Hierarchically structured factor models: an investigation and extension of the bi-factor model

Nils Petras

Inaugural Disseration

Submitted in partial fulfillment of the requirements for the degree of Doctor of Social Sciences in the DFG Research Training Group "Statistical Modeling in Psychology" at the University of Mannheim

1st Supervisor:
Prof. Dr. Thorsten Meiser

2nd Supervisor:
Prof. Dr. Edgar Erdfelder

Dean of the School of Social Sciences:
Prof. Dr. Michael Diehl

Thesis Evaluators:
Prof. Dr. Edgar Erdfelder
Prof. Dr. Benjamin Hilbig

Thesis Defense:
July 12, 2024

For everyone who taught me

Contents

Summary

To represent complex psychological constructs such as multifaceted personality traits, general intelligence, or mental disorders, the bi-factor model is frequently used. Its main advantage over competing models is its clear and often insightful distinction between different parts of multidimensional constructs. It defines a general trait across all observed variables and specific traits representing the various facets of the construct.

The unique characteristics of the bi-factor model's structure come with several challenges that currently need more attention. In this thesis, I tackle three of these issues in three articles. The first article investigates the frequent occurrence of weak specific factors in bi-factor model applications. It explains why the characteristics of the bi-factor model in combination with typical measurement design in psychology should be expected to produce weak specific factors. The meta-analysis shows the pattern of problematic parameter estimates. Using simulations, the article analyses the statistical power and the parameter recovery under realistic conditions and provides guidelines for applied research. The second article investigates the flexibility of bi-factor model variants and their relationships to one another. Whereas previous research has noted the excessive flexibility of the bi-factor model compared to other models, the current work shows in simulations that its different variants can flexibly imitate each other. The most important consequence is that even some of the most basic claims derived from the model need to be questioned, because they may entirely depend on the choice between two equally well-fitting representations of the data. It is discussed that this issue cannot be resolved from a statistical perspective alone and a detailed account of the influence on parameter and trait estimates is provided. The third article proposes an alternative modeling approach for cases in which the underlying assumptions of a full, symmetrical bi-factor structure are violated. On a large example dataset, a set of replications and a multiverse analysis highlight the key strengths and limitations of this proposed approach.

The current work aims to expand the statistical bi-factor model toolbox and to guide the application and interpretation of previously suggested models. For this purpose, I combine statistical insights with a meta-scientific perspective on the bi-factor model's application. In this way, it became clear that an improved understanding of the discussed problems is key to their solution.

Articles

This cumulative thesis is based on three articles, one of which has been published and two have been submitted for publication.

Article I

Petras, N., & Meiser, T. (2024) Problems of domain factors with small factor loadings in bi-factor models, *Multivariate Behavioral Research*, *59*(1), 123-147. https://doi.org/10.1080/00273171.2023.2228757

Article II

Petras, N. (2024a). When factor variance and factor correlations are interchangeable: The relationship between the bi-factor model variants [manuscript submitted for publication]. Department of Psychology, University of Mannheim.

Article III

Petras, N. (2024b). Building hierarchically structured factor models with systematically selected residual correlations [manuscript submitted for publication]. Department of Psychology, University of Mannheim.

1 Psychological measurement: Measureing the unobservable

Measurement in psychology faces theories including constructs that cannot be directly observed, such as well-being, personality traits, or mental disorders. They include thoughts, emotions, and attitudes, which can only be (fully) accessed by the persons themselves. Therefore, these constructs cannot be measured by an objective measurement device alone. To make the unobservable measurable, researchers *operationalize* psychological constructs by selecting observable variables that are assumed to reflect variations on the unobservable target construct. Psychological theories can be researched empirically if they imply testable statistical hypotheses on these observed variables. Commonly, responses to self-report questionnaire items are used as observable variables, especially in research on interindividual differences.

