Quantifying the strength of general factors in psychopathology: A comparison of CFA with maximum likelihood estimation, BSEM and ESEM/EFA bi-factor approaches

Murray, A L ; Booth, T ; Eisner, Manuel ; Obsuth, I ; Ribeaud, D

Abstract: Whether or not importance should be placed on an all-encompassing general factor of psychopathology (or p factor) in classifying, researching, diagnosing, and treating psychiatric disorders depends (among other issues) on the extent to which comorbidity is symptom-general rather than staying largely within the confines of narrower transdiagnostic factors such as internalizing and externalizing. In this study, we compared three methods of estimating p factor strength. We compared omega hierarchical and explained common variance calculated from confirmatory factor analysis (CFA) bifactor models with maximum likelihood (ML) estimation, from exploratory structural equation modeling/exploratory factor analysis models with a bifactor rotation, and from Bayesian structural equation modeling (BSEM) bifactor models. Our simulation results suggested that BSEM with small variance priors on secondary loadings might be the preferred option. However, CFA with ML also performed well provided secondary loadings were modeled. We provide two empirical examples of applying the three methodologies using a normative sample of youth (z-proso, n = 1,286) and a university counseling sample (n = 359).

DOI: https://doi.org/10.1080/00223891.2018.1468338

Quantifying the strength of general factors in psychopathology: A comparison of CFA with maximum likelihood estimation, BSEM and ESEM/EFA bi-factor approaches

Aja Louise Murray[1*], Tom Booth[2], Manuel Eisner[1], Ingrid Obsuth[1] Denis Ribeaud[3]

[1]Violence Research Centre, Institute of Criminology, University of Cambridge,

[2]Department of Psychology, University of Edinburgh

[3]Jacobs Center for Productive Youth Development, University of Zurich

*Corresponding author at Institute of Criminology, Sidgwick Avenue, Cambridge, CB3 9DA.

Email: am2367@cam.ac.uk

Abstract

Whether or not importance should be placed on an all-encompassing general factor of psychopathology (or *p*-factor) in classifying, researching, diagnosing and treating psychiatric disorders depends (amongst other issues) on the extent to which co-morbidity is symptom-general rather than staying largely within the confines of narrower trans-diagnostic factors such as internalising and externalising. In this study we compared three methods of estimating *p*-factor strength. We compared omega hierarchical and ECV calculated from CFA bi-factor models with maximum likelihood (ML) estimation, from ESEM/EFA models with a bifactor rotation, and from BSEM bi-factor models. Our simulation results suggested that BSEM with small variance priors on secondary may be the preferred option. However, CFA with ML also performed well provided secondary loadings were modelled We provide two empirical examples of applying the three methodologies using a normative sample of youth (z-proso, n=1286) and University counselling sample (n= 359).

Keywords: p-factor; general factor of psychopathology, comorbidity, trans-diagnostic factors, bi-factor, BSEM, ESEM

Conceptualising covariation among psychopathological symptoms in terms of broad trans-diagnostic factors, such as internalising and externalising, is now fairly uncontroversial (e.g. Krueger & Markon, 2006). Whether the concept of an all-encompassing general factor of psychopathology (or '*p*-factor') adds scientific and clinical value is less clear. The importance placed on this hypothetical *p*-factor should at least in part be based on the extent to which comorbidity cuts across all symptoms of all disorders, rather than staying largely within the confines of established trans-diagnostic factors. However, there are no definitive guidelines on how best to judge *p*-factor strength. In this study, we use a combination of simulation and real data examples to evaluate Exploratory factor analysis (EFA), Confirmatory factor analysis (CFA) and Bayesian structural equation modeling (BSEM) bifactor-based methods of quantifying general factor strength in the context of psychopathology models.

Across the history of psychopathology research, there has been a shift from a view of psychiatric disorders as distinct categorical entities, to one in which symptoms co-vary across traditional diagnostic boundaries (e.g. see Angold, Costello & Erkanli, 1999; Eaton, Rodriguez-Seijas, Carragher & Krueger, 2015; Kessler et al., 2012). Patterns of covariation suggest that symptoms can be organised hierarchically with a small number of broad 'trans-diagnostic' dimensions at the most general level. Based largely on factor analytic studies, research in this paradigm has primarily focussed on the broad dimensions of internalising and externalising (e.g. Kendler, Prescott, Myers & Neale, 2003; Kramer, Krueger & Hicks, 2008; Krueger & Markon, 2006). Internalising encompasses symptoms from diagnostic categories such as major depression, general anxiety disorder, dysthymia, phobias, post-traumatic stress disorder and panic disorder. Externalising encompasses symptoms belonging to diagnostic categories such as substance use disorder, conduct disorder and others with behavioural disinhibition, impulsivity or 'acting out' as a prominent feature.

However, psychiatric comorbidity frequently occurs even between symptoms belonging to different trans-diagnostic dimensions. At the latent dimension level, for example, correlations among internalising, externalising and thought disorder can be around .40 to .50 (e.g., Wright, Krueger, Hobbs, Markon, Eaton & Slade, 2013). In confirmatory factor analysis (CFA) models, adding a general psychopathology dimension atop or alongside these trans-diagnostic factors tends to yield good- and where tested- better fitting models than those without a general dimension (e.g. Caspi et al., 2014; Noordhof, Krueger, Ormel, Oldehinkel & Hartman, 2015; Patalay, Fonagy Deighton, Belsjy, Vostanis & Wolpert, 2015). Observations of this kind have led to the consideration of the possible role of an all-encompassing general dimension of psychopathology – labelled the *p*-factor by Caspi et al. (2014) - in psychopathology research and clinical practice.

Incorporating a *p*-factor into the systems used to organise psychopathology would have a number of potentially important implications. In clinical settings it would encourage assessment and diagnostic systems that consider symptomology across the entire landscape of psychopathology simultaneously. This would represent a significant contrast to current differential diagnosis processes that attempt to home in on just one 'best-fitting' disorder (e.g. Rodriguez-Seijas, Eaton & Krueger, 2015). In countries where provision of mental health services is linked to having been assigned a specific diagnosis, it may lead to significant policy reform, allowing access to services for those who do not neatly fit into any one specific diagnostic category. In research settings it may promote a search for broadband shared etiological factors and treatments. It could also prompt a re-interpretation of existing empirical findings, considering the possibility that many previous attempts to identify correlates of specific psychopathologies have been confounded by non-disorder-specific influences. For example, some preliminary evidence suggests that sex differences in internalising and externalising may be under-estimated when the *p*-factor is not controlled

for, while many disorder-specific risk factors may, in fact, reflect general psychopathology risk factors (e.g. Caspi et al., 2014; Patalay et al., 2015). Clarifying the generality versus specificity of risk factors has potential to inform early prevention and intervention strategies; where more general risk factors may be viewed as higher priority targets.

There are a number of conceptual and empirical issues that must be considered with regards to whether this potential re-conceptualisation of psychopathology is merited. One fundamental consideration is the quantitative extent of symptom-generality of comorbidity. If the extent of symptom-general covariation is meagre, the idea of placing importance on a *p*-factor is difficult to justify. There is no single definitive method of quantifying *p*-factor strength but a range of methodologies that can contribute to a general picture. Given the popularity of bi-factor models for modelling the *p*-factor, we here discuss indices based on this approach (e.g. Caspi et al., 2014; Murray, Eisner & Ribeaud, 2016; Patalay et al., 2015; Stochl et al., 2014; Tackett, Lahey, van Hulle, Waldman, Krueger & Rathouz, 2013). A bi-factor model is a measurement model in which each indicator loads on two factors: a general factor common to all indicators and a group factor common to a subset of indicators (Reise, 2012). Where psychopathology is concerned, the indicators measure specific symptoms or disorders, the general factor is the '*p*-factor' and the group factors are broad dimensions such as internalising, externalising and thought disorder.

