings in a bifactor model that are at odds with the interpretation of the corresponding factor in a correlated factor model, and greater reliability for correlated factors compared to specific factors from a bifactor model; Eid et al., 2017; Kim et al., 2019; Watts et al., 2019, in press). Our study will use complementary metrics in addition to traditional fit indices to determine the degree to which the "best fitting" model also corresponds to superior factor strength and reliability.

Longitudinal measurement invariance. Longitudinal measurement invariance tests for structural equivalence of latent variable models over time. To be confident we are assessing the same latent constructs longitudinally, the pattern and values of estimated factor loadings need to be the same across waves. If longitudinal measurement invariance holds, it indicates model stability and allows us to compare individuals' levels of the latent variables over time. The correlated factor model has shown measurement invariance over time in psychopathology data (e.g., Vollebergh et al., 2001), but few studies have formally investigated longitudinal measurement invariance for the other models. With respect to the bifactor model, most studies have simply fit the same model at two or more time points

(configural invariance), but have not constrained parameter loadings across waves (metric/scalar invariance; Greene & Eaton, 2017; Snyder et al., 2017). Two studies have examined longitudinal measurement invariance of the bifactor model in psychopathology data. Gluschkoff and colleagues (2019) found that both the bifactor and correlated factor models met criteria for measurement invariance over time when indicators were allowed to be autocorrelated between waves, whereas Olino and colleagues (2018) found that the bifactor model did not meet criteria for metric invariance over time. Failing to meet metric invariance means that the latent factors did not capture the same constructs over time. We sought to extend this work by quantifying and comparing the stability of all four latent variable models over time as a further test of structural validity and reliability.

Concurrent and prospective prediction in relation to important outcomes. In addition to model fit and factor reliability, it is important that the variance summarized by each model factor is in fact useful for understanding psychopathology. Investigations of associations between latent factors and external correlates are helpful for determining which models have strong criterion validity and/or predictive utility in relation to important outcomes. It is also important to consider how such patterns of validity might differ across models. In previous comparisons of the correlated factor and bifactor models, the bifactor model's general factor typically had the strongest links with correlates of psychopathology and adverse outcomes (e.g., Laceulle et al., 2015, 2019; Lahey et al., 2012; Patalay et al., 2015, although see also Michelini et al., 2019; Watts et al., 2019). These results align with recent hypotheses that the general factor of psychopathology may be an index of impairment that is shared among nearly all common disorders (Caspi et al., 2014; Smith et al., 2020; Widiger & Oltmanns, 2016). Three recent studies that specified both a higher-order model and a bifactor model found concurrent external validity for the general factor to be very similar across the two models (Clark et al., under review; Conway et al., 2019; Moore et al., 2020). In contrast, the external

validity of specific factors in a bifactor model is less clear, with suppression effects sometimes emerging after the general factor has accounted for much of the shared variance among indicators (Conway et al., 2019). For example, in a bifactor model, the internalizing specific factor has been found to be positively associated with adaptive outcomes after taking into account a general factor of psychopathology (Lahey et al., 2015; Caspi et al., 2014; Lacuelle, 2019).

One benefit of the bifactor model's orthogonal factors is the ability to examine the unique contribution of specific factors independent from the general factor in the prediction of external criteria. However, the substantive meaning of specific factors from the bifactor model is likely quite different from their counterparts in the correlated factors and higher-order models—despite investigators' common usage of identical labels. Similarly, if a multiple regression framework is used to characterize the unique associations of first-order factors in a correlated factor model, interpretation of the results is based on the unique variance of each specific factor, rather than the original latent construct (Chen et al., 2006; Lynam et al., 2006). With these issues in mind, we will compare the broad (general factor of psychopathology) and narrow (fear, distress, and externalizing) factors from each model—including the single-factor and count variable—as predictors of adaptive functioning, with the aim of understanding whether and how the modeling frameworks differ in terms of concurrent and predictive validity when compared in the same data.

Stability of individuals' profiles on the factors. While the current discourse framing these models is mainly concerned with identifying optimal structural representations of psychopathology (i.e., a group-level question), these models also have implications for individual-level assessment as they allow for estimation of individual scores on the latent variables (Eid 2020). To our knowledge, no study has examined the extent to which each model could generate profiles of psychopathology dimensions that were stable within

individuals over time (i.e., to test whether the between-subjects latent variables also manifested as reliable summaries of individuals' experiences).

The Present Study

In sum, we aimed to compare the predominant modeling frameworks for the structure of psychopathology (i.e., correlated factor, higher-order, and bifactor models) in terms of four criteria: 1) structural validity (model fit and factor reliability), 2) longitudinal measurement invariance, 3) concurrent and prospective predictive validity in relation to important outcomes, and 4) longitudinal consistency in individuals' factor score profiles. As noted earlier, some of these criteria have been considered in the literature, but others have not; most studies have examined or compared only one or two models based on one or two of these criteria (e.g., model fit and concurrent predictive validity). This study thus sought to replicate and extend these comparisons. Further, no study has: 1) considered whether simpler (single-factor and count variable) models can account for the associations of the general factors estimated in more complex models, or 2) compared the stability of any of the models over time, or the longitudinal consistency in individuals' factor score profiles. In contrast with previous examinations, we report on a comprehensive comparison of the predominant models used in the research on the empirical structure of psychopathology and consider this comprehensive set of criteria in concert. Doing so was expected to provide new perspectives on the relative strengths and weaknesses of the various models for different research questions and applications. Further, by applying these criteria to models estimated in the same dataset, we avoided the possibility of differences in model performance across the four criteria being attributable to sample characteristics or the type of correlation matrix used, since all models were compared on even footing. We also preregistered the analytic plan to guard against analytic flexibility that may have biased our results and conclusions.

Thus, the present study built on prior studies that have characterized the latent structure of mental disorders using data from the very large and longitudinal National Epidemiologic Survey on Alcohol and Related Conditions (NESARC; e.g., Albott et al., 2018; Eaton et al., 2013; Greene & Eaton, 2017; Hoertel et al., 2015; Kim & Eaton, 2015; Lahey et al., 2012). As noted earlier and revisited in interpreting the results below, it is important to bear in mind that the observed variables in the present study are not based on empirically derived dimensional phenotypes of psychopathology, but rather on categorical DSM-IV diagnoses, which lose information on individuals' specific symptom profiles and severity and tend to be less reliable than dimensional indicators.

Method

Participants and Procedure

This study used longitudinal data from two waves of NESARC, a representative sample of the adult, civilian, noninstitutionalized United States population: Wave 1 (n = 43,093; collected 2001–2002; response rate: 81% of those eligible) and a wave 2 follow-up (n = 34,653; collected 2004–2005; 86.7% of eligible original sample; 70.2% cumulative response rate). Wave 1 was 57% female; ages ranged from 18 to 98 years. Hispanic/Latinx, Black, and young adults were over-sampled. White participants composed 56.9% of the sample, Hispanic or Latinx 19.3%, African-American 19.1%, Asian or Pacific Islander 3.1%, and American Indian or Alaska Native 1.6%. Wave 2, was 48% female; ages ranged from 20 to over 90 years. White participants composed 70.9% of the sample, African-American 11.1%, Hispanic or Latinx 11.6%, Asian or Pacific Islander 4.3%, and American Indian and Alaska Native 2.2%. Design variables ensured age, racial/ethnic, and gender representativeness of the United States based on the 2000 Census.

Measures

Psychopathology. Lifetime (Wave 1) and "since last interview" (Wave 2) *DSM–IV* diagnoses, assessed via the Alcohol Use Disorder and Associated Disabilities Interview Schedule—*DSM–IV* Version (AUDADIS–IV; Grant et al., 1995), were used in the current study. The AUDADIS-IV is a structured interview designed for highly trained lay interviewers. We examined major depressive disorder, dysthymic disorder, generalized anxiety disorder, panic disorder and agoraphobia, social phobia, specific phobia, alcohol dependence, nicotine dependence, marijuana dependence, and other drug dependence diagnoses. We also examined adult antisocial behavior, which was defined as the presence of at least three antisocial behavior items endorsed since age 15 (Wave 1) and since last interview (Wave 2). The other drug dependence variable combined relatively uncommon forms of drug dependence (i.e., stimulants, opioids, sedatives, tranquilizers, cocaine, solvents, hallucinogens, heroin, and any other drug not assessed) into one variable with sufficient variance; the internal consistency of this variable was adequate (alpha = .77). The reliability of the AUDADIS–IV diagnoses employed in the current work has been reported elsewhere and is generally good to excellent (e.g., kappas ranged from .42 [fair] to .84 [excellent agreement]; see Hasin et al., 2005). Test–retest estimates for AUDADIS–IV disorders are similar to other structured interviews (Wittchen, 1994), although longitudinal stability of the diagnoses observed here tended to be low, reflecting a mix of interrater and test-retest reliability (e.g., kappas ranged from .11 [slight] to .48 [moderate agreement]; Kuder-Richardson 20 ranged from .20 to .65; proportion of cases at Wave 1 with the same diagnosis at Wave 2 ranged from 7.9% to 51.0%—all indices indicated that marijuana dependence had the lowest consistency over time and tobacco dependence the highest).

Adaptive functioning. Eight self-reported indices of adaptive functioning were assessed at both Wave 1 and Wave 2 and examined as potential negative outcomes of psychopathology. To facilitate comparison across predictors, models, and outcomes, all

outcomes (as follows) were coded to be dichotomous: (1) Being fired/laid off from job in past year (yes [1] or no [0]); (2) Unemployed and looking for a job for more than a month in past year (yes [1] or no [0]); (3) Relationship breakdown (separated, divorced, broke off steady relationship) past year (yes [1] or no [0]); (4) Experienced a major financial crisis, declared bankruptcy, or been unable to pay bills past year (yes [1] or no [0]); (5) Fair or poor self-perceived current physical health (fair or poor [1] vs. good to excellent [0]); (6) Accomplished less than would like or did work/other activities less carefully than usual most or all of the time in the past four weeks due to emotional problems (most or all of the time for either item [1] vs. none to some of the time for both items [0]); (7) Chronic illness diagnosis confirmed by health professional in the past year, based on hardening of arteries, high blood pressure, chest pain/angina, rapid heartbeat, heart attack, liver disease/cirrhosis, heart disease, ulcer, gastritis, arthritis (present [1] or absent [0]); and (8) Body mass index indicating obesity (less than 30 [0] vs. 30 or more [1]).

Data Analysis

Preregistration. Consistent with the open science movement aiming to improve research methods, we posted our analytic plan on the Open Science Framework prior to conducting any analyses (see https://osf.io/kzsa4/) with the aim of mitigating bias in model selection, outcome reporting, and hypothesizing after the results are known (HARKing; Kerr, 1998; Schmitt, 2011; Tackett et al., 2017). We subsequently updated the plan (see https://osf.io/rj53d/) due to preliminary models (i.e., based on 12-month diagnoses) failing to converge, as well as feedback from the HiTOP Higher-Order Workgroup at the annual HiTOP meetings in 2017 and 2018. Given the volume of output generated in the preregistered analyses, some results are not included in-text, but are presented in the online supplement. All deviations from the preregistered analytic plan are explicitly noted.

Model estimation. Latent variable models were estimated treating all observed variables as categorical and using the complex survey weighting variables at each wave to maintain a demographically representative sample. Analyses were initially conducted based on both weighted least square mean and variance adjusted (WLSMV) and maximum likelihood with robust standard errors (MLR) estimators to compare model fit and factor reliability. Latent variables were standardised to have a mean of 0 and a variance of 1, with factor loadings freely estimated. The count variable of the general factor was a sum of the number of diagnoses for which an individual met criteria.

Structural validity. Traditional model fit indices were used to assess model fit. To quantify absolute fit, we used the root mean square error of approximation (RMSEA; values < .06 indicating close fit); incremental fit indices included the comparative fit index and Tucker-Lewis index (CFI and TLI; values > .95 indicating close fit). The information criteria were used to directly compare models—the Akaike information criterion (AIC), Bayesian information criterion (BIC), and the sample-size adjusted BIC (SSABIC)—for which lower values indicate better fit (e.g., differences of 10 points strongly favour the model with a lower value; Rafferty, 1995). We also compared the magnitudes of the standard errors for the factor loadings in each model as an indication of the precision of these parameters (Waldman et al., 2017). Our preregistration included a plan to test factor determinacy (i.e., ranging from 0 to 1, with larger values indicating better measurement of the factors by the observed variables). However, as discussed below, this was not possible, which had implications for the concurrent and prospective prediction analyses.