Beyond being only indirectly observable, many psychological constructs are complex. They comprise several qualitatively different facets or abilities. Constructs measured by self- or other-rating, such as most personality traits, often include both attitudes and behaviors and comprise both emotion and cognition. Moreover, personality traits are complex by combining several content sub-dimensions. For example, the massively popular Big Five and HEXACO personality traits are commonly assumed to be meaningful dimensions in themselves and to comprise various, clearly distinguishable facets (Lee & Ashton, 2004; Paunonen & Ashton, 2001). Constructs measured by task performance on the other hand often comprise multiple tasks or task types. For example, general intelligence is understood as a single dimension underlying the performance on a large variety of different (types of) cognitive tasks on which performance systematically varies with more specific abilities, too (Carroll, 2003; e.g., Canivez et al., 2021). This complexity gave rise to two different conceptualizations. First, psychological constructs can be understood as a set of correlated dimensions. In this perspective, each facet of a personality trait, or each task type of a proficiency test, is one dimension of the construct. A construct's unifying characteristic is then the common definition that usually implies substantial correlations between these dimensions. Second, psychological constructs can be understood to be defined by a common core dimension that is measured by all observed variables. Beyond this unifying core dimension, each facet (task type, . . .) provides an additional source of variation that may either be an integral part of the definition of the construct or a consequence of the decisions made regarding the operationalization.

Because the relationships between the different parts of a construct and other constructs can differ in important ways (e.g., Gäde et al., 2017), the question of how to distinguish them is paramount to the testing of psychological theories. In this way, the development of statistical measurement models is parallel to the development of psychological theories and construct definitions. The current work addresses three challenges concerning the bi-factor model (Holzinger & Swineford, 1937; Reise, 2012), which is a key model for disentangling complex psychological constructs into a core dimension and several specific facets.

2 Measurement models of psychological constructs

From definitions of psychological constructs, statistical models can be derived that are hypothesized to account for the covariation of the observed variables.[1] For example, if people truly vary on an agreeableness trait, they should systematically vary in their reports of agreeable behaviors and attitudes. Moreover, if the agreeableness trait is a meaningful dimension that describes interindividual differences within a population, self-reports on multiple agreeableness questionnaire items should correlate positively. Therefore, a common empirical approach is to define and estimate a statistical model of the covariance matrix of the observed variables ($\mathbf{\Sigma}$), in the framework of structural equation modeling (SEM, see e.g., Kline, 2023). Two important frameworks with similar purposes have close similarities to this approach. First, item response theory (IRT, see e.g., Embretson & Reise, 2013; Reckase, 2009) also explains the statistical relationships between the observed variables by relating them to a set of latent variables via a design matrix (=loading matrix). Therefore, many structural features of SEM models can be translated into IRT models and vice versa. Second, network psychometrics (Epskamp et al., 2018) focuses more directly on the pattern of pairwise statistical relationships between the observed variables by modeling them as a network without specifying any latent variables. Therefore, network psychometrics implies a fundamentally different understanding of how psychological constructs are structured and is a major alternative framework to explain observed variable covariances.

When specifying a statistical model of a psychological measure, two potentially conflicting goals need to be considered. On the one hand, the variables in the model

[1] Although this is a guiding principle of empirical research, it is difficult to do in practice. The insufficient formalization of psychological theories and concepts is the subject of an old and ongoing debate (Eronen & Romeijn, 2020; McGuigan, 1953; Scheel, 2022).

should be meaningful. They should closely reflect the theoretical construct of interest. For this purpose, latent variables (also called latent traits) are specified to represent the unobservable construct(s) as common causes[2] of the observed responses in both SEM and IRT. The idea of this reflective measurement approach is that $\mathbf{\Sigma}$ can be accurately modeled by accounting for the influence of one or more constructs of interest. Confirmatory approaches, such as confirmatory factor analysis (CFA) are employed to test a priori assumptions about the relationships between observed and latent variables. CFA allows testing hypotheses in the form of model restrictions. On the other hand, the model should represent the empirical data accurately. If this is the case, all relevant influences in the data are accounted for and biases are limited. To achieve this, some level of exploratory data analysis is usually necessary.