A number of previous studies have reported that fitting this kind of structure using CFA yields good fit by conventional criteria (e.g. Caspi et al., 2014; Kim & Eaton, 2015; Laceulle, Vollebergh & Ormel, 2015; Noordhof et al., 2015; Patalay et al. 2015). However, simply fitting a model that includes a general factor does not provide a direct quantification of how important symptom-general covariance is either in absolute terms or relative to the variance shared among symptoms that belong to more specific trans-diagnostic factors such as internalising and externalising. For this purpose, certain indices can be computed from a bi-

factor model (or a Schmid-Leiman transformation of a second-order model; Schmid & Leiman, 1957) to provide more of a quantification of the importance of symptom-general variance.

First, omega hierarchical (ω_h) can be used to quantify the strength of a *p*-factor controlling for the group factors. ω_h is part of the ω family of model-based estimates of reliability. Different ω coefficients can be computed to estimate the extent to which latent dimensions contribute to the reliability of observed scores. Using the parameter estimates from a bi-factor model in which general and group factors are all orthogonal, ω_h is computed as:

$$\omega_h = \frac{(\sum \lambda_{iP})^2}{(\sum \lambda_{iP})^2 + (\sum \lambda_{iG1})^2 + (\sum \lambda_{iG2})^2 + \cdots (\sum \lambda_{iGK})^2 + \sum {\theta_i}^2}, \quad (1)$$

where λ_{iP} denotes the *p*-factor loading of item *i*; λ_{iG1} to λ_{iGK} denote the group factor loadings of item *i* for group factors 1 to K; and ${\theta_i}^2$ denotes item residual variance. It is an estimate of the proportion of summed score variance attributable to the *p*-factor. Noordhof et al. (2015) was to our knowledge, thus far the only *p*-factor study to report ω_h. They fit a bi-factor model to a selection of the subscales from the Dutch versions of the Child Behavior Checklist (CBCL; Verhulst, van der Ende and Koot, 1996) and the Child Social Behavior Questionnaire (SCBQ; Hartman, De Bildt, and Minderaa, 2013). The value of .75 is large and is comparable to the magnitudes of ω_h found in cognitive ability research where a general factor '*g*' has long been considered of major substantive importance (Revelle & Wilt, 2013; Spearman, 1904; Jensen, 1998). It also satisfies the psychometric rule of thumb that 50% of test variance should be due to the general factor of that test (Revelle & Wilt, 2012).

Another index that can be used to provide a quantification of *p*-factor importance is the explained common variance (ECV) statistic. ECV is the proportion of common variance that is attributable to a general factor. It is computed as:

$$ECV = \frac{\sum {\lambda_{iP}}^2}{\sum {\lambda_{iP}}^2 + \sum {\lambda_{iG1}}^2 + \sum {\lambda_{iG2}}^2 + \cdots \sum {\lambda_{iGK}}^2} \quad (2)$$

where λ_{iP} and λ_{iG1} to λ_{iGK} are as defined for eq.1 and eq.2. It is the proportion of total explained variance that is explained by the *p*-factor and thus provides a quantification of the importance of the *p*-factor relative to the group factors. No study to date has, to our knowledge, computed ECV for a *p*-factor model.

There are several different methods by which a bi-factor model can be estimated to provide the parameters to be entered in equations 1 and 2. In particular, a bi-factor model can be fit as a an exploratory factor analysis (EFA) model (including an exploratory structural equation model; ESEM; Asparouhov & Muthén, 2009) a CFA model (e.g. Jennrich & Bentler, 2011) or a Bayesian structural equation model (BSEM; Asparouhov & Muthén. 2012). No study has yet compared these three approaches to estimating the strength of a general factor such as the *p*-factor; however, it is quite possible that they would perform differently. In brief, the most commonly used approach - CFA with maximum likelihood estimation - may be liable to overestimate *p*-factor loadings because in a CFA model many loadings are conventionally fixed to zero, reflecting the hypothesis that items reflect only certain factors (the *p*-factor plus one group trans-diagnostic factor in the current context). However, the assumption that all remaining loadings are zero is, in general, unrealistic. Many small and substantively meaningful cross-loadings arise in practice (Asparouhov & Muthén,

2009). In psychopathology, symptoms are rarely pure 'indicators' of a single trans-diagnostic factor, but can reflect multiple factors (e.g. Eaton et al., 2011; Keyes et al., 2013; Oleski, Cox, Clara & Hills, 2011). In addition, many small cross-loadings will arise simply because of the practical impossibility of designing completely 'pure' indicators of one construct (e.g. Morin, Arens, and Marsh, 2015).

In principle, traditional CFA approaches can handle cross-loadings. Cross-loadings can be specified *a priori* based on past research and/or theory or modification indices and expected parameter changes (EPCs) can be used to identify local mis-specifications of zero (cross-) loadings and these can be iteratively added to the model (e.g. Saris, Satorra & Van der Veld, 2009). In practice, however, theory provides a poor guide as to cross-loadings and the stepwise use of MIs and EPCs can lead to the incorrect model. Moreover, the inclusion of cross-loadings still tends to be limited to including only 'salient' cross-loadings exceeding a conventional threshold such as |.3|. As such, in practice a large number of small cross-loadings are likely to remain unmodeled.

Fixing the majority of cross-loadings resulting from the non-isomorphic nature of psychopathology symptoms to zero - as is custom in CFA models - forces the covariance to be mediated via an alternative pathway, potentially inflating first-order factor inter-correlations in an oblique model or *p*-factor loadings in a bi-factor or higher-order model (Asparouhov & Muthén, 2009; Murray & Johnson, 2013; Morin et al., 2015). Issues like these mean that even if a bi-factor model is not the 'true' model, it can still fit well (see e.g. Murray & Johnson, 2013)

Muthén and Asparouhov (2012) propose a solution to the problem of cross-loadings in CFA. They recommend using BSEM specifying cross-loadings to be approximately zero, with small variance priors. BSEM differs from traditional applications of CFA estimated with

maximum likelihood (ML) estimation in that parameter estimates are derived from a posterior distribution formed from the combination of a prior distribution and likelihood. Bayesian estimates are the mean, median or mode of the posterior distribution and will be close to ML estimates when the prior distribution is non-informative. In BSEM, cross-loading can be set to be approximately zero by placing a prior distribution on them that is centred on zero and with small variance. This allows cross-loadings to be non-zero while keeping the basic CFA model intact. In traditional applications of CFA, freeing all cross-loadings would lead to a non-identified model; however, in BSEM the priors provide this identification. They are chosen to reflect prior beliefs based on past research or theory. Assuming all indicators are standardised prior to analysis, a reasonable choice of prior for a cross-loading is a normal distribution with mean zero and variance .01. Here, 95% of the loading variance will be in the +/- .2 range. As the variance of the prior is made larger, larger cross-loadings are accommodated but identification suffers.

In an ESEM or EFA model, all loadings are freely estimated; therefore, in principle, the covariance due to cross-loadings is appropriately modelled and does not inflate other parameters, especially those used to gauge the strength of a *p*-factor. However, in practice, Mansolf & Reise, (2016) noted that EFA models with bifactor rotations are liable to over-attribute variance to the general factor because group factors can 'collapse' onto the general factor.

Although BSEM and EFA have been available for a number of years, previous *p*-factor studies have essentially all used CFA with ML estimation. In this study, we therefore compared these three methods of estimating *p*-factor strength using a combination of simulation studies and real data examples.

Simulation Study

Population models

Three population models were considered in which we varied the strength of the general factor. Across all conditions, latent factor variances and total item variances were kept constant at 1 and residual item variances were kept constant at .30. In this way, the summed score variance was also kept constant across population models. We examined three levels of general factor strength, such that population models had either a very weak general factor (general factor loadings of .10; specific factor loadings of .83), a moderate strength general factor (general factor loadings of .46; specific factor loadings of .70), or a strong general factor (general factor loadings of .70; specific factor loadings of .46). Across these population models, the number of items was kept constant at 20. The number of group factors was kept constant at four, with five items per factor. We also included 3 cross-loadings of magnitude 0.25 with the primary loading of cross-loading items correspondingly reduced to maintain the same item total variance. Group factors were orthogonal to one another and to the general factor. The models are summarised in Figures 1-3. Population $\omega_h$and ECV values are provided in Results tables. For each of the three population models, data on N=1000 and N=200 was generated over 1000 replications. All analyses were conducted in *Mplus 7.13* (Muthén & Muthén, 2012). *Mplus* scripts are available from the first author. On the basis of initial results, additional simulation conditions were added *ad hoc* in order to further probe potentially important results identified based on the initial conditions.