Our preregistration also noted that we would test construct replicability (H; Hancock & Mueller, 2001) using the approach of Rodriguez et al. (2016), which represents how well-defined a latent variable is by its indicators, and corresponds to the likelihood that the estimated factors are replicable across studies (ideally H > .8; Hancock & Mueller, 2001). We

expanded this plan to include all of the model-based indices recommended by Rodriguez et al. (2016b) for quantifying the reliability, strength, and dimensionality of the latent variable models. These indices provide particularly rich information about the bifactor model but have implications for understanding the strength and expected replicability of all of the latent variables (Brunner et al., 2012).[2] While there are limited universal cut-off criteria, we present commonly used heuristics for interpreting these indices in isolation and in concert (Reise, Moore, & Haviland, 2010; Reise, Scheines, et al., 2013; Rodriguez et al., 2016a, 2016b).

The computation of reliability for individual latent variables differed based on the model in which they were situated. For the bifactor model, omega hierarchical (ω_h; McDonald, 1999; Reise, Moore, & Haviland, 2013; Zinbarg et al., 2005) estimates the proportion of systematic variance in a count variable of the indicators that is accounted for by the general factor (ideally $\omega_h > .8$); omega hierarchical subscale (ω_{hs}) indexes the reliability of each specific factor after partialling out variance attributable to the general factor ($\omega_{hs} > .75$ indicates sufficient reliability to be used in practice; and $\omega_h/\omega_{hs} < .5$ indicates insufficient precision such that the factor should not be used in practice; Bonifay, Reise, & Haviland, 2013; Gignac & Watkins, 2013). When ω_h is large (>.8) and ω_{hs} values are comparatively small there is evidence that the general factor is more reliable than specific factors. For the single-factor model, omega total (ω_t; ideally > .75) estimates the proportion of variance in the observed total score attributable to all modeled sources of common variance (i.e., the percentage of total variance accounted for by a single latent construct; McDonald, 1999; Revelle & Zinbarg, 2009; Zinbarg et al., 2005). For the correlated factor model, omega subscale (ω_s; ideally > .75) focuses on one subset of indicators at a time to estimate the

[2] These indices cannot be calculated for the higher-order model directly; a Schmid Leiman transformation (orthogonalization) is required to calculate them, making interpretation challenging. For example, in this transformation the general factor is residualized out of the specific factors, which removes much of their variance in a manner similar to the bifactor model. The indices for the transformed second-order model are given in the supplement (Table S1), and are nearly identical to the bifactor model results given below across all factors and indices (with no differences in interpretation).

proportion of variance in the observed subscale score that is attributable to the corresponding first-order factor.

To ascertain the relative strength of factors and characterize the degree of essential unidimensionality, we calculated the explained common variance (ECV; Reise, Scheines et al., 2013; Reise et al., 2010; Sijtsma, 2009; Ten Berge & Socan, 2004). The percent of common variance across all indicators that is explained by the general factor indexes the importance of the general factor relative to the specific factors (i.e., ideally ECV > .7, and > .85 if there is evidence of unidimensionality; Stuckey & Edelen, 2014). To quantify the uniqueness of a specific factor, we used ECV_S to estimate the percent of explained variance for only those indicators loading on that specific factor (ideally ECV_S >.7).

The percentage of uncontaminated correlations (PUC; Reise, Scheines et al., 2013; Bonifay et al., 2015) represents the proportion of correlations that only reflect variance from the general factor (i.e., are "uncontaminated by multidimensionality"; Reise et al., 2013, p. 5), indexing potential bias resulting from fitting a unidimensional model to multidimensional data (PUCs > .7 provide evidence for unidimensionality). Together, ECV and PUC indicate whether the common variance in the model can be interpreted as essentially unidimensional, thereby reducing the specific factors to disturbance factors that are not conceptually meaningful (e.g., when both ECV and PUC are > .70). Finally, average parameter bias (APB) quantifies the difference between an item's loading in the unidimensional solution and its general factor loading in the bifactor model (10-15% is acceptable; Muthén, Kaplan, and Hollis, 1987).

The structural validity and reliability analyses for the WLSMV and MLR estimators were highly similar (see Table 1 and Table S2 for standardized factor loadings using WLSMV and MLR, respectively, Table 2 for model fit, and Table 3 and Table S3 for reliability

indices). The more computationally efficient WLSMV estimator was thus used in subsequent latent variable analyses.

Longitudinal measurement invariance. Longitudinal measurement invariance was tested according to the approach described by Widaman and colleagues (2010), using default Mplus recommendations for WLSMV and delta parameterisation (Muthén & Muthén, 2012). Unconstrained models were compared to models with factor loadings and thresholds constrained to equality between waves based on changes in CFI, using the more stringent criterion of a decrease ≤ .002 points to indicate strict measurement invariance (Meade et al., 2008), and the less stringent criterion of a decrease ≤ .01 points to indicate general support for measurement invariance (Cheung & Rensvold, 2002).[3] Additional exploratory (not preregistered) analyses were conducted to quantify the similarity of the estimated factor scores for latent variables between models based on Spearman correlations (e.g., correlating the estimated factor scores for the Wave 1 fear latent variables from the correlated factor, higher-order, and bifactor models).

Concurrent and prospective predictive validity. We examined concurrent and prospective validity for each latent variable, and for the count of diagnoses, in the statistical prediction of important outcomes. Concurrent validity was evaluated by examining the variance accounted for in each outcome by each factor at Wave 1, adjusting for age and sex (i.e., the difference between the full model R^2 and the R^2 for covariates only). Prospective validity was evaluated by examining the variance accounted for in each outcome at Wave 2 by each factor at Wave 1, adjusting for age, sex, and the presence of each outcome at wave 1

[3] These criteria were derived based on simulations of continuous data examined in a multiple group measurement invariance framework (Cheung & Rensvold, 2002; Meade et al., 2008). Sass and colleagues (2014) examined the performance of these criteria using ordinal data and WLSMV estimation in a multiple group format and found that with sample sizes of at least n = 500 per group, the Meade et al. criteria in particular performed well at detecting non-invariance between groups. Power to detect non-invariance also increased as a function of sample size (n = 150, 300, and 500 per group), so it seems likely that the very large sample sizes used here will afford sufficient power to detect non-invariance in the models. However, it is important to keep in mind that these thresholds have not been validated specifically for longitudinal measurement invariance using dichotomous indicators in very large samples. The chi-square difference tests based on the DIFFTEST function in Mplus were likely overpowered in these sample sizes, and were significant at $p < .05$ for all model comparisons.

(i.e., the difference between the full model R^2, and the R^2 for a model including only these covariates).

As mentioned earlier, our preregistered analytic plan indicated that we would use factor determinacy to choose between an estimated factor score framework (i.e., using factor scores estimated in the longitudinal measurement invariance models as the observed predictor variables in a logistic regression framework) versus a structural equation modeling framework (i.e., using the latent variable models with all parameters fixed based on the longitudinal measurement invariance testing in a probit regression framework). However, after commencing analyses we learned that factor determinacy cannot be accurately estimated when using dichotomous indicators (Beauducel & Hilger, 2016; Ferrando & Lorenzo-Seva, 2017). We thus decided to conduct the analyses in both frameworks, and treated variation in the interpretation of the two frameworks as an indicator of low determinacy. Our preregistration included planned comparisons of the standardized regression coefficients and confidence intervals, but given the overlap in these results we report the R^2 results here and report the betas, odds ratios, and confidence intervals in Tables S4-S7. R^2 values in the SEM and factor score frameworks quantify the proportion of explained variance in a latent response variable underlying the binary outcome for each model, but the latent response variables have different distributions, so the precise R^2 values are not directly comparable across SEM and factor score frameworks. As such, we focus on substantive similarities and differences when comparing the results of the SEM versus factor score frameworks.

We also do not report the planned multiple regression analyses (i.e., simultaneously entering all latent variables as predictors) for the bifactor and higher-order hierarchical models predicting each outcome. For the bifactor model, this is due to conceptual redundancy, as the latent variables are orthogonal (i.e., represent the same variance whether partialled or

unpartialled).[4] For the higher-order model, the multiple regression models were not identified. The multiple regression analyses for the correlated factor model, and entering all diagnoses as simultaneous predictors (per the preregistered analyses), are presented in Tables S8-S12. Correlation coefficients for the latent variables and maximum *a posteriori* estimated factor scores with each external criterion are also presented in Tables S13-S16.

Consistency in individuals' profiles. Factor scores were used to create psychopathology profiles at Waves 1 and 2 for each individual in each model (i.e., estimating their levels of general psychopathology and fear, distress, and externalizing, as appropriate). For each individual, these values were compared within each model between waves based on multiple metrics of consistency in the profiles over time (see McCrae, 2008; Woods et al., 2020), including profile elevation and scatter (i.e., mean and variance of the individual's factor scores at each wave), profile shape (i.e., Pearson correlation between the individual's factor score profiles at each wave), omnibus profile similarity between waves (as indexed by a double-entry intraclass correlation coefficient for the individual's factor score profiles at each wave), and rank-order stability (as indexed by a Spearman rank-order correlation comparing the individual's rank on each variable between waves). The full results for these metrics are presented in Table S17, and the Spearman rank-order correlations for each variable (i.e., all latent variables and the count of diagnoses) are presented below.

Results

Structural Validity

Standardized factor loadings for the four latent variable models at each wave, using WLSMV estimation, are shown in Table 1 (see Table S2 for MLR estimation). All factor loadings were positive, and most were substantial, although generalized anxiety disorder was a particularly weak indicator of the distress specific factor in the bifactor model at both waves

[4] We do report, in Table S12, the total change in R^2 accounted for by the bifactor versus correlated factor models with all latent variables entered as simultaneous predictors of each outcome, and with all diagnoses entered as simultaneous predictors.

(from λ = .16 to λ = .25 between the two estimators). According to the traditional fit indices in Table 2, all models fit the data well except for the single-factor model. Similarly, the correlated factor, higher-order, and bifactor models each accounted for very similar amounts of variance in the indicators across waves and estimators (R^2 = 60% to R^2 = 65% on average), whereas the single-factor model tended to account for less variance (R^2 = 47% to R^2 = 51% on average). The bifactor model exhibited the best fit to the data, based on the information criteria (AIC, BIC, SSABIC). However, the bifactor model factor loadings also had appreciably larger standard errors than the other models, particularly for the specific factors (see Figure 2 for WLSMV estimation and Figure S1 for MLR).

[Table 1]

[Table 2]

[Figure 2]

The greater imprecision of parameter estimates for the bifactor model was further corroborated by reliability indices (Table 3 and Table S3). Specifically, ECV (0.57 to 0.61) and APB (0.17 to 0.25) values did not meet the criteria for an acceptable bifactor model (ECV > .70 and APB < 0.15; Stucky and Edelen, 2014), and the results for ω_h (.72 to .75) implied that these data were somewhat multidimensional (i.e., ω_h < .8). However, the bifactor model's specific factors did not reliably capture this multidimensionality (i.e., ECV_S < .0.70 and ω_{hs} < .75). The externalizing specific factor had just-acceptable reliability (i.e., ω_{hs} > 0.50), but the distress and fear specific factors did not capture substantial unique or reliable variance and were poorly defined, meaning they should not be used in practice (e.g., < 30% of the variance in the composite scores for these factors was due to the target construct, making it very difficult to interpret these scores in applied settings). Overall, the general factor in the bifactor model was more reliable and better defined than the specific factors ($\omega_h > \omega_{hs}$).

[Table 3]

In contrast with the poor reliability of the specific factors in the bifactor model, the factors in the correlated factor model were reliable, interpretable, and well-defined, accounting for > 83% of the systematic variance in the indicators. The single-factor model was also reliable, interpretable, and well defined by the indicators, although, as noted earlier, forcing a unidimensional structure creates some bias and lost information due to the multidimensionality of the data; the same is true of using a count variable of diagnoses, which forces a unit-weighted unidimensional structure.

Longitudinal Measurement Invariance

All models required a single correlated residual across waves for the tobacco dependence variable (r = .74)—corresponding to marked stability in the diagnosis over time—to fix a non-positive definite latent variable covariance matrix. The correlated factor model met Cheung and Rensvold's (2002) criterion for longitudinal measurement invariance (ΔCFI = -.06). However, a very large modification index (193.4) indicated that the threshold for adult antisocial behaviour (AASB) should be freed between waves, reflecting a change in the meaning of antisocial behaviour as an indicator of externalizing between incidence since age 15 at Wave 1 (23.9%) and incidence since the last interview at Wave 2 (3.7%). After freeing this threshold, the correlated factor model also met Meade et al.'s (2008) stricter criterion for measurement invariance (ΔCFI = -.002). The higher-order model met Cheung and Rensvold's criterion for longitudinal measurement invariance (ΔCFI = -.01), but did not meet Meade et al.'s stricter criterion even after freeing the AASB threshold (ΔCFI = -.003). The bifactor and single-factor models both met Meade et al.'s strict criterion for longitudinal measurement invariance without freeing the AASB threshold (ΔCFI = -.002 and .001, respectively). However, to put the models on even footing in subsequent analyses based on estimated factor scores, all models included a correlated residual for tobacco dependence and a freed threshold for AASB between waves.