The different perspectives on the structure of psychological constructs are manifested in the statistical models used to represent them (Figure 1). Whereas in the past, the group-factor model was favored in most applications, the bi-factor model (Holzinger & Swineford, 1937) has seen a surge in popularity recently (Figure 2, see also Reise, 2012; B. Zhang et al., 2021). The group-factor model distinguishes facets of a construct by assigning a factor to each of them. The notion of a unified construct is mostly absent in this model: it is only reflected in the latent correlations between the factors. The bi-factor model's popularity can be partially explained by resolving that. It distinguishes a general trait that represents the core construct from several facet-specific traits (Chen et al., 2012). The orthogonality of (all of) its factors further helps to separate the different parts of a construct, especially because their relationships with other variables are independent of one another. The bi-factor model's structure reflects a dual-perspective understanding of psychological constructs. They are a singular entity in the sense of having one meaningful underlying core dimension. At the same time, they are more than a single unobservable variable – they are multifaceted and therefore multidimensional. Conceptually, the orthogonality of the general and specific factors allows researchers to see the construct as the sum of its parts, or exclusively focus on its common core. It also allows researchers to separately account for some forms of systematic variance that are merely a consequence of the measurement design (e.g., testlets, Rijmen, 2010). The bi-factor structure is not specific to the literature of CFA measurement models: exploratory factor analysis (EFA, Jennrich & Bentler, 2011, 2012), and IRT (Cai et al., 2011) both use it as well.

[2]For a discussion of the meaning of causality and its direction in this context, see e.g., Bagozzi (2007)

Figure 1

Path diagrams of confirmatory factor models.

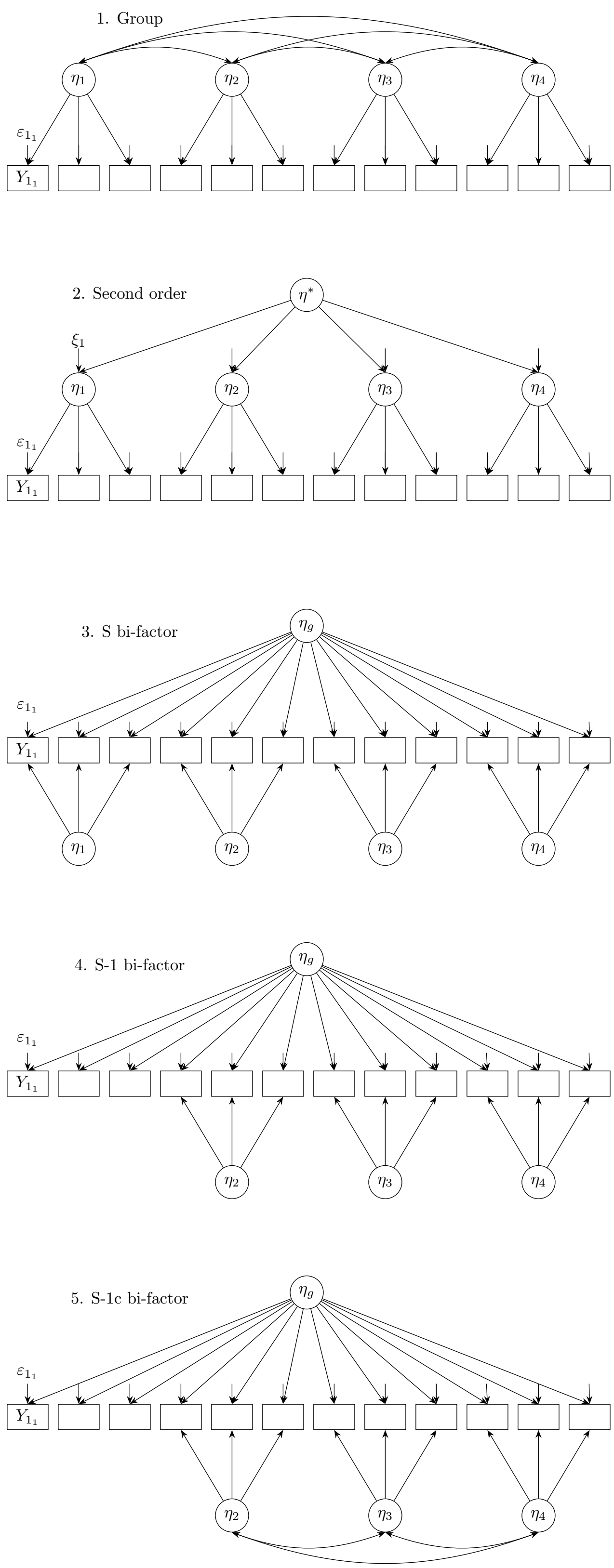
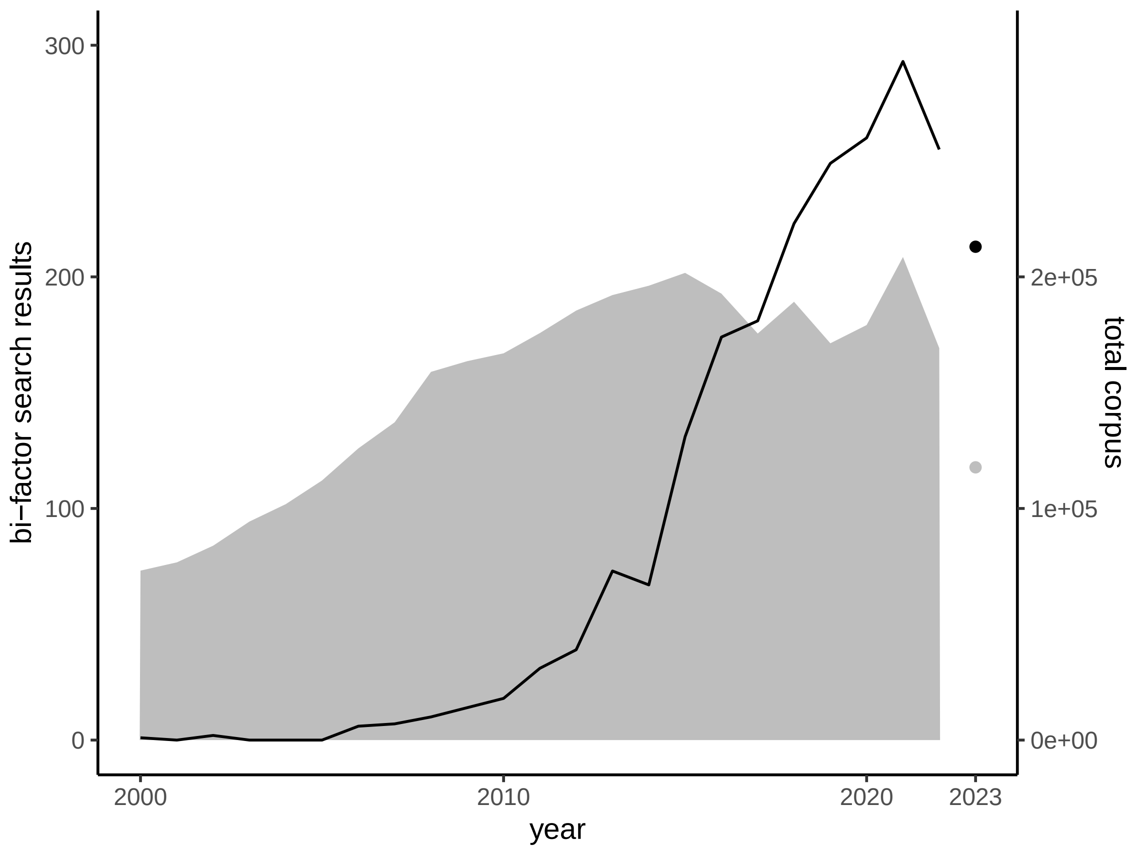