Fitted models

For each population model, three approaches to calculating ω_h and ECV were applied to the N=1000 and N=200 replicates: a confirmatory factor analysis bi-factor model ('CFA'), an exploratory structural equation bi-factor model ('ESEM/EFA') and a Bayesian structural equation bi-factor model ('BSEM'). These are described below in more detail.

In the *CFA* conditions, a bi-factor confirmatory factor analysis model with one general factor and four specific factors were fit to each dataset. General and specific factors were all set orthogonal to one another and scaling and identification achieved by fixing the latent variable variances to 1. Models were estimated using ML estimation. We included one set of conditions in which cross-loadings were freely estimated and one set where they were fixed to zero. The latter set represents mis-specified models in the sense that parameters that are present in the population models are not present in the fitted models.

In the *BSEM* conditions, bi-factor models with one general factor and four specific factors were fit to the data with scaling and identification achieved by fixing latent variances to 1. Models were estimated using Bayesian estimation. Analogous to the CFA models, two sets of models were fit: one including only the primary loadings (i.e. mis-specified models) and one including small variance priors [~N(0,0.01)] on all potential secondary loadings. We also included the software default priors, specifically inverse gamma prior distributions with alpha= -1 and beta= 0 for observed variable residual variances. This encodes the assumption that their values are positive

In the *ESEM/EFA* conditions, an ESEM/EFA model using a bifactor rotation was used. The technical details of ESEM/EFA are comprehensively described in Asparouhov & Muthen (2009) and the technical details of bifactor rotations are provided in Jennrich & Bentler (2011). In brief, 'ESEM' describes an exploratory factor analysis measurement model within a structural equation model although the term is often used even when only the measurement model is estimated. Here we estimated an ESEM/EFA measurement model with an orthogonal bifactor rotation (bi-geomin) in which we specified four group factors and one general factor. The bi-geomin rotation is recommended in cases where cross-loadings are likely to be present (e.g. Mansolf & Reise, 2016). Scaling and identification were achieved by fixing the latent factor variances to 1. As ESEM/EFA by definition allows all items to load on

all factors, there were no 'mis-specified' models fit to the datasets. Models were estimated using ML estimation.

In all cases, ω_h and ECV were computed as shown in equations 1 and 2 based on the estimated solutions.

Simulation Outcomes

We considered three simulation outcomes: the percentage of convergence failures, bias in ω_h, and bias in ECV. In ML estimation, convergence is defined by a vanishingly small difference between estimates from successive iterations. For Bayesian estimation, convergence is defined by similarity across chains (each formed by successive draws) as indexed by a comparison of within and between chain variance. Bias in ω_h and ECV were computed in terms of per cent bias:

$$\frac{est - pop}{pop} \times 100 \tag{3}$$

where *est* refers to the average ω_h or ECV parameter estimate over the 1000 replications and *pop* refers to the corresponding population parameter.

Real Data Examples

We include two real data examples to illustrate the different approaches to estimating *p*-factor strength in empirical psychopathology data. We provide one real data analysis utilising a dataset in which there was evidence for a strong general factor ('counselling CORE-OM') and one utilising a dataset in which no true general factor could be extracted ('z-proso SBQ').

Counselling CORE-OM

Participants and Measures.

Participants contributing data for the first real data example were n=359 users of University counselling services at a large UK higher education institution. The dataset has been described in several existing publications (Murray et al., 2016a; Murray et al., 2016b; McKenzie et al., 2016). In brief, participants (108 male, 249 female, 1 transgender) with a mean age of 22.7 (SD=4.3) were administered the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) before receiving a counselling intervention. The CORE-OM is supported by psychometric evaluations across a large number of previous studies (e.g. Barkham, Mellor-Clark, Connell & Cahill, 2006; Connell et al., 2007; Murray, McKenzie, Murray, Richelieu, 2014). It is a 34 item self-report instrument. Items refer to internalising symptoms such as loneliness, panic, feeling unhappy as well as externalising symptoms such as threatening or intimidating others, taking dangerous risks with health. They also refer to somatic symptoms, insomnia, suicidal ideation and plans, intrusive thoughts and social support. Participants rated the extent to which they have experienced symptoms on a 5-point Likert scale from *Not at all* to *Most or all the time*.

Statistical Procedure

The basic factor structure for the CORE-OM real data analyses was adopted from previous research (Murray, McKenzie & Richelieu, 2018). Exploratory factor analyses in the previous study indicated that an optimal factor structure for this set of items was one in which all items loaded on a general factor as well as subsets of items loading on one of three specific factors. The specific factors were labelled 'externalising', 'internalising' and 'self-harm' based on the contents of the highest loading items in each case. Using this basic structure, the 3 previously described approached to estimating ω_h and ECV were applied: CFA, BSEM and ESEM/EFA. All indicators were standardised prior to analysis.

z-proso SBQ

Participants and Measures

Data for the second real data examples comes from the Zurich project on social development from Childhood to Adulthood (z-proso): a longitudinal cohort study based in Zurich, Switzerland focussed on positive youth development. A full description of the study, including recruitment and assessment procedures can be found in various prior publications (Eisner & Ribeaud, 2007; Ribeaud & Eisner, 2010) and on the study website (http://www.jacobscenter.uzh.ch/en/research/zproso/aboutus.html). The current study focusses on the 6th main data collection wave when the participants were aged 15-16 (median = 15.68). At this stage, data on the constructs relevant for the current study were available on between 1271 and 1286 participants, depending on the specific item. Analyses were based on 17 items of the Social Behavior Questionnaire (SBQ; Tremblay et al., 1991). These items provided measures of internalising (anxiety, depression), externalising (reactive aggression, relational aggression, proactive aggression, physical aggression) and attention-deficit hyperactivity disorder (attention deficit, hyperactivity/impulsivity). All items were administered in German. Individuals were asked to respond with respect to their feelings or behaviour in the last month in the case of anxiety and depression and in the last year in the case of externalising and ADHD symptoms. Responses were on a five-point scale from *Never* to *Very Often.*

Statistical Procedure

In a first step, the appropriate number of factors to include in the main analyses was determined using EFA. The number of group factors (K) to retain was guided by parallel analysis with principal components analysis (PA-PCA), the minimum average partial (MAP) test and visual inspection of a scree plot. PA-PCA was used rather than PA-PAF (parallel

analysis with principal axis factoring) because although the latter is theoretically aligned with EFA, it has a greater tendency to over-extract than PA-PCA (e.g. Crawford et al, 2010). We evaluated factor solutions with a range of numbers of factors centred on the consensus from the factor retention criteria to check for evidence of over- or under- extraction of group factors. Factor solutions were estimated using minimum residuals (minres) estimation and oblimin rotation. The factors were interpreted based on the contents of high-loading indicators. These preliminary analyses were used to guide model specification in the main analyses with items with loadings >|.3| in the preliminary analyses were used to define the K group factors in the main analyses. All items, whether or not they loaded >|.3| on the *p*-factor in the preliminary EFA analyses, were used to define the *p*-factor in the main analyses.

Results

Simulation Study

In the *CFA* condition, estimation failures occurred 18-19% of the time when a bi-factor model was fit to a set of items with a very weak general factor and n=1000. They occurred at an even higher rate with n=200 (up to 42% when the model was mis-specified). In these very weak general factor conditions, even among the replications that converged, there were a large number of solutions in which the residual covariance matrix was non-positive definite. Convergence problems with bifactor and similar psychometric models using ML estimation have previously been noted, especially at smaller sample sizes (e.g. Maydeu-Olivares & Coffman, 2006; Helm, Castro-Schilo & Oravecz, 2017). They may be more likely occur in the conditions in which the general factor is low in strength and where the sample size is small because factor loading estimates are here liable to be close to zero in samples. Indeed, estimation failures did not tend to occur when the general factor was moderate or strong even when the model was mis-specified, irrespective of sample size.