Spearman correlations among the estimated factor scores indicated particularly high similarity of the general factor of psychopathology between models: Estimated factor scores from the single-factor, higher-order, and bifactor general factors all correlated $\rho \geq .98$ at wave 1 and $\rho \geq .99$ at wave 2. Estimated factor scores for fear, distress, and externalizing were perfectly correlated ($\rho = 1.00$) at both waves when comparing the correlated factor and higher-order models. However, the estimated factor scores from the specific factors in the bifactor model tended to diverge from the fear, distress, and externalizing factors in the correlated factor and higher-order models, respectively: They had weak negative correlations for fear ($\rho = -.23$ and $\rho = -24$ at wave 1, and $\rho = -.19$ and $\rho = -.18$ at wave 2), very weak correlations for distress ($\rho = .11$ and $\rho = .10$ at wave 1, and $\rho = -.07$ and $\rho = -.06$ at wave 2), strong correlations for externalizing at wave 1 ($\rho = .64$ for both models), and weak correlations for externalizing at wave 2 ($\rho = .34$ for both models).

Concurrent and Prospective Predictive Validity

The variance accounted for by each latent variable is presented in Figure 3. In the SEM framework, the fear variable was similar for correlated and higher-order models, as expected, and the fear specific factor from the bifactor model inconsistently predicted higher and lower levels of variance, compared to the other two models. In the factor score framework, the fear variables from the correlated factor and higher-order models were again similar, although they predicted less variance than in the SEM framework; the fear specific factor from the bifactor model consistently predicted trivial variance ($R^2 < 2\%$) for all outcomes. The marked differences between the SEM and factor score frameworks for the bifactor model likely indicate low factor determinacy, and reflect the indications of low reliability in earlier analyses. Similar findings were evident for the distress and externalizing specific factors, except that the bifactor distress factor often had very high R^2 values in the

SEM framework (up to 80%; see Figure S2), compared to consistently low values in the factor score framework (< 5%).

[Figure 3]

The multiple regression analyses for the correlated factor model (i.e., with fear, distress, and externalizing entered as simultaneous predictors) are given in Tables S8-S11. Figure S3 shows the regression coefficients generated within a SEM framework for these multiple regression analyses compared to the bifactor model. Notably, results for the fear, distress, and externalizing factors diverge, despite the shared aims of these analyses in highlighting the unique contribution of each factor as a predictor of external criteria. These differences between models underscore challenges in the interpretation of the specific factors' unique variance in both modeling frameworks. Overall, the latent variables in each model accounted for similar amounts of variance in each outcome (see Table S12). In an SEM framework, the four latent variables in the bifactor model together predicted slightly more variance, on average, compared to the three latent variables in the correlated factor model (14.4% vs. 12.9% at Wave 1; 7.3% vs. 5.4% at Wave 2).

The different operationalizations of the general factor were more consistent between models, in line with their good reliability. The higher-order and bifactor general factors accounted for the most (and very similar amounts of) variance in the outcomes within the SEM framework (see Figure 4), followed by the single-factor model, and the count of diagnoses. In the factor score framework, all models accounted for similar amounts of variance in outcomes (see Figure 4), although the single-factor model tended to account for marginally more variance, followed by the bifactor and higher-order general factors, and then the count of diagnoses. The results for all of the factors in the higher-order model are also presented together in Figure S4, showing that the first-order fear, distress, and externalizing

factors did not differentially predict the outcomes examined here, and the general factor parsimoniously captured the associations.

[Figure 4]

Consistency in Individuals' Profiles

The consistency of individuals' factor score profiles varied enormously over time (e.g., from perfect negative to perfect positive correlations between waves; see Table S17 and Figure S5). The bifactor model showed somewhat poorer consistency across the various characteristics of individuals' profiles (e.g., very large values for scatter at Wave 2, and lower omnibus index values; see Table S17), but all models exhibited only moderate within-subjects consistency between waves on average, in line with the generally low longitudinal stability of the diagnoses over time (e.g., excluding tobacco dependence, all diagnoses had 'slight' or 'fair' consistency between waves and generally less than a quarter of cases for each diagnosis at Wave 1 met criteria for the same diagnosis at Wave 2). Rank-order stability of the factor scores was moderate, with nearly all variables correlating $\rho = .32$ to $\rho = .39$ (see Table 4). The exceptions were the bifactor fear and distress specific factors: their lower reliability was evident again in rank-order stability nearly half that of the first-order fear and distress factors in the correlated factor and higher-order models ($\rho s = .19$ versus $\rho s = .35$).

Given the unusual profiles for the bifactor model at Wave 2 (Figure 5), we conducted exploratory analyses comparing factor scores estimated based on the unconstrained and constrained models tested in the longitudinal measurement invariance framework described earlier. We compared the bifactor model to the higher-order model, as both specify fear, distress, externalizing, and general psychopathology factors (see Figure 5). On average, factor scores differed very little between constrained and unconstrained models at Wave 1 for both the bifactor and higher-order models (all $|mean_{diff}| < .03$, $SD < .13$). By contrast, at Wave 2, factor scores differed substantially for constrained and unconstrained models: For the higher-

order model, factor scores were lower in the constrained versus unconstrained models ($mean_{diff}$[SD] = -1.0[.07], -0.8[.17], -0.4[.09], and -0.5[.04] for distress, fear, externalizing, and general psychopathology, respectively), likely reflecting decreases in prevalence of the diagnoses between waves (i.e., shifting from lifetime incidence at Wave 1 to incidence since the previous wave at Wave 2). For the bifactor model, factor scores were also lower for fear and externalizing in the constrained model compared to the unconstrained model ($mean_{diff}$[SD] = -0.8[.09]) and -0.9[.06], respectively), substantially lower for distress ($mean_{diff}$[SD] = -2.3[.05]), but higher for the general factor ($mean_{diff}$[SD] = 0.9[.07]).

[Figure 5]

Discussion

In a framework that integrates and extends existing research on the NESARC data, we compared four latent variable models used in research on the structure of psychopathology—a correlated factor model, a higher-order model, a bifactor model, and a single-factor model—alongside a count variable of diagnoses. Each of these approaches has methodological strengths and weaknesses in terms of their applications to psychopathology (e.g., Bonifay et al., 2017; Bornovalova et al., 2020; Markon, 2019; Watts et al., 2019). In an applied example, we compared the reliability and validity of these alternative structures using a variety of approaches, including model fit, structural properties and reliability, concurrent and prospective criterion validity, and prediction of illness course. We made these comparisons to evaluate whether one model would perform best across the various tests of reliability and validity, and to understand if different research questions and contexts might be better suited to specific models (e.g., from a clinical perspective, the least complex and most interpretable model should be preferred). To our knowledge, this study is the first to comprehensively compare these models using this thorough set of group- and individual-level criteria.

Overall, we found that the NESARC diagnostic data were not unidimensional: narrower factors (fear, distress, and externalizing) were needed to capture the multidimensionality in the data while the unidimensional single-factor model and count variable lost important information. The fear, distress, and externalizing factors in the correlated factor and higher-order models were reliable, largely invariant over time, and exhibited consistent associations with indicators of adaptive functioning as well as moderate stability within individuals over time. By contrast, the fear and distress specific factors in the bifactor model did not show good reliability or validity across analyses. These differences in performance did not emerge when examining the general factors of psychopathology in the higher-order and bifactor models, which were similarly reliable, stable, and predictive of adaptive functioning. A noteworthy finding was that the simplest (single-factor) latent variable model performed comparably well across most tests, with the exception of model fit (i.e., in line with the multidimensionality of the data). Indeed, the general factor was essentially isomorphic across these models, with near-perfect factor score correlations between models within waves ($\rho \geq .98$). The results have implications for research, and assessment in clinical practice, extending the literature in several ways that we describe below.

Consistent with the existing literature, model fit tended to indicate support for the bifactor model, in both an absolute and relative sense (Greene et al., 2019; Morgan et al., 2015; Murray & Johnson, 2013). Except for the single-factor model, all models technically fit the data well, with fit being strongest for the bifactor model. These findings broadly accord with those of simulation studies, which have shown that traditional model fit indices tend to indicate support for a bifactor structure even when the "true" model follows a higher-order or correlated factor structure, suggesting that model fit may be an unreliable indicator of the underlying structure of psychopathology (e.g., Greene et al., 2019; Maydeu-Olivares &

Coffman, 2006; Morgan et al., 2015; Murray & Johnson, 2013). As discussed earlier, this is particularly important in the context of the commonplace (and logical) practice in the literature of comparing these competing models' fit to select which one to carry forward in subsequent analyses (e.g., Carragher et al., 2016; Olino et al., 2018; Snyder et al., 2017). We echo the calls of others that researchers should consider additional properties to adjudicate structural models of psychopathology (e.g., Hopwood & Donnellan, 2010; Murray et al 2018; Schmitt et al 2018; Sellbom & Tellegen, 2019; Waldman et al., 2017; Watts et al., 2019).

One pitfall of relying on model fit is that important structural limitations might be masked. This was exemplified by the bifactor model's excellent fit, despite the particularly low reliability for fear and distress specific factors. Another striking inconsistency between fit and function was found for the interpretation of the bifactor model's estimated factor scores when the model parameters were unconstrained versus constrained in the longitudinal measurement invariance framework: The bifactor model met the strictest criteria for longitudinal measurement invariance of all of the models, but had some of the most substantial shifts in factor loadings between waves and in the estimated factor scores for the model holding measurement invariance between waves. For example, dysthymia was the strongest indicator of the distress specific factor at Wave 1, but a weaker indicator at Wave 2. Similarly, the distress variables were the strongest indicators of the general factor at Wave 1, whereas the fear variables were more prominent indicators of the general factor at Wave 2. These between-wave differences for the bifactor model were obscured by the strongest evidence supporting longitudinal measurement invariance for any of the models. We believe that the larger standard errors associated with the bifactor model's parameter estimates likely accounted for this finding; large standard errors create large confidence intervals, allowing for sometimes sizable differences between constrained and unconstrained models to be deemed 'not significantly different'—even in the context of the very large sample used here. This

elasticity in the bifactor parameter estimates is consistent with other work mentioned earlier on the tendency of the bifactor mode to (over)fit any data, including noise, leading to unstable parameter estimates (Bonifay, 2015; Bonifay & Cai, 2017; Reise et al., 2016).

The bifactor model generated a reliable and valid general factor, which is vital for the recent surge of research on the general factor of psychopathology. However, compared with the correlated factor and higher-order variables, there were marked variations for the bifactor model within and between both the estimated factor score and SEM regression frameworks in terms of the fear, distress, and externalizing specific factors' external validity, consistent with their lower reliability and determinacy (see also Watts et al., in press). Interpreting the substantive meaning of the bifactor specific factors represents an additional challenge (i.e., capturing the shared variance within a spectrum that is *not* shared with indicators of other spectra), which warrants caution when using the bifactor model as a tool to partition general and specific sources of psychopathology variance (e.g., Brandes et al., 2019). This is true for all models that include the narrower fear, distress, and externalizing factors when investigators are interested in isolating their unique (i.e., partialled) associations with external correlates. For example, after partialling out the general factor in the higher-order model—or the shared variance among the correlated factors—the remaining residual variance captures the shared variance within a spectrum that is not shared with indicators of other spectra, just like the bifactor model. Similar issues regarding reliability and interpretability therefore apply.

Given the conceptual similarity between examining partialled associations of the correlated factor with external criteria and counterpart associations for the specific factors from the bifactor model, it was surprising that these approaches yielded different patterns of association. For example, the fear factor of the correlated factor model predicted lower levels or no significant differences for nearly all outcomes, whereas the fear factor of the bifactor model predicted higher levels of nearly all of the outcomes. These differences are difficult to

parse, as the true population model is unknown; the regression coefficients had similar reliability based on the confidence intervals, and we cannot infer that one model was more accurate or valid in identifying "true" specificity between domains and correlates of psychopathology. Simulation studies that compare the two modeling frameworks in identifying known specific associations between latent dimensions and external correlates would help to clarify this point. In the meantime, we propose that if the unique associations of specific or first-order factors with external criteria are of substantive interest in a study, researchers should include the unpartialled correlated factors for comparison to bifactor and/or multiple regression results (i.e., to quantify and interpret any differences between coefficients based on the partialled and unpartialled constructs; Lynam et al., 2006). We found that the unpartialled correlated factors estimated here were well-defined, composed of reliable and interpretable variance, and captured the multidimensionality observed in the data.