Figure 2

Articles in PsycINFO that mention "bifactor" or "bi-factor" in their titles or abstracts.



Note. The gray background area and the right axis refer to the total corpus size. Values for 2023 (dots) are preliminary (Nov 27).

The bi-factor model enables new insights but also brings new challenges. Due to the relative recency of its rise to popularity, research on the (CFA) bi-factor model itself is still relatively scarce (Bader & Moshagen, 2022; Bornovalova et al., 2020; DeMars, 2013; Eid et al., 2017; Markon, 2019; Reise, 2012; Rodriguez et al., 2016; B. Zhang et al., 2021). The current work addresses three major open challenges: weak specific factors, excessive model flexibility, and inflexible application. First, specific factors are often found to have small factor loadings and little variance (Eid et al., 2017). The first article in this dissertation discusses the likely origin of this problem and analyzes the role of effect size and statistical power to derive guidelines for applied research (Petras & Meiser, 2024). Second, the bi-factor model is well-known to be extremely flexible (Bader & Moshagen, 2022; Bonifay & Cai, 2017), which raises the question if the different model variants (Eid et al., 2017) can be clearly distinguished. The second article in this dissertation clarifies the relationship between the bi-factor model's variants (Petras, 2024a). Third, most applications of the bi-factor model use the basic scheme of the full bi-factor structure (all observed variables relate to one specific factor), even if this is highly questionable. The third article in this dissertation provides a principled model-specification workflow that optimizes the representation of specific content beyond the general trait more flexibly (Petras, 2024b).