Bias in ω_h was substantial when the general factor was very weak and cross-loadings were present in the population but not estimated model. Here for n=1000, the average estimate was .25 (.20 for n=200) where the population value was only .05. Bias in ECV was also most pronounced in this condition (average estimate of .20 for n=1000 and .16 for n=200 compared with a population value of .01). ECV % bias was substantial across all conditions with a very weak general factor, even where the model was correctly specified although the difference in absolute values were generally modest and would be unlikely to lead to major distortions of substantive conclusions. Examining the patterns of estimated factor loadings suggested that the overestimation of ω_h and ECV was due both to an overestimation of general factor loadings and an underestimation of specific factor loadings. Having unmodeled cross-loadings led to a mis-attribution not only of unmodeled variance to the general factor, but also to a fundamental shift in the content of factors so that further specific factor variance was also attributed to the general factor. For example, the average *p*-factor loading for item 14 was .22 (compared with population value of .10) in the n=200 model while its average specific factor loading was .73 (compared with population value of .83).

Estimation failures occurred in the *BSEM* bi-factor models, in which the general factor was very weak and cross-loadings were present in the population model at n=1000 (12% failure rate when the cross-loadings were modelled; 21.8% when they were not) but were otherwise rare. The better convergence rates in BSEM than in CFAs with ML in some conditions was likely due to the additional information provided by the priors (those on the residual variances in all models and on the secondary loadings specifically in the condition in which cross-loadings were modelled; e.g. Helm et al., 2017). ω_h was substantially overestimated when the general factor was very weak and cross-loadings were present in the population model, especially when cross-loadings were not modelled (where ω_h was estimated at .23 for n=1000 and .20 for n=200). ECV was substantially overestimated in all

three conditions on which the general factor was very weak with the effect again being most marked when cross-loadings were present in the population model but not estimated in the fitted model (where ECV was .17 for both n=1000 and n=200). Examining the average factor loading estimates across replications suggested that these biases were due to a combination of overestimated general factor loadings and underestimated specific factor loading, with loading biases showing a similar pattern to those in the corresponding CFA conditions.

ESEM/EFA

Estimation failures occurred at a relatively constant rate of 17-18% across all conditions at n=1000 and of 22-26% at n=200. This was in contrast to BSEM and CFA with ML, both of which were considerably more likely to fail when the population model was characterised by low general factor loadings and/or the model was mis-specified. Both ω_h and ECV were substantially overestimated in the conditions in which the general factor was very weak, but there was some overstatement of general factor variance across all conditions. ω_h was estimated at .23 and .26 for the n=1000 conditions and at .23 and .27 for the n=200 conditions (compared with .05 population value), while the corresponding ECV estimates were .22 and .25 at both sample sizes (compared with .01 population value). A similar pattern of overestimated general factor loadings and underestimated specific factor loadings was also seen to be responsible for the ω_h and ECV overestimates; however, while the BSEM and CFA models generally only erred substantially when mis-specified, none of the ESEM/EFA models were technically incorrectly specified.

Additional conditions

Given the above results, we added supplementary conditions to further explore some of the observations from the initial set of simulations. First, given the poor performance of the ESEM/EFA models we increased the random starts for the rotation algorithm, from the

software default of 30 to 1000. Past research has suggested that bi-factor rotations in ESEM/EFA are prone to local minima and that within these solutions, general factor variance is liable to be overstated (Mansolf & Reise, 2016). We used a sample size of n=200.

Second, given that CFA with ML and BSEM did not evidence substantial bias when the general factor was moderate or strong provided the number of cross-loadings were limited, we also explored some conditions in which population models presented greater factorial complexity , in order to identify the point at which their performance is likely to break down. To do this, we relocated some of the variance in primary loadings to secondary loadings. Specifically, an additional 12 cross-loadings of .10 were added, adjusting primary factor loading parameters downwards to maintain the same population item total and residual variances. In order to evaluate whether ESEM/EFA might outperform CFA and BSEM in conditions with more complex structures, we also evaluated its performance with these more complex underlying population structures. The population models are summarised in Supplementary Materials. Our model fitting strategies were here designed to mimic common or recommended strategies in practice. For the CFA models we followed the standard recommendation of including standardised loadings <|.3| and thus did not include the .10 nor the .25 cross-loadings in the fitted models. For the BSEM models, we followed the recommendation of Muthén & Asparouhov (2012) and included small variance priors on all secondary loadings. For the ESEM/EFA models, all secondary loadings were freely estimated.

Results for the above-described additional conditions are provided in Supplementary Materials. Increasing the number of random starts to 1000 (Mansolf & Reise, 2016) in the rotation algorithm improved neither convergence rates nor bias in the ESEM/EFA models (see Table S1). This suggests the problems with ESEM/EFA are broader than local minima. The convergence failures are not necessarily surprising given the complexity- in terms of

number of free parameters - of the ESEM/EFA models (the BSEM models also contained large numbers of freely estimated parameters but convergence was assisted by the small variance priors on the secondary loadings). A likely explanation for the bias in factor loadings seem to be the shifts of group factor variance to the general factor outlined in Mansolf & Reise (2016), not only in local minima solutions but in the solution at the global minimum as well.

Results of fitting CFA, BSEM and ESEM/EFA models to more complex factorial structures are provided in Table S2. As expected, overestimation in ECV and omega hierarchical estimates increased for both CFA with ML and BSEM. Bias also increased in ESEM/EFA, and was similar to that observed in the CFA with ML and BSEM conditions, suggesting that it was no better able to handle more complex factorial structures.

Real Data Examples

ω_h and ECV values computed from the factor solutions of each method for the two datasets are provided in Table 4. For the counselling CORE-OM data, ω_h values were highly similar across the 3 methods, ranging from .90 (ESEM/EFA) to .92 (BSEM). ECV ranged from .70 (ESEM/EFA) up to .76 (CFA with ML). For the z-proso SBQ data, ω_h ranged from .16 (ESEM/EFA) up to .34 (CFA with ML) while ECV ranged from .23 (BSEM) up to .28 (ESEM/EFA).

Discussion

The extent to which symptom-general co-morbidity is a dominant feature of psychopathological symptoms has potential implications for the research, assessment and treatment of psychiatric disorders. However, to date there have been no studies comparing different method of estimating *p*-factor importance. We thus conducted a simulation study complemented by two real data examples to compare estimates of $\omega_h$and ECV derived from

CFA models estimated with ML, CFA models estimated with Bayesian estimation and ESEM/EFA models with a bifactor rotations. All three methods overestimated *p*-factor strength when the *p*-factor was weak. Overall, CFA performed well provided it was correctly specified (including major secondary loadings in the model). BSEM is likely to be useful when there is limited a priori knowledge of these secondary loadings. ESEM/EFA did not offer an advantage over these two methods despite freely estimating all loadings. In all cases, as would be expected, the overestimation of *p*-factor strength depended on the extent of unmodeled factorial complexity (i.e. secondary loadings).

As independent cluster structure would not generally be expected in psychopathology data (e.g. Cote et al., 2016), we specified cross-loadings in the population in all of our simulation conditions. BSEM and CFA showed some robustness to the effects of omission of cross-loadings when *p*-factor strength was at least moderate. When *p*-factor strength was weak and there were cross-loadings present in the population that were not modelled (i.e. the model was mis-specified) BSEM and CFA with ML overestimated ECV and ω_h. ESEM/EFA estimates all loadings, therefore, there was no condition in which the ESEM/EFA model was mis-specified in this way. Thus, it is notable that ESEM/EFA models performed worse than mis-specified BSEM and CFA models in many cases. Finally, it is instructive to compare a mis-specified CFA model to a BSEM model with small variance priors on all secondary loadings because mis-specified CFA models that omit cross-loadings may be common in the literature, and BSEM models with small variance priors on secondary loadings have been recommended as a solution (Muthén & Asparouhov, 2012). Here BSEM performed better, overestimating $\omega_h$and ECV to a lesser extent than the mis-specified CFA model.