By contrast, if the general factor of psychopathology is the main focus of a study, our results suggest that it may not be particularly important which model is used to estimate it. Specifically, the various latent variable operationalizations of the general factor (i.e., higher-order and bifactor, and the simpler single-factor model) were largely similar in terms of reliability, interpretability, and predictive validity. There was marked similarity in the general factors derived from the higher-order and bifactor models (see also Clark et al., under review; Conway et al., 2019; Moore et al., 2020). While the single-factor model did not exhibit good fit to the data—indicating that the data were multidimensional—it was reliable, equivalent to other models in terms of accounting for variance in external criteria, and yielded estimated factor scores that were nearly perfectly correlated with the higher-order and bifactor general factors. In the current data, the general factors also parsimoniously captured the non-specificity of the correlated factor model's fear, distress, and externalizing factors in terms of concurrent and prospective prediction of adaptive functioning. This finding would be

interesting to continue to explore in other data with indicators of psychopathology that would be expected to have differential associations with different domains of psychopathology (cf. Brislin & Patrick, 2019; Venables et al., 2018). We view the observed similarity in the general factor across models as encouraging for researchers who are interested in the nature, causes, and consequences of a general factor of psychopathology: Latent variable general factors may be robust to the modeling framework chosen by the researcher. Future investigations should test this inference further using other data and indicators of psychopathology (e.g., Clark et al., under review; Moore et al., in press), and further explore whether the differences in the profiles of indicators' factor loadings (e.g., for the bifactor general factor versus the single-factor model) are substantive.

An important exception to the apparent isomorphism of latent variable general factors was that the count of diagnoses was a substantially weaker predictor of concurrent and subsequent functioning in the SEM framework. By contrast, it had similar utility in predicting external criteria when compared to the estimated factor scores, and exhibited similar consistency within individuals over time. We suggest that researchers interested in concurrent or prospective validity of latent dimensions of psychopathology pay close attention to factor determinacy if using factor scores (i.e., for continuous indicators, quantify factor determinacy directly; for categorical indicators, look at item response information curves and/or compare SEM with factor score regressions for convergence).

Limitations and Future Directions

Like many examinations of the structure of psychopathology using large-scale, epidemiological data, the present study was based on a single dataset containing binary indicators of psychiatric disorders, due to the skip-out structure of the diagnostic interview used in NESARC—an issue common to most epidemiological samples. This precludes the construction of dimensional symptom counts or homogeneous symptom clusters, which are

demonstrably more reliable than categorical diagnoses, and may have led to more reliable latent variables as well as more consistency in individuals' factor score profiles over time. The particularly low reliability of distress and fear in the bifactor model may reflect the lower reliability of categorical diagnoses, but this is unclear given the mixed evidence for lower reliability of specific factors in bifactor models across samples and indicator types. Two recent analyses of psychopathology symptom dimensions in large samples of children and adolescents did not find the bifactor specific factors to be unreliable (e.g., as judged by construct replicability [H]; Moore et al., 2020; Sunderland et al., 2020), whereas other studies of dimensional indicators have found indications of unreliability in specific factors (e.g., Olino et al., 2018; Snyder et al., 2017; Watts et al., 2019). We encourage ongoing efforts to quantify and compare the reliability of estimated latent variables using a variety of indices to continue to build our understanding of how factor reliability varies as a function of study characteristics.

Available measures of adaptive functioning were limited in terms of their specificity to particular domains of psychopathology: Each of the external criteria exhibited very similar patterns of association with the fear, distress, externalizing, and general factors in the higher-order model, for example. While this limited our ability to test the sensitivity of each model for teasing apart differential associations, the consistency of the relationships was a useful benchmark for comparing models (e.g., comparing factors within and between models and statistical frameworks). Future research should also examine domain-specific correlates of psychopathology to compare these models' abilities to uniquely characterize such relationships.

Additionally, both psychopathology and external criteria were collected from the same informant, so it is likely that response biases inflated the covariation among psychopathology and indices of functioning, as well as between diagnoses. However, the finding that the

psychopathology dimensions were not uniformly associated with the outcomes suggested some differentiation among psychopathology and external criteria reported by the same informant (e.g., obesity versus being fired or laid off versus experiencing a financial crisis). We encourage and look forward to additional studies using bipolar indicators of dimensional psychopathology diagnoses assessed in a multi-method/multi-informant context.

The disorders included in our models were also relatively limited in number. This presents two notable challenges to modeling the latent structure of psychopathology. The first is a rather circumscribed issue with the specification of the higher-order model. Only three first-order factors were used to indicate the higher-order general psychopathology factor, which renders it a just-identified direct reparametrization of the inter-factor correlations in the correlated factor model. This means that the correlated factor and higher-order models cannot be distinguished statistically. The second challenge pertains to the extent to which a general factor in our models is truly general to all of psychopathology, given that our models did not include disorders comprising thought disorder (e.g., schizophrenia, mania; Caspi et al., 2014, Keyes et al., 2013), somatoform (e.g., somatic symptom disorder; Kotov, Ruggero, et al., 2011), detachment (e.g., schizoid PD; Wright & Simms, 2015), and neurodevelopmental (e.g., autism; Noordhof et al., 2015) dimensions. Other research has included these dimensions in structural models of psychopathology, although it is rare for a study to include thorough coverage of all of the aforementioned dimensions.

Given the large proportion of traditional internalizing disorders in our models, the general factor tended to be marked by distress and fear disorders, similarly to previous studies (e.g., Lahey et al., 2012; Kim & Eaton, 2015). However, the lack of coverage of all major dimensions of psychopathology suggests that our general factor may differ from those of studies that include a broader or different set of disorders (e.g., Caspi et al., 2014; Noordhof et al., 2015). Future studies should consider including a broader variety of common and

uncommon psychopathology, including four or more first-order factors to differentiate the correlated factor and higher-order models, and could formally test the replicability of general factors between studies with varied measurement of psychopathology.

We used the NESARC data despite their limitations given the substantial literature on the empirical structure of psychopathology in these data, the very large sample size to increase precision in parameter estimates and facilitate comparison of the models of interest, the opportunity to examine longitudinal stability and prospective prediction of the models, and with the rationale that all of the models would be affected by the same limitations of the data so they could be compared on even footing. We encourage future comparisons of these models in other data to test the robustness and generalizability of our results.

Conclusion

Taken together, our results support three recommendations. First, researchers are urged to go beyond model fit in adjudicating which structural model to use in their analyses. The bifactor model fit best in our analyses, but the specific factors had marked unreliability that would not have been detected had we relied on traditional fit indices for model selection. Second, it is important to attend to (un)reliability of any latent variables estimated, and to quantify factor determinacy where possible. If factor scores are used as predictors or outcomes of external correlates, we suggest that sensitivity analyses are conducted by estimating the latent variables and regression coefficients in an SEM framework. Third, researchers should be cautious in partitioning the unique variance in narrower (e.g., fear, distress, and externalizing) factors to examine associations with external correlates—whether using the bifactor model, residuals of first-order factors in the higher-order model, or correlated factors in a multiple regression framework—given challenges in interpretation and the potential for lower reliability. If these approaches are used, results for an unpartialled correlated factor model should also be examined for comparison. We also found that all latent

general factors of psychopathology tended to perform similarly well across tests of reliability and validity here, suggesting more than one modeling framework may be suitable for research focusing exclusively on a general factor of psychopathology. The single-factor model did not have close fit to these (multidimensional) data, but could be useful to compute optimally weighted scores to approximate a general factor without much loss of reliability or validity. Similarly, the count of diagnoses evidently lost important information, but the number of common mental disorders for which an individual meets criteria may also have utility as a crude index of general psychopathology (e.g., in under-resourced clinical contexts). Overall, these results provide new insights into popular statistical models used to understand the nature, causes, and consequences of psychopathology that we hope will strengthen research in this area.

References

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: a factor-analytic study. *Psychological Monographs: general and applied, 80*(7), 1.

Afzali, M. H., Sunderland, M., Carragher, N., & Conrod, P. (2018). The structure of psychopathology in early adolescence: study of a canadian sample: la structure de la psychopathologie au debut de l'adolescence: etude d'un echantillon canadien. The *Canadian Journal of Psychiatry, 63*(4), 223-230.

Albott, C. S., Forbes, M. K., & Anker, J. J. (2018). Association of childhood adversity with differential susceptibility of transdiagnostic psychopathology to environmental stress in adulthood. *JAMA Network Open*, *1*, e185354-e185354.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Beauducel, A., & Hilger, N. (2017). On the bias of factor score determinacy coefficients based on different estimation methods of the exploratory factor model. *Communications in Statistics-Simulation and Computation*, *46*(8), 6144-6154.

Blanco, C., Wall, M. M., He, J.-P., Krueger, R. F., Olfson, M., Jin, C. J., . . . Merikangas, K. R. (2015). The space of common psychiatric disorders in adolescents: comorbidity structure and individual latent liabilities. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(1), 45-52.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*, 465-484.

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*, 184-186.

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: appreciating benefits and limitations. *Biological Psychiatry*.

Brandes, C. M., Herzhoff, K., Smack, A. J., & Tackett, J. L. (2019). The p factor and the n factor: Associations between the general factors of psychopathology and neuroticism in children. *Clinical Psychological Science*, *7*, 1266-1284.

Brislin, S. J., & Patrick, C. J. (2019). Callousness and Affective Face Processing: Clarifying the Neural Basis of Behavioral-Recognition Deficits Through the Use of Brain Event-Related Potentials. *Clinical Psychological Science, 7*(6), 1389-1402.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of personality*, *80*(4), 796-846.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Carragher, N., Krueger, R. F., Eaton, N. R., & Slade, T. (2015). Disorders without borders: current and future directions in the meta-structure of mental disorders. *Social psychiatry and psychiatric epidemiology, 50*(3), 339-350.

Carragher, N., Teesson, M., Sunderland, M., Newton, N., Krueger, R., Conrod, P., . . . Slade, T. (2016). The structure of adolescent psychopathology: a symptom-level analysis. *Psychological Medicine, 46*(5), 981-994.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., . . . Poulton, R. (2014). The p factor one general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science, 2*(2), 119-137.

Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *American Journal of Psychiatry, 175*(9), 831-844.

Castellanos-Ryan, N., Brière, F. N., O'Leary-Barrett, M., Banaschewski, T., Bokde, A., Bromberg, U., ... & Garavan, H. (2016). The structure of psychopathology in adolescence and its common personality and cognitive correlates. *Journal of abnormal psychology, 125*(8), 1039.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*(2), 233-255.

Clark, D. A., Hicks, B. M., Angstadt, M., Rutherford, S., Taxali, A., Hyde, L. W., … Sripada, C. (under review). The General Factor of Psychopathology in the Adolescent Brain Cognitive Development (ABCD) Study: A Comparison of Alternative Modeling Approaches. https://doi.org/10.31234/osf.io/b6uy7

Conway, C. C., Mansolf, M., & Reise, S. P. (2019). Ecological validity of a quantitative classification system for mental illness in treatment-seeking adults. *Psychological Assessment*.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the" problem" of sample size: A clarification. *Psychological bulletin, 109*(3), 512.

Del Giudice, M. (2016). The life history model of psychopathology explains the structure of psychiatric disorders and the emergence of the p factor: a simulation study. Clinical *Psychological Science, 4*(2), 299-311.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13*(4), 354-378.

Eaton, N. R. (2014). Transdiagnostic psychopathology factors and sexual minority mental health: Evidence of disparities and associations with minority stressors. *Psychology of Sexual Orientation and Gender Diversity, 1*(3), 244-254.

Eaton, N. R., Keyes, K. M., Krueger, R. F., Balsis, S., Skodol, A. E., Markon, K. E., . . . Hasin, D. S. (2012). An invariant dimensional liability model of gender differences in mental disorder prevalence: evidence from a national sample. *Journal of Abnormal Psychology, 121*(1), 282-288.

Eaton, N. R., Keyes, K. M., Krueger, R. F., Noordhof, A., Skodol, A. E., Markon, K. E., . . . Hasin, D. S. (2013). Ethnicity and psychiatric comorbidity in a national sample: evidence for latent comorbidity factor invariance and connections with disorder prevalence. *Social Psychiatry and Psychiatric Epidemiology, 48*(5), 701-710.

Eaton, N. R., Krueger, R. F., Markon, K. E., Keyes, K. M., Skodol, A. E., Wall, M., . . . Grant, B. F. (2013). The structure and predictive validity of the internalizing disorders. *Journal of Abnormal Psychology, 122*(1), 86-92.

Eaton, N. R., Krueger, R. F., & Oltmanns, T. F. (2011). Aging and the structure and long-term stability of the internalizing spectrum of personality and psychopathology. *Psychology and Aging, 26*(4), 987-993.

Eaton, N. R., Rodriguez-Seijas, C., Carragher, N., & Krueger, R. F. (2015). Transdiagnostic factors of psychopathology and substance use disorders: a review. *Social psychiatry and psychiatric epidemiology, 50*(2), 171-182.

Eaton, N. R., South, S. C., & Krueger, R. F. (2010). The meaning of comorbidity among common mental disorders. In T. Millon, R. F. Krueger & E. Simonsen (Eds.), *Contemporary Directions in Psychopathology: Scientific Foundations of the DSM-V and ICD-11* (pp. 223-241). New York, NY, US: Guilford Press.