It follows a detailed introduction of the discussed models, a summary and synopsis of the work within the three articles, and a discussion within the larger context of psychological measurement. After the conclusion, the full texts of the three articles of this dissertation are appended.

2.1 Confirmatory factor models

Confirmatory factor analysis (CFA) models assume that there is an underlying set of latent variables (factors, $\boldsymbol{\eta}$), which relate to the observed variables ($\mathbf{Y}$) via the matrix of factor loadings ($\boldsymbol{\Lambda}$). In the j'th column of $\boldsymbol{\Lambda}$, the strength of the influence of the j'th latent variable on the i'th observed variable is indicated in the i'th row. Equation 1 shows the computation of the model-implied covariance matrix $\boldsymbol{\Sigma}_\theta$.

$$\boldsymbol{\Sigma}_\theta = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_e \tag{1}$$

$\boldsymbol{\Sigma}_\theta$ depends on the estimates of the free parameters in $\boldsymbol{\Lambda}$ and in the covariance matrix of the latent variables ($\boldsymbol{\Phi}$). These are estimated to optimize a criterion regarding $\boldsymbol{\Sigma}_\theta$, such as maximizing the likelihood of obtaining the observed $\boldsymbol{\Sigma}$ assuming that the model-implied $\boldsymbol{\Sigma}_\theta$ was true in the population (maximum likelihood estimator). Equation 2 shows how this CFA model predicts the observed responses of person p.

$$\mathbf{Y}_p = \boldsymbol{\Lambda}\boldsymbol{\eta}_p + \boldsymbol{\epsilon} \tag{2}$$

The measurement error ($\boldsymbol{\epsilon}$) is usually assumed to follow a multivariate normal distribution in the population. A common assumption is the independence of errors, making its covariance matrix $\boldsymbol{\Theta}_e$ a diagonal matrix.

CFA models restrict the pattern of relationships between observed and latent variables in the matrix of factor loadings ($\boldsymbol{\Lambda}$) based on prior assumptions to identify the model for estimation and to test hypotheses. CFA is primarily used to compare and select models defined by certain restrictions and to examine parameter values within selected models. That means if there are multiple latent variables, many entries in $\mathbf{\Lambda}$ are fixed to 0. CFA models usually assume a metric scale of the observed variables with linear relationships between latent and observed variables. Nevertheless, the structure of CFA models (i.e. $\mathbf{\Lambda}$ and $\mathbf{\Phi}$) can generally be translated into an IRT model, which assumes a nonlinear relationship between response categories and latent variables. Importantly, the bi-factor CFA model has been translated into an item bi-factor model (Cai et al., 2011), meaning that the implications of the current work on the bi-factor model are not necessarily specific to the SEM framework.

Figure 1 shows different CFA models that represent psychological constructs. group-factor models are common models derived from exploratory factor analysis (EFA) or hypothesized in CFA models. group-factor models define one or more correlated factors (first diagram in Figure 1). Despite the split of the target construct into multiple dimensions, this modeling approach has been most popular for decades. More recently, hierarchically structured models were popularized: the higher-order factor model and the bi-factor model. The higher-order factor model provides a factor structure at the second level: it estimates one or more higher-order factors from the first-order factors (second diagram in Figure 1). The factors at the second level account for the correlations between the factors at the first level. The bi-factor model instead relates the observed variables directly to the general factor, as well as specific factors representing the facets of the target construct (third diagram in Figure 1).

2.2 The bi-factor model

The idea of the bi-factor model is to represent a common target trait of all observed variables, as well as specific influences that are reflected by subgroups of observed variables. It decomposes the systematic variance in the observed variables into general and (domain-)specific variance. Therefore, it is characterized by a $\mathbf{\Lambda}$-matrix with two nonzero entries per row: one for the general factor and one for the respective specific factor. Equation 3 shows an example with twelve observed variables and four specific factors (equal to the third path diagram in Figure 1).