ESEM/EFA models were arguably the poorest performing method while the currently dominant method of fitting CFA models with ML estimation performed well provided they were correctly specified and the general factor was of at least moderate

strength. The ESEM/EFA models had a tendency to shift specific factor variance onto the general factor. Mansolf & Reise (2016) provide a comprehensive account of how this can occur. In brief, available bifactor rotation criteria are minimised on the basis of the rotation of the group factors alone; however, in order to achieve permissible solutions, variance may shift between group and general factors. These shifts can lead to an overestimation of the general factor. Mansolf & Reise (2016), illustrated that this can often arise in local minima solutions; however, in the current study, general factor overestimation was not remedied by including a large number of random starts to account for local minima. This suggests that that global minima solutions may also be affected.

Overall, we would recommend estimating a BSEM model with small variance priors on secondary loadings in order to identify cross-loadings that need to be specified and estimating ω_h and ECV from either this model or a CFA model that includes the identified cross-loadings. EFA/ESEM may be useful in an exploratory phase but would not necessarily be the ideal model from which to compute ω_h and ECV.

We also provide two real data examples across which the three approaches to estimating *p*-factor strength can be compared. In our first real data example, there was evidence for a relatively strong *p*-factor. In fact, many items loaded only on the *p*-factor and not on any specific factor. In this dataset, there was minimal difference across methods in estimates of ω_h, which ranged from .90 to .92. This is in line with our simulation study that suggested that all three methods are reasonably robust in cases where the *p*-factor is strong. For ECV, the range of estimated was from .70 (ESEM/EFA) to .76 (CFA) with BSEM yielding an estimate of .73. This difference reflects the fact that to calculate ECV, loadings are squared before summing. In both BSEM and ESEM/EFA, there were a large number of small negative specific factor loadings, which would have contributed to the denominator of ECV.

In the second real data example, there was no evidence for a true general factor in the sense that any general factor extracted was defined by only a limited subset of the items in the set. Arguably, this renders the ω_h and ECV values meaningless as estimates of *p*-factor strength. These results provide a cautionary note against giving importance to the *p*-factor on the basis of good fitting bi-factor models alone. Some previous studies have cited good fit of a bi-factor model as evidence for a *p*-factor, however, all of our bi-factor models fit reasonably by conventional fit criteria in these data but none showed evidence of a truly general factor. In fact, in this dataset we observed that *at most* 53% of items loaded saliently on the *p*-factor. Thus, another recommendation from the current study is to ensure that patterns of loadings are examined in order to evaluate whether there is evidence for a latent factor that is general to all items, not just a subset.

On balance, past studies appear to be more in line with our counselling than z-proso real data example. Most studies find that only a minority – one or two – items, if any show non-salient *p*-factor loadings (e.g. Caspi et al., 2014; Lacuelle et al., 2015; Lahey et al., 2015; Patalay et al., 2015; Stochl et al., 2015; Tackett et al., 2013). However, the question of the importance of symptom-general covariance requires further study. Routinely providing ECV and ω_h as indices of *p*-factor strength will help towards this goal.

Finally, *p*-factor strength is only one issue relevant for *p*-factor interpretation. There are also broader issues related to the interpretation of the common variance that is captured by the general factor in bi-factor and closely related model general factor models that remain to be resolved. At a basic level, researchers must decide whether a bi-factor model provides an appropriate model for their data; a task made difficult by its practical indistinguishability from other possible models on the basis of model fit (e.g. Murray & Johnson, 2013). It is well known, for example, that the appearance of a general factor can result from a range of other underlying causal structures that produce equivalent covariance structure (e.g. van der Maas

et al., 2006). Current psychopathological theory provides little strong justification for any particular interpretation of general factor variance over others, although it appears likely that different psychopathological symptoms seem to *both* share causes and create increases risks for one another, suggesting that the true causal structure underlying the so-called *'p*-factor' is mixed (e.g. van Lier et al., 2012; Wertz et al., 2015).

Finally, though the work presented here were framed in terms of general psychopathology, they will also apply to a range of different research areas where there are questions about the relative importance of general and specific dimensions and difficulties in obtaining measures that demonstrate independent cluster structure. This includes areas such as personality research (e.g. Booth & Hughes, 2014) or research into specific clinical phenotypes (e.g. Garner et al., 2017; Murray et al., 2015).

Limitations

The primary limitations of the current study relate to the limited number of conditions explored in the simulation study. We focussed on a small number of population models that we believe are broadly representative of the types of data observed in psychopathology research. Future studies will be required to study the effects of features such as non-normal indicators, different numbers of items, patterns of loadings, and residual covariances that could influence estimates of *p*-factor strength. Our coverage of possible methods of estimating *p*-factor strength was not exhaustive. Other methods have been suggested to estimate general factor strength such as the average general factor loading (e.g. Gignac, 2014); a comparison of Revelle's worst split half reliability β at different levels of aggregation (e.g. Revelle, 1979); utilising bifactor loadings from a Schmid-Leiman transformed higher-order model; or using target rotations in ESEM/EFA. Future simulation studies will be required to assess the utility of these indices for assessing *p*-factor strength.

Finally, the utility of simulation studies depends on the extent to which the studied models are useful representations of real world; however, this is difficult to directly verify and generally has to be assumed.

Conclusion

Provided that *p*-factor strength is moderate to high, ECV and ω_h can be used as estimates of *p*-factor strength. Lower values should be treated with caution as they are liable to overestimate *p*-factor strength when there are unmodeled secondary loadings. It is also important to check that ECV and ω_h reflect a genuine general factor, in the sense that most items should load saliently on this factor. Our results suggest that the best method of quantifying *p*-factor strength is either a BSEM bi-factor model with small variance priors on cross-loadings or a CFA model with major secondary loadings freely estimated. BSEM and CFA should be preferred to a bifactor rotation strength as the latter tends to overestimate *p*-factor strength even in the most favourable conditions.

Acknowledgements

We are grateful to the children, parents and teachers who provided data for the z-proso study and the research assistants involved in its collection. Funding from the Jacobs Foundation (Grant 2010-888) and the Swiss National Science Foundation (Grants 100013_116829 & 100014_132124) is also gratefully acknowledged.

Conflict of Interest

The authors have no conflicts of interest to declare.

Compliance with ethical standards

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry*, *40*, 57-87.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397-438.

Asparouhov, T. & Muthén, B., (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313-335.

Barkham, M., Mellor-Clark, J., Connell, J., & Cahill, J. (2006). A core approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System. *Counselling and Psychotherapy Research*, *6*, 3-15.

Booth, T., & Hughes, D. J. (2014). Exploratory structural equation modeling of personality data. *Assessment*, *21*, 260-271.

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

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*¸ 885–901.

Côté, S. M., Orri, M., Brendgen, M., Vitaro, F., Boivin, M., Japel, C., ... & Tremblay, R. E. (2017). Psychometric properties of the Mental Health and Social Inadaptation Assessment for Adolescents (MIA) in a population-based sample. *International Journal of Methods in Psychiatric Research*.

Connell, J., Barkham, M., Stiles, W. B., Twigg, E., Singleton, N., Evans, O., & Miles, J. N. (2007). Distribution of CORE–OM scores in a general population, clinical cut-off points and comparison with the CIS–R. *The British Journal of Psychiatry*, *190*, 69-74.

Eaton, N. R., Krueger, R. F., Keyes, K. M., Skodol, A. E., Markon, K. E., Grant, B. F., & Hasin, D. S. (2011). Borderline personality disorder co-morbidity: relationship to the internalizing–externalizing structure of common mental disorders. *Psychological Medicine*, *41*, 1041-1050.

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*, 171-182.

Eisner, M., & Ribeaud, D. (2007). Conducting a Criminological Survey in a Culturally Diverse Context Lessons from the Zurich Project on the Social Development of Children. *European Journal of Criminology*, *4*, 271-298.

Garner, A. A., Peugh, J., Becker, S. P., Kingery, K. M., Tamm, L., Vaughn, A. J., ... & Epstein, J. N. (2017). Does sluggish cognitive tempo fit within a bi-factor model of ADHD?. *Journal of Attention Disorders*, *21*, 642-654.

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence*, *42*, 89-97.

Hartman, C. A., De Bildt, A., & Minderaa, R. B. (2013). CSBQ –children's social behavior questionnaire. In Volkmar, F. R. (Ed.), *Encyclopedia of Autism Spectrum Disorders* (pp. 825-829). New York: Springer.