Eid, M. (2020). Multi-Faceted Constructs in Abnormal Psychology: Implications of the Bifactor S-1 Model for Individual Clinical Assessment. *Journal of Abnormal Child Psychology*, 1-6.

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological methods, 22*(3), 541.

Forbes, M. K., Rapee, R. M., & Krueger, R. F. (2019). Opportunities for the prevention of mental disorders by reducing general psychopathology in early childhood. *Behaviour research and therapy, 119*, 103411.

Forbes, M. K., Tackett, J. L., Markon, K. E., & Krueger, R. F. (2016). Beyond comorbidity: Toward a dimensional and hierarchical approach to understanding psychopathology across the life span. *Development and psychopathology, 28*(4pt1), 971-986.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research*, *48*(5), 639-662.

Gluschkoff, K., Jokela, M., & Rosenström, T. (2019). The general psychopathology factor: structural stability and generalizability to within-individual changes. *Frontiers in psychiatry*, *10*, 594.

Grant, B. F., Dawson, D. A., Stinson, F. S., Chou, P. S., Kay, W., & Pickering, R. (2003). The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. *Drug and alcohol dependence*, *71*(1), 7-16.

Greene, A. L., & Eaton, N. R. (2016). Panic disorder and agoraphobia: A direct comparison of their multivariate comorbidity patterns. *Journal of affective disorders, 190*, 75-83.

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., . . . Docherty, A. R. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *Journal of Abnormal Psychology, 128*(7), 740.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliabilitywithin latent variable systems. In R. Cudeck, S. du Toit, & D. Sorbom(Eds.),*Structural equation modeling: Present and future* (pp. 195-216). Lincolnwood, IL: ScientificSoftware International.

Hasin, D. S., Goodwin, R. D., Stinson, F. S., & Grant, B. F. (2005). Epidemiology of major depressive disorder: results from the National Epidemiologic Survey on Alcoholism and Related Conditions. *Archives of general psychiatry*, *62*(10), 1097-1106.

Hasin, D., & Kilcoyne, B. (2012). Comorbidity of psychiatric and substance use disorders in the United States: current issues and findings from the NESARC. *Current opinion in psychiatry, 25*(3), 165.

Hayduk, L. (2014). Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement, 74*(6), 905-926.

Hoertel, N., Franco, S., Wall, M. M., Oquendo, M. A., Kerridge, B. T., Limosin, F., & Blanco, C. (2015). Mental disorders and risk of suicide attempt: a national prospective study. *Molecular Psychiatry*, *20*, 718-726.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated?. *Personality and Social Psychology Review, 14*(3), 332-346.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry, 62*(6), 617-627.

Keyes, K. M., Eaton, N. R., Krueger, R. F., Skodol, A. E., Wall, M. M., Grant, B., ... & Hasin, D. S. (2013). Thought disorder in the meta-structure of psychopathology. *Psychological Medicine*, *43*, 1673-1683.

Kim, H., & Eaton, N. R. (2015). The hierarchical structure of common mental disorders: Connecting multiple levels of comorbidity, bifactor models, and predictive validity. *Journal of Abnormal Psychology*, *124*, 1064-1078.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., ... & Eaton, N. R. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *Journal of abnormal psychology, 126*(4), 454.

Kotov, R., Ruggero, C. J., Krueger, R. F., Watson, D., Yuan, Q., & Zimmerman, M. (2011). New dimensions in the quantitative classification of mental illness. *Archives of General Psychiatry*, *68*, 1003-1011.

Krueger, R. F., Caspi, A., Moffitt, T. E., & Silva, P. A. (1998). The structure and stability of common mental disorders (DSM-III-R): a longitudinal-epidemiological study. *Journal of abnormal psychology, 107*(2), 216.

Krueger, R. F., & Eaton, N. R. (2015). Transdiagnostic factors of mental disorders. *World Psychiatry, 14*(1), 27-29.

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2*, 111-133.

Laceulle, O. M., Chung, J. M., Vollebergh, W. A., & Ormel, J. (2019). The wide-ranging life outcome correlates of a general psychopathology factor in adolescent psychopathology. *Personality and mental health.*

Laceulle, O. M., Vollebergh, W. A., & Ormel, J. (2015). The structure of psychopathology in adolescence replication of a general psychopathology factor in the TRAILS Study. *Clinical Psychological Science, 3*(6), 850-860.

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? *Journal of Abnormal Psychology, 121*(4), 971.

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2016). A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin, 143*(2), 142-186.

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. *Journal of Child Psychology and Psychiatry, 56*(4), 415-422.

Lahey, B. B., Zald, D. H., Perkins, S. F., Villalta-Gil, V., Werts, K. B., Van Hulle, C. A., ... & Watts, A. L. (2018). Measuring the hierarchical general factor model of psychopathology in young adults. *International journal of methods in psychiatric research, 27*(1), e1593.

Levin-Aspenson, H. F., Watson, D., Clark, L. A., & Zimmerman, M. (in press). What is the general factor of psychopathology? Consistency of the p factor across samples. *Assessment.*

Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The perils of partialling: Cautionary tales from aggression and psychopathy. *Assessment*, *13*, 328-341.

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual review of clinical psychology, 15*, 51-69.

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychological bulletin, 137*(5), 856.

Martel, M. M., Pan, P. M., Hoffmann, M. S., Gadelha, A., do Rosário, M. C., Mari, J. J., ... & Rohde, L. A. (2017). A general psychopathology factor (P factor) in children: structural model analysis and external validation through familial risk and child global executive function. *Journal of Abnormal Psychology, 126*(1), 137.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344-362.

McCrae, R. R. (2008). A note on some measures of profile agreement. *Journal of personality assessment, 90*(2), 105-109.

McCrae, R. R., Zonderman, A. B., Costa Jr, P. T., & Paunonen, S. V. (1996). Evaluating Replicability of Factors in the Revised NEO Personality Inventory: Confirmatory Factor Analysis Versus Procrustes Rotation. *Journal of Personality and Social Psychology, 70*(3), 552-566.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: L. Erlbaum Associates.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of applied psychology, 93*(3), 568.

Michelini, G., Barch, D. M., Tian, Y., Watson, D., Klein, D. N., & Kotov, R. (2019). Delineating and validating higher-order dimensions of psychopathology in the Adolescent Brain Cognitive Development (ABCD) study. *Translational psychiatry, 9*(1), 1-15.

Moore, T. M., Kaczkurkin, A. N., Durham, E. L., Jeong, H. J., McDowell, M. G., Dupont, R. M., Applegate, B., Tackett, J. L., Cardenas-Iniguez, C., Kardan, O., Akcelik, G. N., Stier, A. J., Rosenberg, M. D., Hedeker, D., Berman, M. G., & Lahey, B. B. (2020). Criterion validity and relationships between alternative hierarchical dimensional models of general and specific psychopathology. *Journal of Abnormal Psychology.* https://doi.org/10.1037/abn0000601

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (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.

Morin, A. J., Arens, A. K., & Marsh, H. W. (2016). 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.

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.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*(3), 431-462.

Muthén & Muthén (2012). V7.1 MPlus Language Addendum.

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. *Journal of Abnormal Child Psychology*, *43*, 577-587.

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. *Journal of Abnormal Child Psychology*, *46*, 1373-1383.

Patalay, P., Fonagy, P., Deighton, J., Belsky, J., Vostanis, P., & Wolpert, M. (2015). A general psychopathology factor in early adolescence. *The British Journal of Psychiatry, 207*(1), 15-22.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research, 48*(1), 28-56.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry, 170*(1), 59-70.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research, 47*(5), 667-696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of personality assessment*, *95*(2), 129-140.

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 behavioral research, 51*(6), 818-838.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2013). Applying unidimensional item response theory models to psychological data. In K. F. Geisinger, B. A. Bracken, J. F.

Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbooks in psychology®. APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (p. 101–119).

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5-26.

Revelle, W. (2018a). Package 'psych'.

Revelle, W. (2018b). Using the psych package to generate and test structural models.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223-237.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137.

Rodriguez-Seijas, C., Eaton, N. R., & Krueger, R. F. (2015). How transdiagnostic factors of personality and psychopathology can inform clinical assessment and intervention. *Journal of personality assessment*, 1-11.

Rodriguez-Seijas, C., Stohl, M., Hasin, D. S., & Eaton, N. R. (2015). Transdiagnostic Factors and Mediation of the Relationship Between Perceived Racial Discrimination and Mental Disorders. *JAMA Psychiatry, 72*(7), 706-713.

Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., ... & Docherty, A. (2019). Integrating the Hierarchical Taxonomy of Psychopathology

(HiTOP) into clinical practice. *Journal of consulting and clinical psychology, 87*(12), 1069.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321.

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the "best" factor structure and moving measurement validation forward: An illustration. *Journal of personality assessment*, *100*(4), 345-362.

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, 31(12), 1428-1441.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107.

Smith, G. T., Atkinson, E. A., Davis, H. A., Riley, E. N., & Oltmanns, J. R. (2020). The General Factor of Psychopathology. *Annual Review of Clinical Psychology*, *16*.

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. *Clinical Psychological Science*, *5*, 98-110

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. *Psychological medicine, 45*(7), 1483-1493.

Sunderland, M., Forbes, M. K., Mewton, L., Baillie, A. J., Carragher, N., Lynch, S., ... & Teesson, M. (2020). The structure of psychopathology and association with poor sleep, self-harm, suicidality, risky sexual behaviour, and low self-esteem in a population sample of adolescents. *Development and Psychopathology.*

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for

and from clinical psychological science. *Perspectives on Psychological Science, 12*(5), 742-756.

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613-625.

Venables, N. C., Foell, J., Yancey, J. R., Kane, M. J., Engle, R. W., & Patrick, C. J. (2018). Quantifying inhibitory control as externalizing proneness: A cross-domain model. *Clinical Psychological Science, 6*(4), 561-580.

Verheul, R. (2005). Clinical utility of dimensional models for personality pathology. *Journal of personality disorders, 19*(3), 283-302.

Vollebergh, W. A., Iedema, J., Bijl, R. V., de Graaf, R., Smit, F., & Ormel, J. (2001). The structure and stability of common mental disorders: the NEMESIS study. *Archives of General Psychiatry, 58*(6), 597-603.

Waldman, I.D., Poore, H.E., Watts, A.L., Rathouz, P., Van Hulle, C., Zald, D., & Lahey, B. (2017). Issues in the validation of the general factor of psychopathology. Paper presented at the annual meeting of the Behavior Genetics Association, Oslo, Norway, June 21-23, 2017.

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier tests of the validity of the bifactor model of psychopathology. *Clinical Psychological Science*, *7*, 1285-1303.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child development perspectives, 4*(1), 10-18.

Widiger, T. A., & Oltmanns, J. R. (2016). The general factor of psychopathology and personality. *Clinical Psychological Science*.

Widiger, T. A., & Trull, T. J. (2007). Plate tectonics in the classification of personality disorder: Shifting to a dimensional model. *American Psychologist, 62*(2), 71.

Woods, W.C., Wright, A.G.C., Skodol, A.E., Morey, L.C., & Hopwood, C.J. (2020). Deconstructing individual differences in long-term personality disorder and trait change. *Clinical Psychological Science, 8*(1) 184–197.

Wright, A. G., & Simms, L. J. (2015). A metastructural model of mental disorders and pathological personality traits. *Psychological Medicine*, *45*, 2309-2319.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β and Mcdonald'sωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133.

Table 1. Standardized loadings on the narrow (fear, distress, and externalizing) and general factors for the four latent variable models using the weighted least square mean and variance adjusted (WLSMV) estimator

Indicator variable	Wave 1						Wave 2					
	Corr. Factors	Higher-Order		Bifactor		Single-Factor	Corr. Factors	Higher-Order		Bifactor		Single-Factor
	Narrow	Narrow	General	Narrow	General	General	Narrow	Narrow	General	Narrow	General	General
Major depressive disorder	0.90	0.90	-	0.43	0.78	0.77	0.85	0.85	-	0.47	0.76	0.76
Generalized anxiety disorder	0.78	0.78	-	0.20	0.75	0.71	0.82	0.82	-	0.25	0.75	0.76
Dysthymia	0.84	0.84	-	0.51	0.71	0.77	0.78	0.78	-	0.36	0.68	0.74
Panic disorder/ agoraphobia	0.87	0.87	-	0.60	0.69	0.73	0.94	0.94	-	0.51	0.80	0.85
Social phobia	0.78	0.78	-	0.38	0.64	0.61	0.85	0.85	-	0.36	0.73	0.75
Specific phobia	0.69	0.69	-	0.42	0.56	0.54	0.64	0.64	-	0.48	0.52	0.56
Adult antisocial behavior	0.80	0.80	-	0.64	0.47	0.69	0.77	0.77	-	0.63	0.46	0.63
Nicotine dependence	0.69	0.69	-	0.52	0.44	0.61	0.60	0.60	-	0.43	0.38	0.46
Alcohol dependence	0.78	0.78	-	0.70	0.42	0.70	0.71	0.71	-	0.63	0.40	0.56
Cannabis dependence	0.82	0.82	-	0.62	0.53	0.77	0.79	0.79	-	0.60	0.50	0.69
Other drug dependence	0.86	0.86	-	0.64	0.55	0.79	0.84	0.84	-	0.60	0.55	0.71
Distress	-	-	0.89	-	-	-	-	-	0.90	-	-	-
Fear	-	-	0.81	-	-	-	-	-	0.84	-	-	-
Externalizing	-	-	0.60	-	-	-	-	-	0.61	-	-	-
Inter-factor correlations												
Fear with Distress	0.72	-	-	0.00	0.00	-	0.76	-	-	0.00	0.00	-
Fear with Externalizing	0.48	-	-	0.00	0.00	-	0.51	-	-	0.00	0.00	-
Externalizing with Distress	0.53	-	-	0.00	0.00	-	0.55	-	-	0.00	0.00	-

Note. Corr. Factors = correlated factor model.