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{1,g} & \lambda_{1,1} & 0 & 0 & 0 \\ \lambda_{2,g} & \lambda_{2,1} & 0 & 0 & 0 \\ \lambda_{3,g} & \lambda_{3,1} & 0 & 0 & 0 \\ \lambda_{4,g} & 0 & \lambda_{4,2} & 0 & 0 \\ \lambda_{5,g} & 0 & \lambda_{5,2} & 0 & 0 \\ \lambda_{6,g} & 0 & \lambda_{6,2} & 0 & 0 \\ \lambda_{7,g} & 0 & 0 & \lambda_{7,3} & 0 \\ \lambda_{8,g} & 0 & 0 & \lambda_{8,3} & 0 \\ \lambda_{9,g} & 0 & 0 & \lambda_{9,3} & 0 \\ \lambda_{10,g} & 0 & 0 & 0 & \lambda_{10,4} \\ \lambda_{11,g} & 0 & 0 & 0 & \lambda_{11,4} \\ \lambda_{12,g} & 0 & 0 & 0 & \lambda_{12,4} \end{pmatrix} \tag{3}$$

The corresponding factor covariance matrix ($\mathbf{\Phi}$) is diagonal, meaning that the factors are orthogonal to each other. This ensures that the model is estimable and the specific factors can be interpreted as unique influences beyond the general factor.

This model is particularly well suited to represent psychological constructs that consist of multiple facets but still represent one overarching trait. For example, a comprehensive definition of agreeableness might comprise compassion, respectfulness, and trust as subordinate facets of agreeableness without rejecting agreeableness as a singular target dimension (Soto & John, 2017). In that case, the general factor of the bi-factor model would be interpreted to represent the core of agreeableness. The specific factors of the facets would be interpreted to represent the unique characteristics of the facets above and beyond this core. For example, the specific compassion factor would be interpreted to represent what is unique to compassion – beyond the general notion of agreeableness and the other facets. A person with a high score on this factor would show a higher level of compassion than is typical for persons with the same level of core agreeableness (cf. DeMars, 2013). In that sense, the specific factors are residuals relative to the general agreeableness trait. This stands in marked contrast to the group-factor model, in which a high compassion factor score simply indicates a high level of compassion of the person. All factors in the model – the agreeableness factor and the three facet-specific factors – can be interpreted as relevant dimensions under a comprehensive definition of agreeableness.

2.2.1 Variants of the bi-factor model

Several variants of the bi-factor model have been introduced by Eid et al. (2017) to improve interpretability in cases with specifically selected domains (e.g., facets of a trait, subscales of a questionnaire), as opposed to randomly sampled domains (e.g., randomly sampled raters). The core idea of the proposed S-1 and S*I-1 models is to select a reference domain (or item) that defines the meaning of the general factor. The fourth path diagram in Figure 1 shows an S-1 bi-factor model, in which the first specific factor is omitted compared to the full symmetrical bi-factor model (S-model, third path diagram above). In the terminology of classical test theory, the meaning of the general factor is then based on the true score of the reference domain for which there is no specific factor. This true score comprises the shared general true score and the true score specific to the reference domain (Eid et al., 2017). This model adaptation is similar to the $CTC(M-1)$ variant of the $CTCM$ multitrait-multimethod model, in which the reference refers to the reference method of measurement, such as a gold standard measure (Eid et al., 2003, 2022). The selection of a reference is different if the specific factors refer to content domains: the reference should then most closely resemble the general trait of interest, to obtain the most meaningful interpretation of the factors (Eid et al., 2017). A variation of the S-1 bi-factor model is shown in the fifth path diagram of Figure 1: in the S-1c bi-factor model, the remaining specific factors have their correlations estimated freely. Finally, the S*I-1 variant uses a single item as the reference, instead of a whole domain. Eid et al. (2017) suggest interpreting the specific factors in all these variants as follows[3]:

> For each domain (with exception of the reference domain) a specific factor is defined as a residual factor. Such a specific factor represents that part of

[3]This interpretation is challenged by the observation that factor scores of the specific factors in the S-1 and S-1c model systematically correlate with those of the specific factor of the reference domain from the S bi-factor model of the same data (Petras, 2024a). This means that, at least when an S-1 or S-1c model is estimated in practice, the further specific factors are related to the specific true score variance of the reference domain, contradicting the interpretation given by Eid et al. (2017).