Helm, J. L., Castro-Schilo, L., & Oravecz, Z. (2017). Bayesian Versus Maximum Likelihood Estimation of Multitrait–Multimethod Confirmatory Factor Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 17-30.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*, 537-549.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kendler, K. S., Prescott, C. A., Myers, J., & Neale, M. C. (2003). The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. *Archives of General Psychiatry*, *60*, 929-937.

Kessler, R. C., Avenevoli, S., McLaughlin, K. A., Green, J. G., Lakoma, M. D., Petukhova, M., ... & Merikangas, K. R. (2012). Lifetime co-morbidity of DSM-IV disorders in the US national comorbidity survey replication adolescent supplement (NCS-A). *Psychological Medicine*, *42*, 1997-2010.

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.

Kramer, M. D., Krueger, R. F., & Hicks, B. M. (2008). The role of internalizing and externalizing liability factors in accounting for gender differences in the prevalence of common psychopathological syndromes. *Psychological Medicine*, *38*, 51-61.

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., 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*, 850-860.

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*, 415-422.

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research*, *51*, 698-717.

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

McKenzie, K., Murray, K. R., Murray, A. L., & Richelieu, M. (2015). The effectiveness of university counselling for students with academic issues. *Counselling and Psychotherapy Research*, *15*, 284-288.

Morin, A. J., Arens, A. K., & Marsh, H. W. (2015). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, *23,* 116-139.

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The development of the general factor of psychopathology 'p-factor' through childhood and adolescence. *Journal of Abnormal Child Psychology*, *44*, 1573-1586.

Murray, A. L., McKenzie, K., Murray, K. R., & Richelieu, M. (2014). Mokken scales for testing both pre-and postintervention: An analysis of the Clinical Outcomes in Routine Evaluation—Outcome Measure (CORE–OM) before and after counseling. *Psychological Assessment*, *26*, 1196.

Murray, A. L., McKenzie, K. & Richelieu (2018). Treatment-related response shifts in the CORE-OM. Manuscript submitted for publication.

Murray, A. L., McKenzie, K., Kuenssberg, R., & Booth, T. (2015). Do the autism spectrum quotient (AQ) and autism spectrum quotient short form (AQ-S) primarily reflect general ASD traits or specific ASD traits? A bi-factor analysis. *Assessment*, *24*, 444-57.

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,* 407-422.

Muthén, L. K., & Muthén, B. O. (2012). Mplus Version 7 user's guide. *Los Angeles, CA: Muthén & Muthén*.

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.

Oleski, J., Cox, B. J., Clara, I., & Hills, A. (2011). Pathological gambling and the structure of common mental disorders. *The Journal of Nervous and Mental Disease*, *199*, 956-960.

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*, 15-122.

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

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*, 57-74.

Revelle, W., & Wilt, J. (2012). On when a factor is a general factor: Presented at the European Association for Personality Psychology experts meeting on the structure of personality, Dubrovnik, Croatia.

Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality*, *47*, 493-504.

Ribeaud, D., & Eisner, M. (2010). Risk factors for aggression in pre-adolescence: Risk domains, cumulative risk and gender differences-Results from a prospective longitudinal study in a multi-ethnic urban sample. *European Journal of Criminology*, *7*, 460-498.

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*, *97*, 425-435.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications?. *Structural Equation Modeling*, *16*, 561-582.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53-61.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*, 201-292.

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.

Tackett, J. L., Lahey, B. B., Van Hulle, C., Waldman, I., Krueger, R. F., & Rathouz, P. J. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *Journal of Abnormal Psychology*, *122*, 1142-1153.

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivee, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *Journal of Abnormal Child Psychology*, *19*, 285-300.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842-861.

Van Lier, P. A., Vitaro, F., Barker, E. D., Brendgen, M., Tremblay, R. E., & Boivin, M. (2012). Peer victimization, poor academic achievement, and the link between childhood externalizing and internalizing problems. *Child Development, 83*, 1775-1788.

Verhulst, F. C., van der Ende, J., & Koot, H. (1996). Handleiding voor de CBCL/4-18. [Manual for the CBCL/4-18]. Rotterdam: Afdeling Kinder- en jeugdpsychiatrie, Sophia Kinderziekenhuis/Academisch Ziekenhuis Rotterdam/Erasmus.

Wertz, J., Zavos, H., Matthews, T., Harvey, K., Hunt, A., Pariante, C. M., & Arseneault, L. (2015). Why some children with externalising problems develop internalising symptoms: testing two pathways in a genetically sensitive cohort study. *Journal of Child Psychology and Psychiatry*, *56*, 738-746.

Wright, A. G., Krueger, R. F., Hobbs, M. J., Markon, K. E., Eaton, N. R., & Slade, T. (2013). The structure of psychopathology: toward an expanded quantitative empirical model. *Journal of Abnormal Psychology*, *122*, 281-294.

Table 1: Simulation results for CFA

Population model	Fitted model	Estimation failure rate (%)	Population ω_h	Average ω_h estimate	ω_h % Bias	Population ECV	Average ECV estimate	ECV % bias
				n=1000				
Very weak *p*-factor/cross-loadings	Bi-factor with cross-loadings	18.0	0.05	0.25	436.81	0.01	0.04	1278.95
Moderate *p*-factor/cross-loadings	Bi-factor with cross-loadings	0.0	0.59	0.64	8.02	0.30	0.30	10.36
Strong *p*-factor/cross-loadings	Bi-factor with cross-loadings	0.0	0.87	0.89	2.56	0.70	0.70	3.15
Very weak *p*-factor/cross-loadings	Bi-factor ICS	19.2	0.05	0.04	690.55	0.01	0.20	283.51
Moderate *p*-factor/cross-loadings	Bi-factor ICS	0.0	0.59	0.59	24.74	0.30	0.33	-0.64
Strong *p*-factor/cross-loadings	Bi-factor ICS	0.0	0.87	0.88	7.22	0.70	0.72	-1.12
				n=200				
Very weak *p*-factor/cross-loadings	Bi-factor with cross-loadings	27.2	0.05	0.07	36.94	0.01	0.07	395.97
Moderate *p*-factor/cross-loadings	Bi-factor with cross-loadings	0.8	0.59	0.58	-1.67	0.30	0.30	0.36
Strong *p*-factor/cross-loadings	Bi-factor with cross-loadings	0	0.87	0.87	-0.08	0.70	0.69	-0.73

Very weak *p*-factor/cross-loadings	Bi-factor ICS	41.7	0.05	0.20	313.56	0.01	0.16	1049.82
Moderate *p*-factor/cross-loadings	Bi-factor ICS	4	0.59	0.62	6.12	0.30	0.34	11.72
Strong *p*-factor/cross-loadings	Bi-factor ICS	0	0.87	0.89	2.41	0.70	0.72	2.59

Note. ICS=independent cluster structure.

Table 2: Simulation results for BSEM

Population model	Fitted model	Estimation failures (%)	Population ω_h	Average ω_h estimate	ω_h % Bias	Population ECV	Average ECV estimate	ECV % bias
			n=1000					
Very weak *p*-factor/cross-loadings	Bi-factor with small variance priors	12	0.05	0.13	165.88	0.01	0.09	525.21
Moderate *p*-factor/cross-loadings	Bi-factor with small variance priors	0.4	0.59	0.63	6.93	0.30	0.33	9.04
Strong *p*-factor/cross-loadings	Bi-factor with small variance priors	0.3	0.87	0.89	2.31	0.70	0.71	1.62
Very weak *p*-factor/cross-loadings	Bi-factor ICS	21.8	0.05	0.23	382.87	0.01	0.17	1084.75
Moderate *p*-factor/cross-loadings	Bi-factor ICS	0	0.59	0.64	8.31	0.30	0.34	11.25
Strong *p*-factor/cross-loadings	Bi-factor ICS	0	0.87	0.89	2.70	0.70	0.72	3.47
			n=200					
Very weak *p*-factor/cross-loadings	Bi-factor with small variance priors	0	0.05	0.01	165.45	0.01	0.11	687.27
Moderate *p*-factor/cross-loadings	Bi-factor with small variance priors	0	0.59	0.30	6.20	0.30	0.34	11.22
Strong *p*-factor/cross-loadings	Bi-factor with small variance priors	0	0.87	0.70	2.63	0.70	0.71	1.32

Very weak *p*-factor/cross-loadings	Bi-factor ICS	0.4	0.05	0.01	315.88	0.01	0.17	1105.71
Moderate *p*-factor/cross-loadings	Bi-factor ICS	0	0.59	0.30	7.73	0.30	0.35	15.83
Strong *p*-factor/cross-loadings	Bi-factor ICS	0	0.87	0.70	3.16	0.70	0.73	4.33

Note. ICS=independent cluster structure.