Table 2. Model fit indices

		Wave 1				Wave 2			
Estimator	Fit Index	CF	HO	Bifactor	1F	CF	HO	Bifactor	1F
	k	25	25	33	22	25	25	33	22
WLSMV	RMSEA	0.011	0.011	**0.010**	0.049	0.008	0.008	**0.007**	0.029
	CFI	0.994	0.994	**0.996**	0.871	0.994	0.994	**0.996**	0.903
	TLI	0.992	0.992	**0.994**	0.839	0.992	0.992	**0.993**	0.879
MLR	AIC	223527	223529	**223322**	229795	120092	120092	**120041**	122048
	BIC	223743	223746	**223608**	229985	120303	**120303**	120320	122234
	SSABIC	223664	223666	**223503**	229915	120224	120224	**120215**	122164

Note: Wave 1 (n = 43093) measures lifetime disorders; Wave 2 (n = 34653) measures disorders since last interview. CF = correlated-factors model, HO = higher order model, 1F = single-factor model, WLSMV = weighted least square mean and variance adjusted, MLR = robust maximum likelihood, k = number of estimated parameters, RMSEA = root mean square error of approximation, CFI = comparative fit index, TLI = Tucker-Lewis index, AIC = Akaike's information criterion, BIC = Bayesian information criterion, SSA = sample-size adjusted. Results in bold denote the closest fit for each index within each wave, although only the information criteria can be used to compare models' fit.

Table 3. Reliability indices for the bifactor, correlated factor, and single-factor model.

Index	Factor	Bifactor		Correlated factor		Single-factor	
		Wave 1	Wave 2	Wave 1	Wave 2	Wave 1	Wave 2
H	General psychopathology	**0.88**	**0.89**			**0.92**	**0.92**
	Distress	0.38	0.33	**0.89**	**0.86**		
	Fear	0.49	0.44	**0.85**	**0.92**		
	Externalizing	0.77	0.73	**0.9**	**0.87**		
ω_h / ω_t	General psychopathology	0.72	0.73			**0.91**	**0.91**
ω_{hs} / ω_s	Distress	*0.18*	*0.17*	**0.88**	**0.86**		
	Fear	*0.30*	*0.26*	**0.83**	**0.86**		
	Externalizing	0.56	0.53	**0.89**	**0.86**		
ECV	General psychopathology	0.57	0.6				
	Distress	0.07	0.06				
	Fear	0.10	0.09				
	Externalizing	0.27	0.25				
ECV_S	Distress	0.22	0.20				
	Fear	0.36	0.30				
	Externalizing	0.62	0.61				
PUC		**0.71**	**0.71**				
APB		*0.25*	*0.18*				

Note: H = index of construct replicability (ideally H > .8), ω_h = omega hierarchical (used for bifactor GP; ideally ω_h > .8), ω_t = omega total (used for single-factor GP; ideally > .75), ω_{hs} = omega hierarchical subscale (used for bifactor D, F, and E; ω_{hs} > .75 indicates sufficient reliability to be used in practice; and ω_h/ω_{hs} < .5 indicates insufficient precision such that the factor should not be used in practice), ω_s = omega subscale (used for correlated-factors D, F, and E; ideally > .75), ECV = explained common variance (i.e., ideally ECV > .7, and > .85 if there is evidence of unidimensionality), ECV_S = ECV for specific factors (ideally ECV_S >.7), PUC = percent of uncontaminated correlations (PUCs > .7 provide evidence for unidimensionality), APB = average parameter bias (10-15% is acceptable). Values indicating an appropriate level of reliability are bolded. Values of ω and APB indicating unreliability are italicized. Greyed out cells indicate indices only applicable to the bifactor model. Indices could not be calculated for the untransformed higher-order model; indices for the Schmid-Leiman transformed model are presented in Table S1.

Table 4. Spearman rank-order correlations for individuals' estimated factor scores or observed count at wave 1 and wave 2

Latent variable	Correlated factor	Higher-Order	Bifactor	Single-Factor	Count Variable
General Psychopathology	-	0.36	0.36	0.37	0.39
Distress	0.35	0.35	0.19	-	-
Fear	0.35	0.35	0.19	-	-
Externalizing	0.38	0.38	0.32	-	-

Figure Captions.

Figure 1. The four latent variable models to be compared, alongside a count variable of diagnoses. Model 1 is the correlated factor model. Model 2 is the higher-order factor model. Model 3 is the bifactor model. Model 4 is the single-factor model.

Figure 2. Standardized standard errors of factor loadings for wave 1 (top) and wave 2 (bottom) models using the WLSMV estimator. MDD = major depressive disorder, GAD = generalized anxiety disorder, DYS = dysthymia, PDA = panic disorder and agoraphobia, SOC = social phobia, SPH = specific phobia, AAB = adult antisocial behaviour, NIC = nicotine dependence, ALC = alcohol dependence, CAN = cannabis dependence, DRG = other drug dependence.

Figure 3. Variance accounted for in each outcome by the specific factors in the correlated factor, higher-order, and bifactor models. The R^2 value regressing Wave 1 'accomplished less at work' on bifactor distress was undefined due to a non-positive definite psi matrix. Values labeled with a star exceed the scale of the y-axis; see the supplemental materials (Figure S2) for a full-scale version of this figure. Fired = being fired/laid off from a job in the past year; Unemployed = unemployed and looking for a job for more than a month in the past year; Relationship Breakdown = separated, divorced, or broke off steady relationship in the past year; Financial Crisis = experienced a major financial crisis, declared bankruptcy, or was unable to pay bills in the past year; Fair/Poor Physical Health = fair or poor self-perceived current physical health; Accomplished Less At Work = accomplished less than would like or did work/other activities less carefully than usual most or all of the time due to emotional problems in the past four weeks; Illness = chronic illness diagnosis (hardening of arteries, high blood pressure, chest pain/angina, rapid heartbeat, heart attack, liver disease/cirrhosis, heart disease, ulcer, gastritis, or arthritis) confirmed by a health professional in the past year; Obesity = body mass index ≥ 30.

Figure 4. Variance accounted for in each outcome by the general factors in the bifactor, higher-order, and single-factor models as well as the count of diagnoses. Fired = being fired/laid off from a job in the past year; Unemployed = unemployed and looking for a job for more than a month in the past year; Relationship Breakdown = separated, divorced, or broke off steady relationship in the past year; Financial Crisis = experienced a major financial crisis, declared bankruptcy, or was unable to pay bills in the past year; Fair/Poor Physical Health = fair or poor self-perceived current physical health; Accomplished Less At Work = accomplished less than would like or did work/other activities less carefully than usual most or all of the time due to emotional problems in the past four weeks; Illness = chronic illness diagnosis (hardening of arteries, high blood pressure, chest pain/angina, rapid heartbeat, heart attack, liver disease/cirrhosis, heart disease, ulcer, gastritis, or arthritis) confirmed by a health professional in the past year; Obesity = body mass index ≥ 30.

Figure 5. Left: the profiles of the ten randomly selected individuals for the constrained (red lines) versus the unconstrained (green lines) longitudinal measurement invariance models for the bifactor (top) and higher-order (right) models. Right: The differences in the estimated factor scores for the constrained versus the unconstrained models

Figure 1.

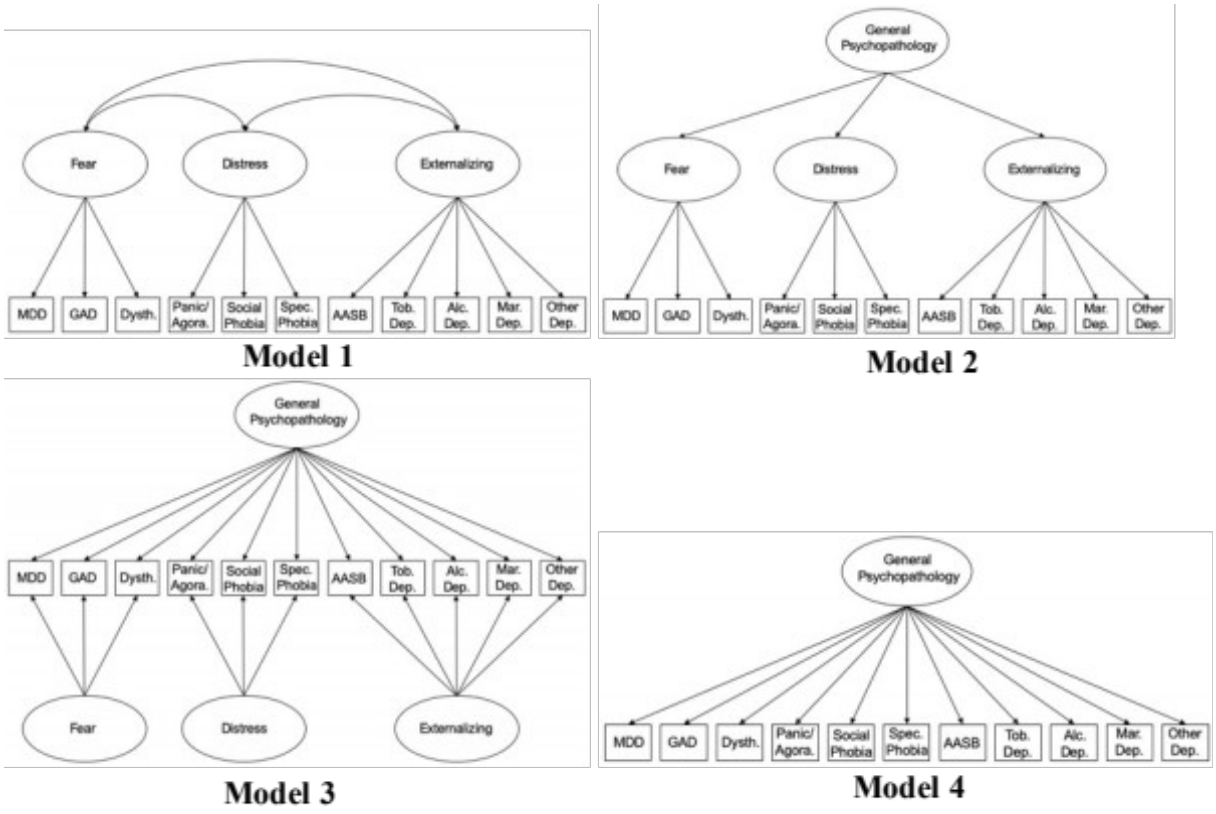

Fear
Distress
Externalizing
MDD
GAD
Dysth.
Panic/ Agora.
Social Phobia
Spec. Phobia
AASB
Tob. Dep.
Alc. Dep.
Mar. Dep.
Other Dep.
Model 1
General Psychopathology
Model 2
Model 3
Model 4

Figure 2.