> a domain that is not shared with the reference domain. (Eid et al., 2017, p. 550)

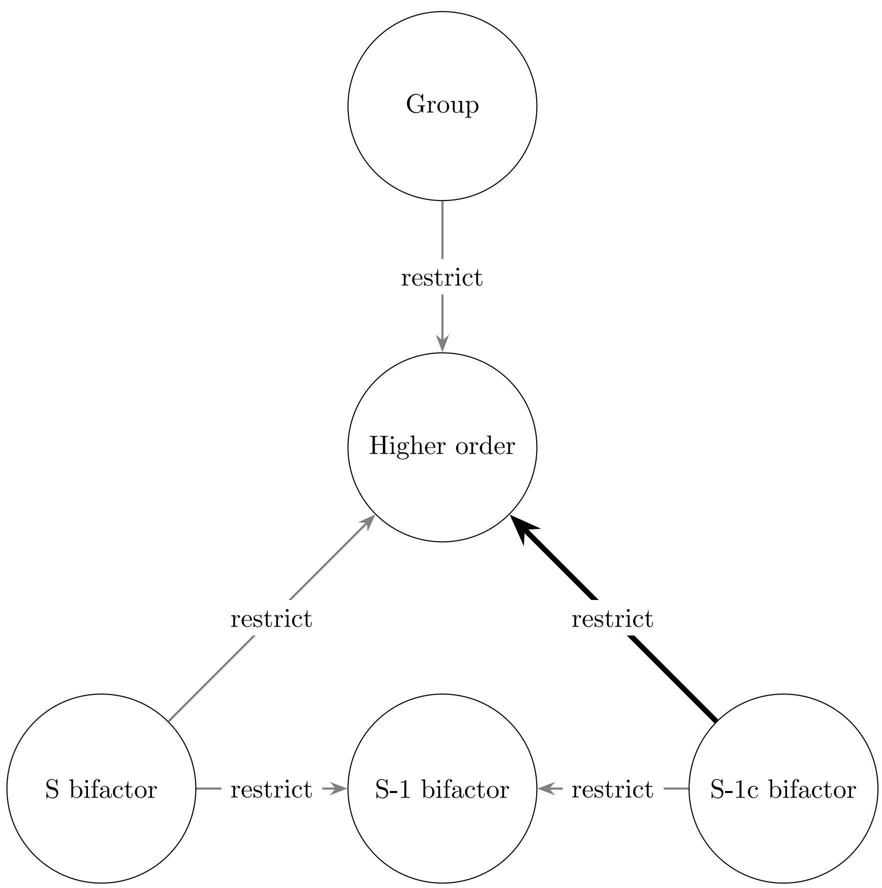
These model variants do not only come with theoretical reasons to select them but also have an interesting statistical relationship to one another.

2.2.2 Nesting structure

The relationship between the model variants needs to be understood to interpret the variants and the differences in their model fit to a given dataset. Figure 3 (Petras, 2024a, fig. 2) shows the nesting structure of the bi-factor model variants and the related higher-order and group-factor models.

Figure 3

Nesting structure of confirmatory factor models.



Most of these nesting relationships have been established previously in the literature. Trivially, the higher-order factor model imposes a structure on the correlations between the first-order factors, thereby restricting the group-factor model in cases with four or more first-order factors. Similarly, the S bi-factor model can be restricted to be equivalent to the higher-order factor model by imposing a proportionality constraint on the factor loadings (Yung et al., 1999). The S-1 bi-factor model is nested trivially in both the S and the S-1c bi-factor model since it is defined by restrictions of individual parameters relative to the other two variants. The relationship between the S-1c model and the higher-order model is less trivial and has not been described previously. Petras (2024a) shows how an S-1c parameterization can be computed for every higher-order factor model, meaning that the higher-order factor model is nested in the S-1c bi-factor model variant. The exact restrictions that need to be imposed on the S-1c model to be equivalent to the higher-order model are non-trivial, though. The more complex relationship between the S and S-1c variants is further explored in Petras (2024a).