Table 3: Simulation results for ESEM/EFA

Population model	Fitted model	% Estimation failures	Population ω_h	Average ω_h estimate	% ω_h Bias	Population ECV	Average ECV estimate	% ECV bias
			n=1000					
Very weak *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	16.6	0.05	0.26	456.82	0.01	0.25	1644.29
Moderate *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	16.8	0.59	0.73	23.57	0.30	0.41	36.20
Strong *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	17.9	0.87	0.92	5.96	0.70	0.73	5.05
			n=200					
Very weak *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	22.2	0.05	0.27	463.14	0.01	0.23	1531.82
Moderate *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	22.3	0.59	0.71	20.97	0.30	0.40	33.51
Strong *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	24.1	0.87	0.92	5.69	0.70	0.72	2.90

Table 4: Comparison of ω_h and ECV across approaches in real data

Method	ω_h	ECV
Counselling CORE-OM		
CFA	.91	.76
BSEM	.92	.73
ESEM/EFA	.90	.70
z-proso SBQ		
CFA	.34	.26
BSEM	.32	.23
ESEM/EFA	.16	.28

Figure 1: Population model for 'Very weak *p*-factor/cross-loadings' conditions

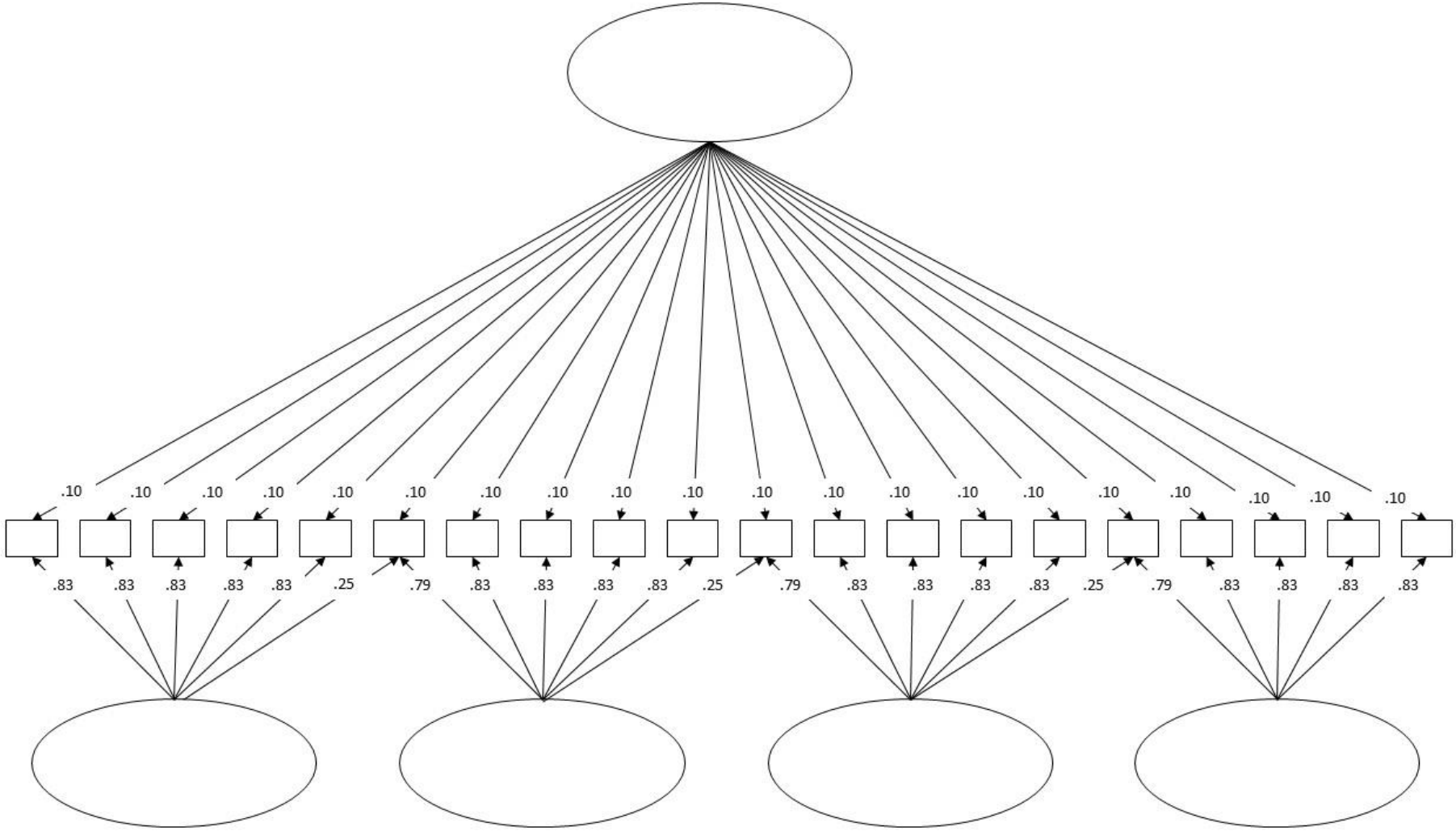

Figure 2: Population model for 'Moderate *p*-factor/cross-loadings' conditions

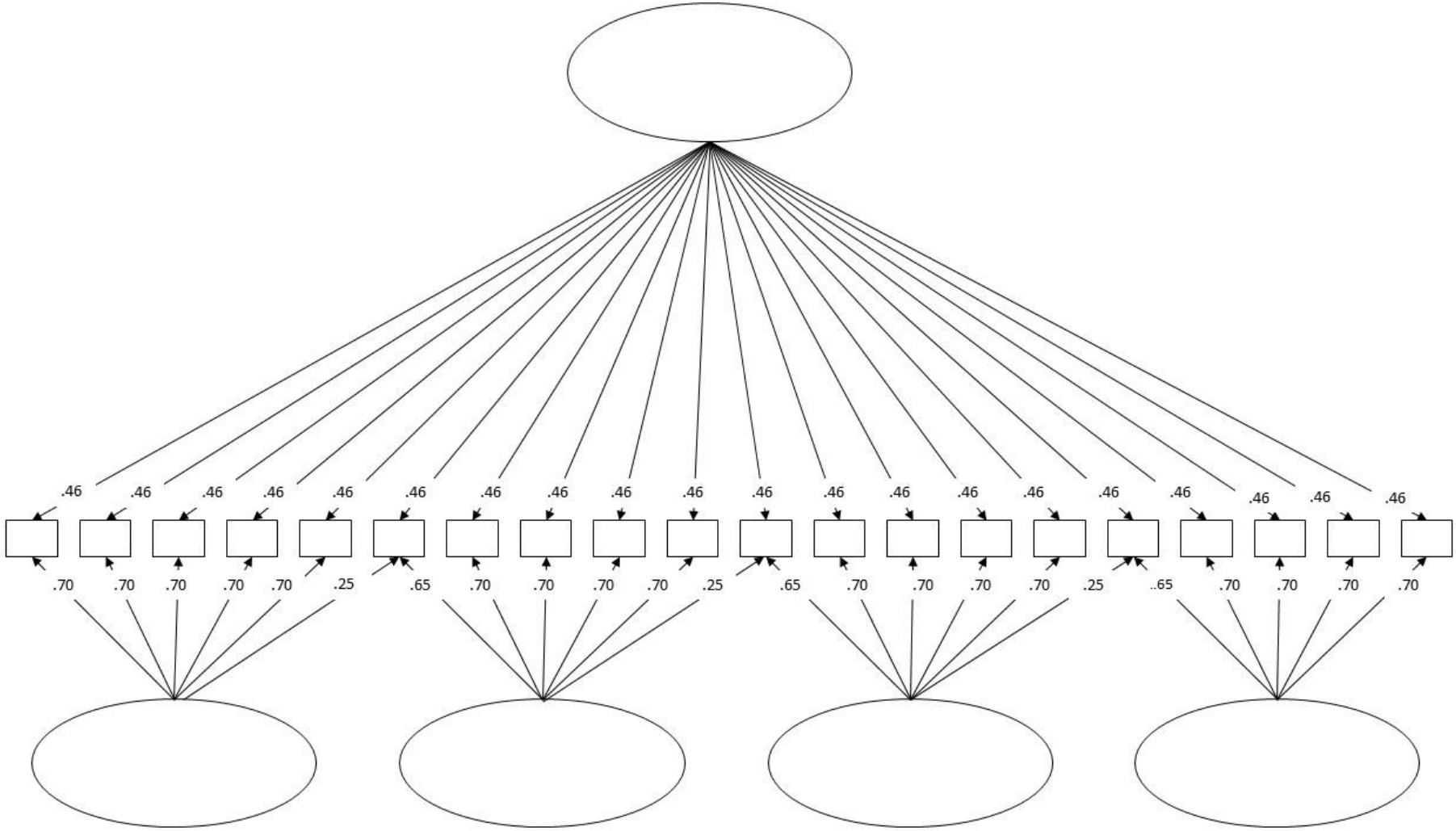

Figure 3: Population model for 'Strong *p*-factor/cross-loadings' condition

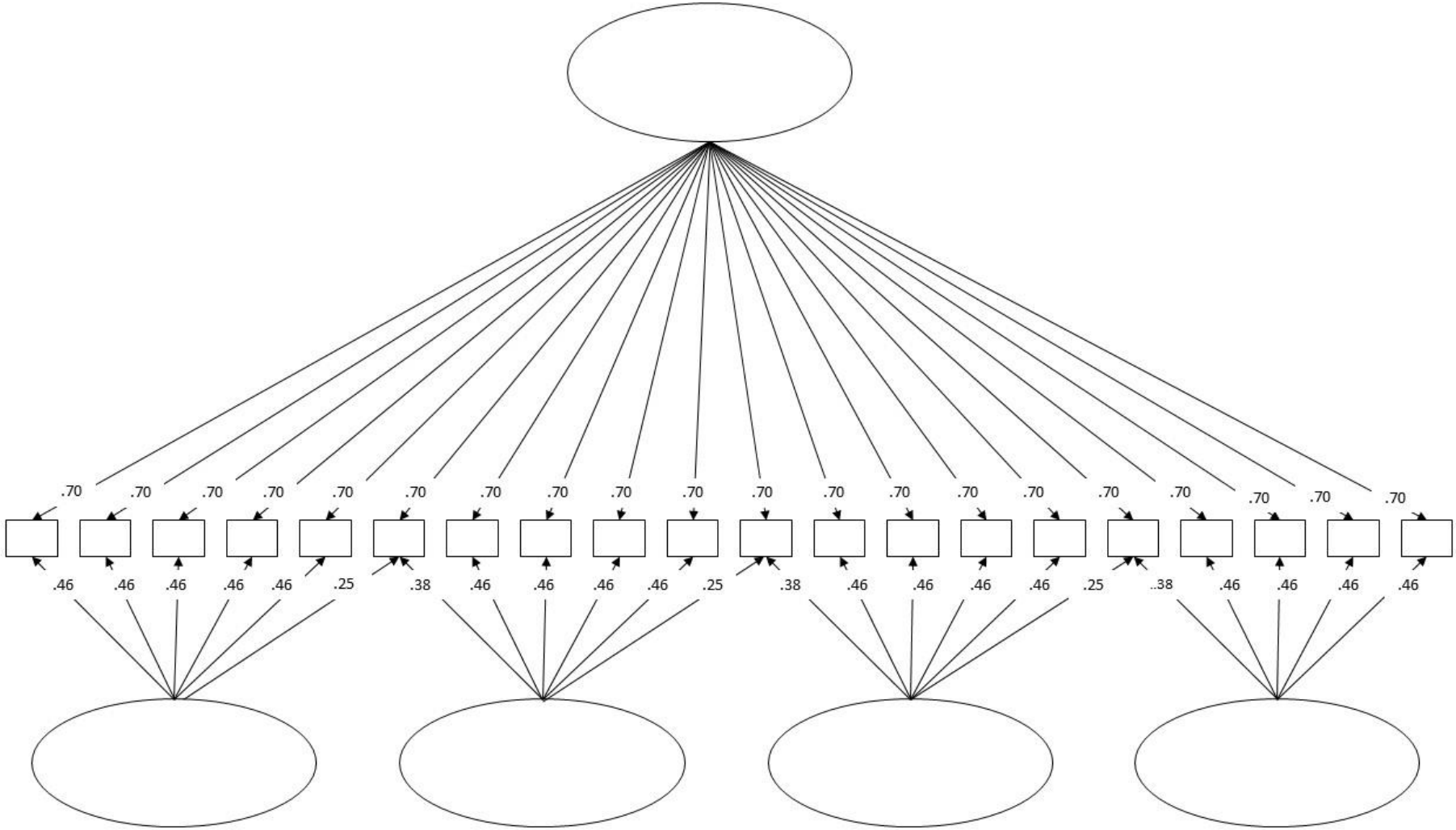

Table S1:

ESEM performance with 1000 random starts for rotation algorithm and n=200

Population model	Fitted model	% Estimation failures	Population ω_h	Average ω_h estimate	% ω_h Bias	Population ECV	Average ECV estimate	% ECV bias