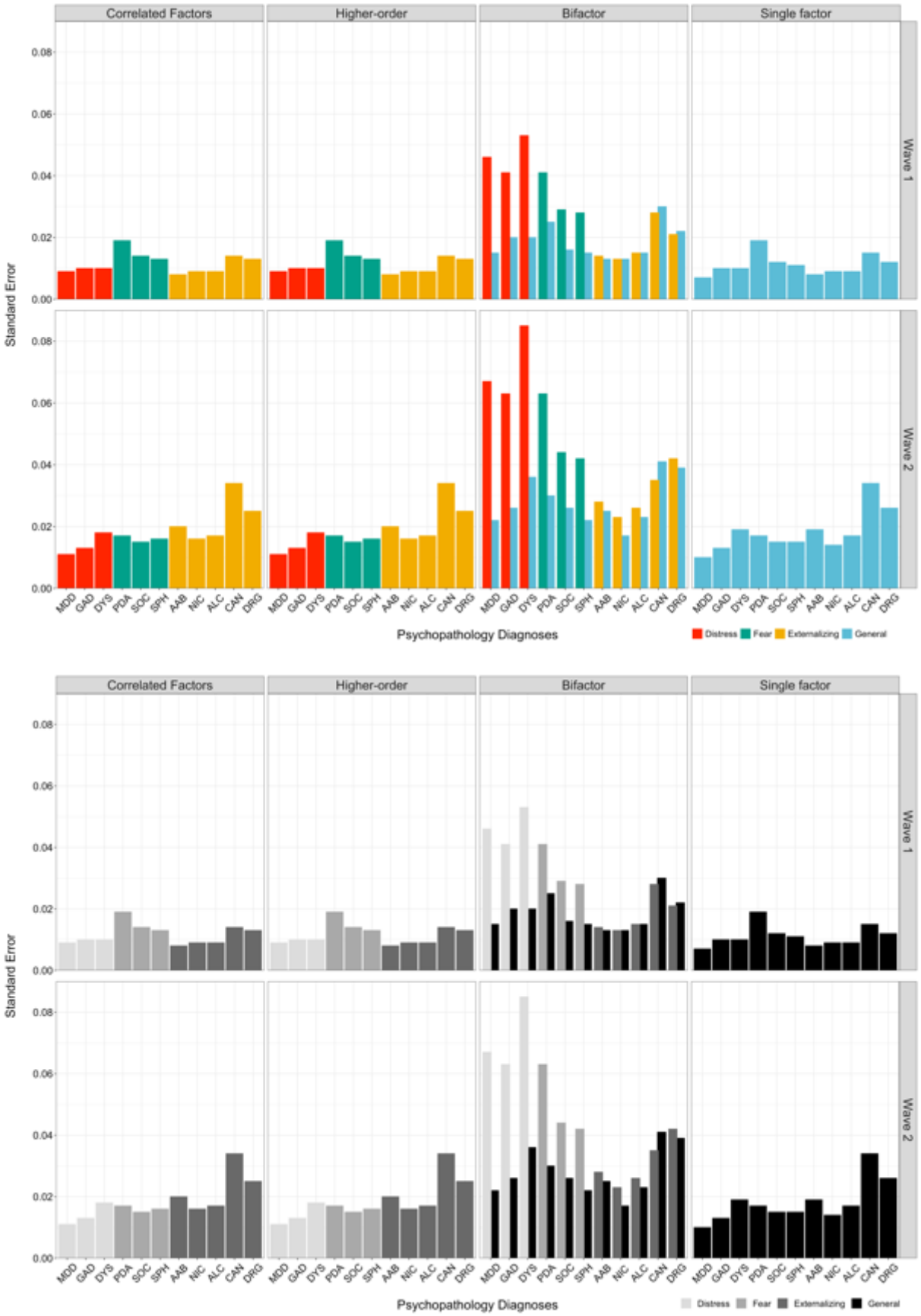

Correlated Factors
Higher-order
Bifactor
Single factor
Wave 1
Wave 2
Standard Error
0.00
0.02
0.04
0.06
0.08
MDD GAD DYS PDA SOC SPH AAB NIC ALC CAN DRG
Psychopathology Diagnoses
Distress
Fear
Externalizing
General

Figure 3 (colour).

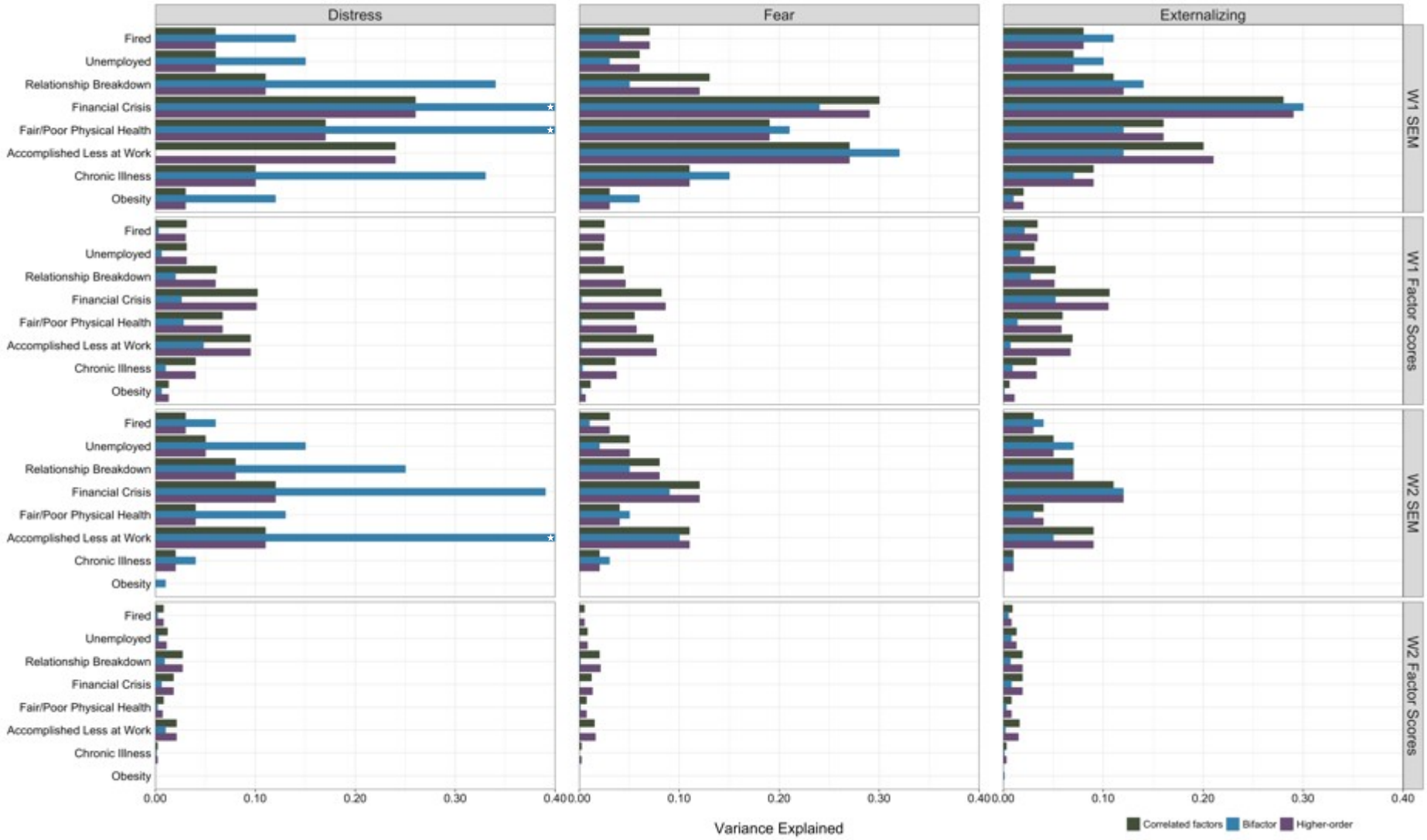


Figure 3 (black & white).

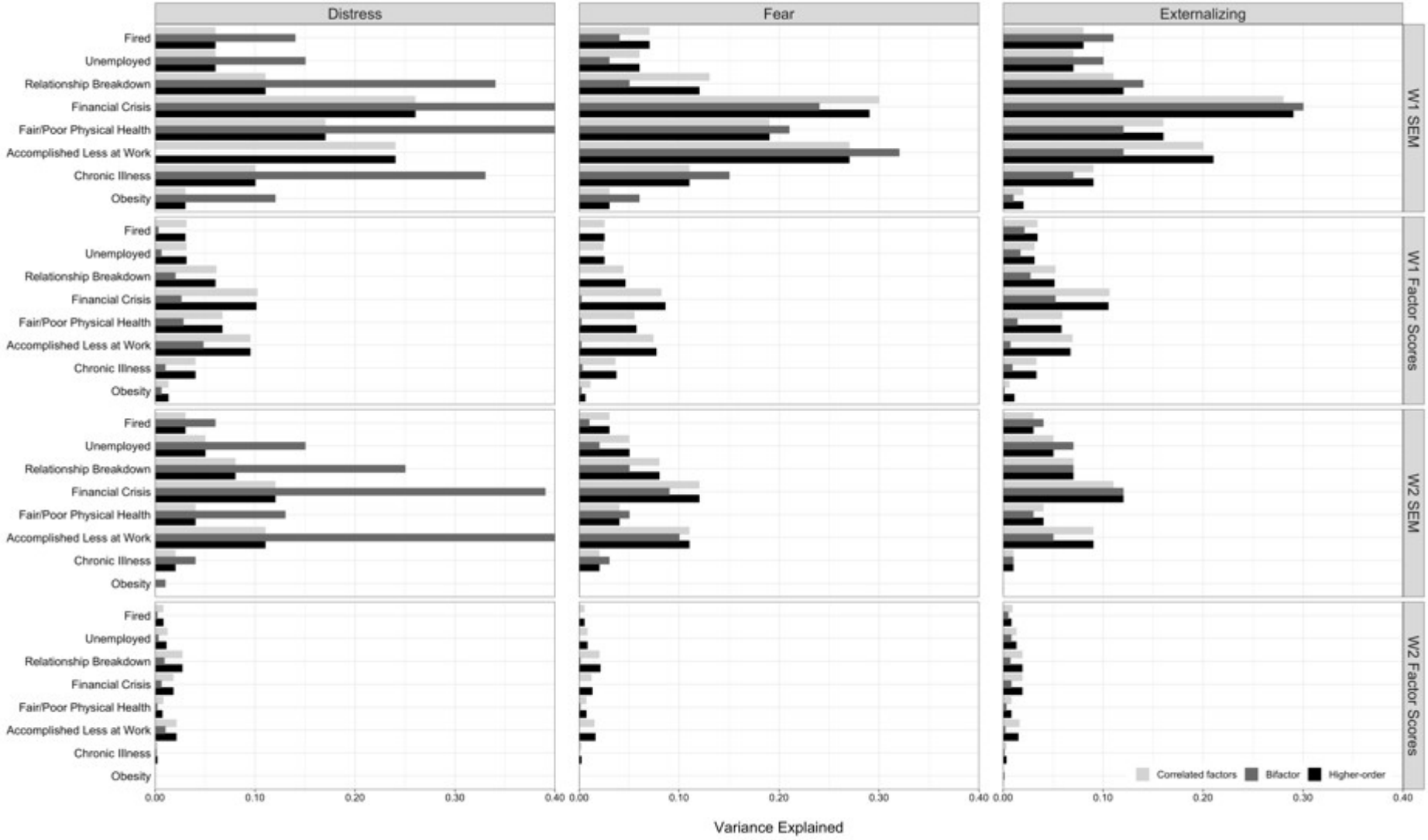


Figure 4 (colour).

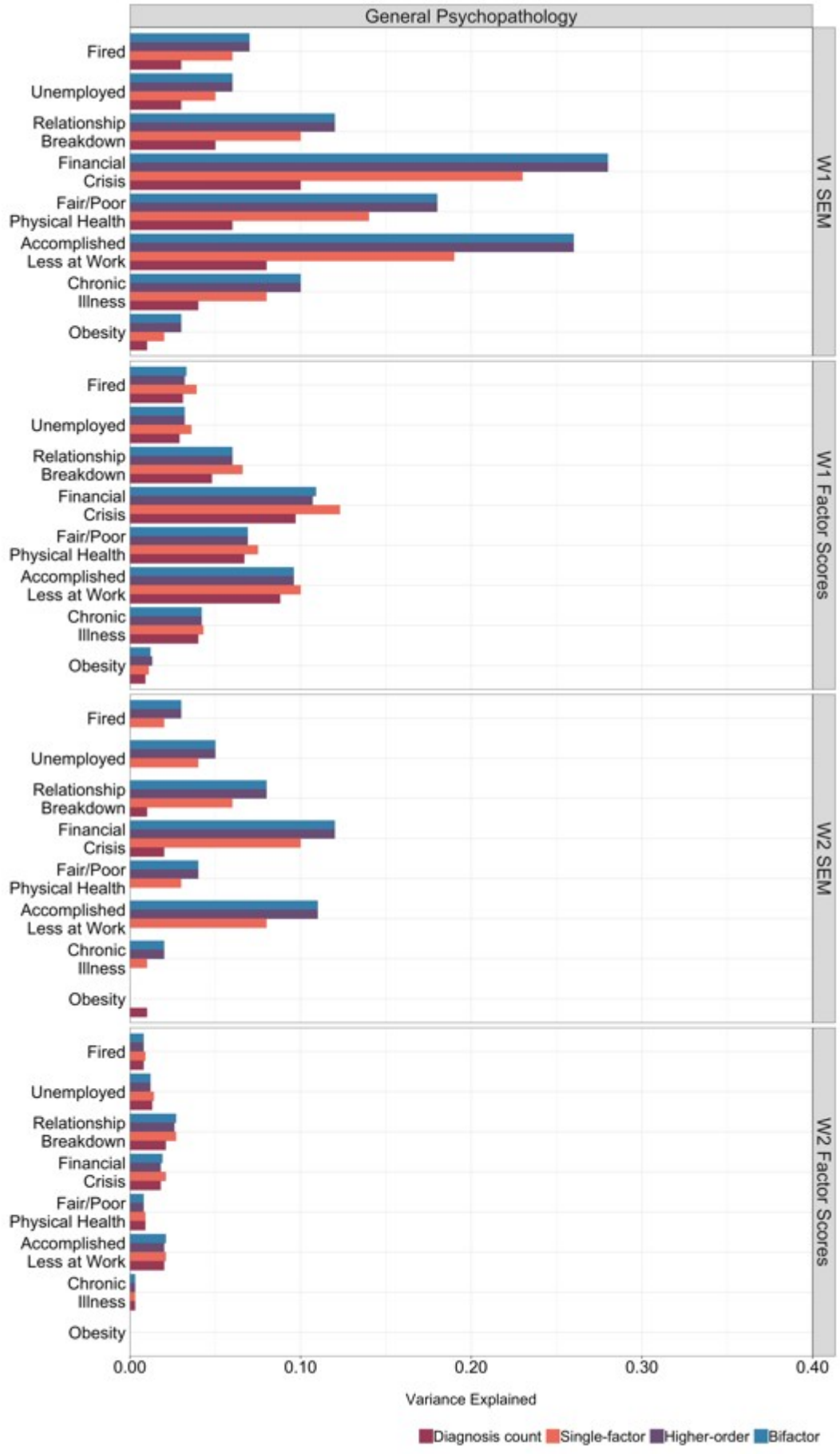


Figure 4 (black & white).

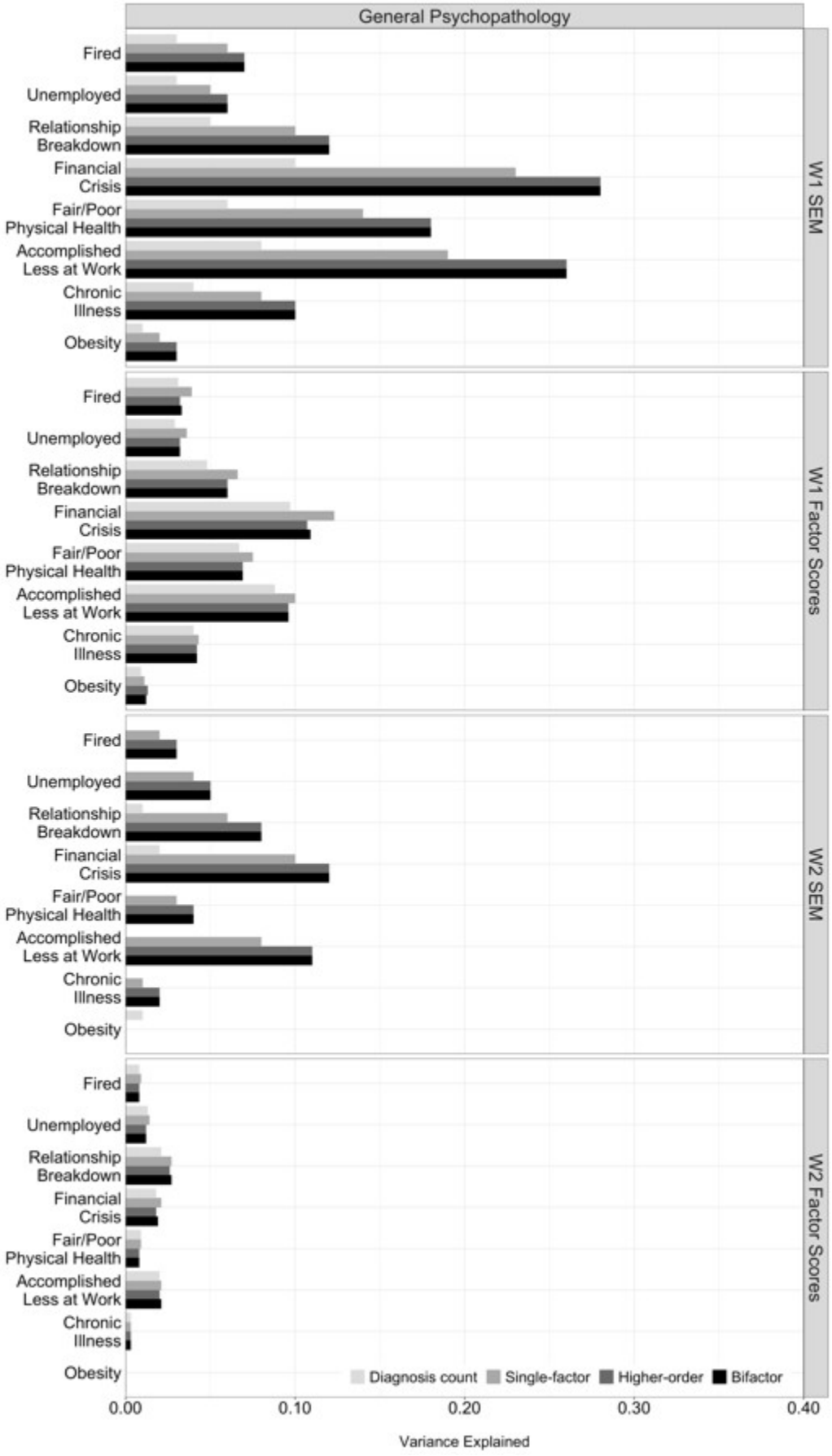
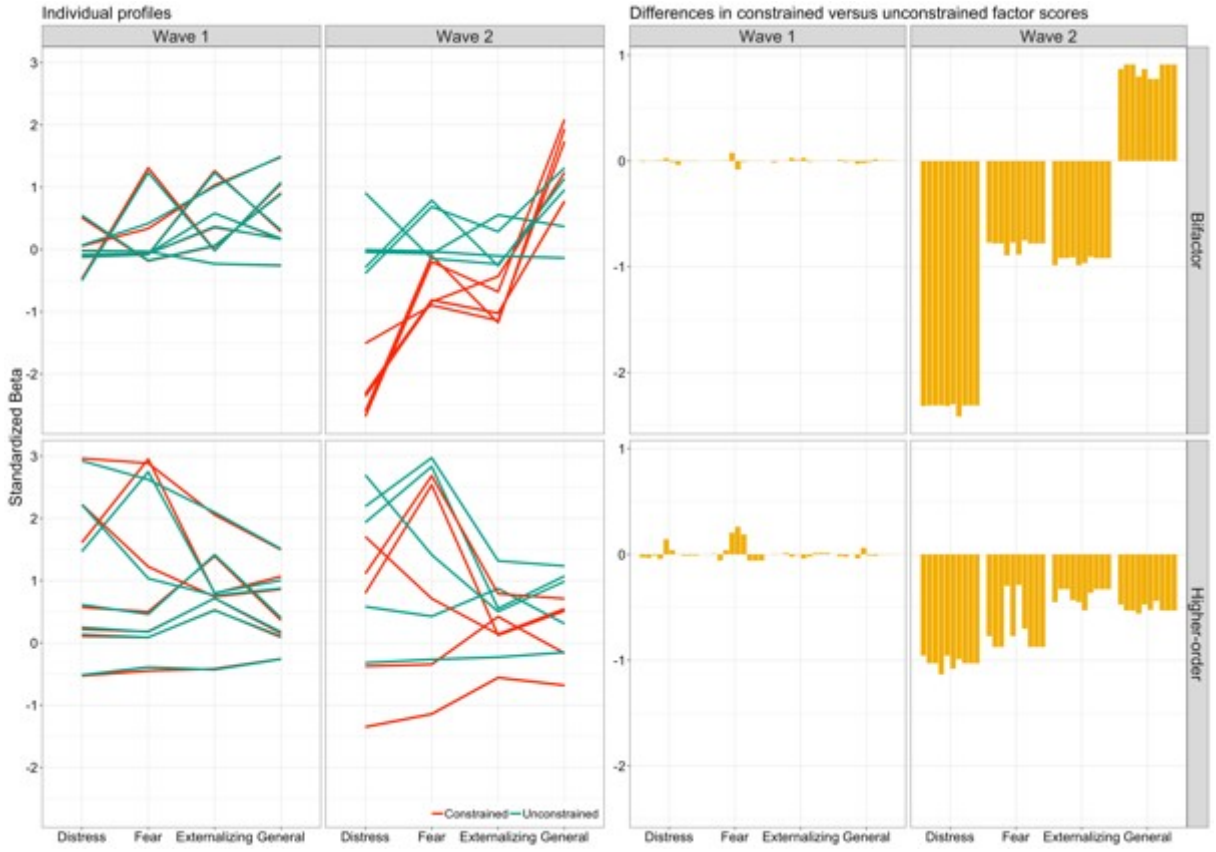
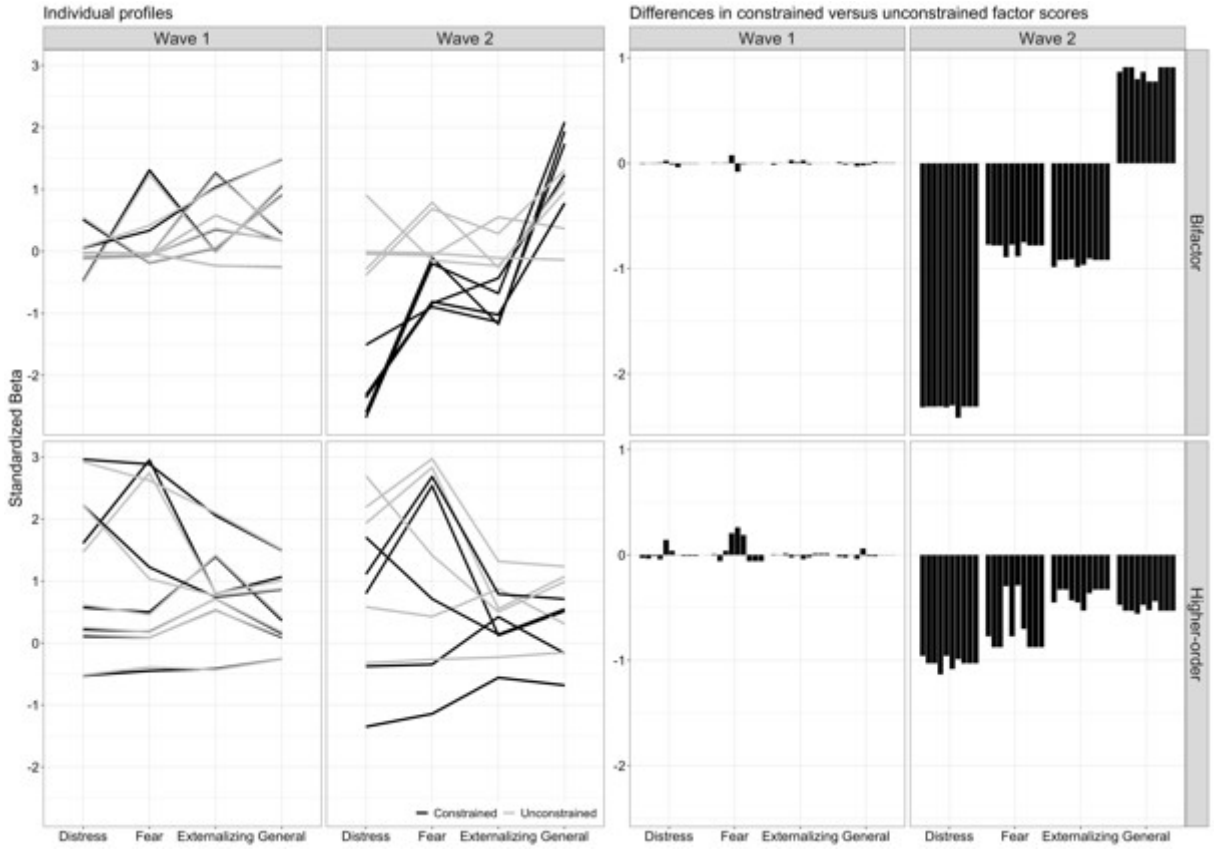


Figure 5.


Individual profiles
Differences in constrained versus unconstrained factor scores
Wave 1
Wave 2
Standardized Beta
Bifactor
Higher-order
Constrained
Unconstrained
Distress
Fear
Externalizing General


Individual profiles
Differences in constrained versus unconstrained factor scores
Wave 1
Wave 2
Standardized Beta
Bifactor
Higher-order
Constrained
Unconstrained
Distress
Fear
Externalizing General