Very weak *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	22.2	0.05	0.27	433.20	0.01	0.23	1564.29
Moderate *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	22.3	0.59	0.71	20.51	0.30	0.40	33.44
Strong *p*-factor/cross-loadings	Bi-factor all loadings freely estimated	24.1	0.87	0.92	5.53	0.70	0.72	3.07

Table S2:

CFA with ML, BSEM and ESEM performance with different numbers of cross-loadings

Population model	Fitted model	n	ω_h	ECV	Average ω_h estimate	Average ECV estimate	% bias ω_h	% bias ECV	% estimation failure
Moderate *p*-factor/cross-loadings	CFA/Bi-factor ICS	1000	.55	.30	.69	.39	24.74	27.60	0
Strong *p*-factor/cross-loadings	CFA/Bi-factor ICS	1000	.85	.70	.91	.76	7.22	9.02	0
Moderate *p*-factor/cross-loadings	BSEM/Small variance priors on cross-loadings	1000	.55	.30	.68	.38	23.65	25.41	100.00
Strong *p*-factor/cross-loadings	BSEM/Small variance priors on cross-loadings	1000	.85	.70	.90	.74	6.95	6.37	0.10
Moderate *p*-factor/cross-loadings	ESEM/EFA	1000	.55	.30	.74	.44	35.28	46.34	14.60
Strong *p*-factor/cross-loadings	ESEM/EFA	1000	.85	.70	.93	.76	9.83	9.07	15.20
Moderate *p*-factor/cross-loadings	CFA/Bi-factor ICS	200	.55	.30	.67	.38	21.91	27.23	6.10
Strong *p*-factor/cross-loadings	CFA/Bi-factor ICS	200	.85	.70	.91	.76	7.06	8.31	0.00
Moderate *p*-factor/cross-loadings	BSEM/Small variance priors on cross-loadings	200	.55	.30	.68	.39	23.80	28.59	0.00
Strong *p*-factor/cross-loadings	BSEM/Small variance priors on cross-loadings	200	.85	.70	.91	.75	7.32	6.40	0.00
Moderate *p*-factor/cross-loadings	ESEM/EFA	200	.55	.30	.68	.39	23.80	28.59	0.00
Strong *p*-factor/cross-loadings	ESEM/EFA	200	.85	.70	.91	.75	7.32	6.40	0.00