3 Challenges in bi-factor model research

3.1 Weak specific factors and loadings

Petras, N., & Meiser, T. (2024) Problems of domain factors with small factor loadings in bi-factor models, *Multivariate Behavioral Research*, *59*(1), 123-147. https://doi.org/10.1080/00273171.2023.2228757

In the first article of this dissertation, I tackle the previously reported problem of weak specific factors and loadings in bi-factor model applications (Eid et al., 2017). The original work by Eid et al. (2017) suggests the model's improper account of the sampling structure (specific domains selected and not randomly sampled) was responsible for this problem but does not systematically consider other reasons for estimates of factor variances and factor loadings to be surprisingly small (or even negative). In their analysis of the prevalence and role of non-significant estimates, they do not consider the statistical power underlying the significance tests or the true strength of the factors in the population. In this context, it is important to realize that bi-factor model applications for many measures analyzed with a bi-factor model were developed using a different model, such as a group-factor model derived from exploratory factor analysis. Therefore, the expectation that the bi-factor model should yield substantial estimates of specific factor loadings should be questioned in those cases. The current work brings these two observations together. The meta-analysis of factor loading estimates shows that non-significant and negative specific factor loadings are part of a larger pattern that can be explained by the way questionnaire items were most likely selected. As a meaningful general metric of the strength of a factor, I propose to use the sum of squared standardized loadings. The simulation study shows that the statistical power to detect a specific factor is usually sufficient, even if the factor is rather weak. This is especially true when using the most relevant but rarely computed likelihood ratio test that compares the model with the factor in question to the model without it. The simulation study shows that it is relevant to consider simulation results even if analytical power estimates (Moshagen, 2021; Moshagen & Bader, 2023) are available. The frequent occurrence of non-convergence can lower the effective statistical power substantially below the analytically computed value, especially if cases of non-convergence are counted as failures in combination with the errors (β). Another previously undiscussed issue regarding weak factors is the reliability of estimates: it is often not enough to show that a parameter is non-zero since researchers also want to interpret the estimated values of factor loadings, factor variances, and factor scores. The simulation shows that the factor scores of weak specific factors are relatively unreliable regardless of sample size, and that weak factors in the population can lead to a much higher frequency of anomalous results, such as non-convergence and negative variance estimates. For loadings, however, low reliability cannot explain small estimates as random variations from supposedly substantial population parameters, corroborating the meta-analysis finding of many near-zero factor loadings. The S-1 and S-1c model variants, which were suggested to fix the problem of irregular estimates and non-significant estimates, showed almost no difference on any metric compared to the standard S variant. The only major difference is that defining a superfluous specific factor, which is more likely to happen in the S variant than in the other variants, produces more convergence problems than defining one factor too few. The main conclusion from this work is that the crucial step to obtain a useful bi-factor model with substantial parameter estimates is to provide adequate data from measures that are designed for the use of the bi-factor model. This issue would be less subtle if measures were developed with strict adherence to formalized theories that include a precise understanding of the constructs' dimensionality. Moreover, several guidelines for planning, troubleshooting, and interpreting bi-factor models are derived for applied research. This work does not invalidate the suggested S-1 or S-1c models, since they still offer a more straightforward interpretation whenever domains are specifically selected (i.e. not at random).

3.2 Extreme flexibility

Petras, N. (2024a). When factor variance and factor correlations are interchangeable: The relationship between the bi-factor model variants [manuscript submitted for publication]. Department of Psychology, University of Mannheim.

In the second article, I examine the high flexibility of the bi-factor model. A preference for the bi-factor model has previously been reported in studies on the selection between the bi-factor model and competing models (Bonifay & Cai, 2017; Greene et al., 2019). Bader and Moshagen (2022) clarified that fit indices are not biased in favor of the standard bi-factor model, but the bi-factor model is flexible to imitate other models. It is thereby an equally valid account of the data. The second article investigates how well the different bi-factor model variants can be distinguished and how their estimates behave when they attempt to imitate each other.

The current work adds to the understanding of the nesting structure (Figure 3, Petras, 2024a) by proving that the higher-order model is nested within the S-1c bi-factor model. Therefore, the crucial remaining comparison involves the non-nested S and S-1c models. These add to the more restricted S-1 and higher-order models in two different ways. Compared to the S-1 variant, the S variant adds a specific factor, whereas the S-1c variant adds freely estimated correlations between the specific factors. This can be equivalent in the special case of the higher-order model. Therefore, the corresponding model parameters in the different models change their values – and thereby their interpretation – when swapping between the parameterizations. For example, reparameterizing a higher-order model as an S-1c bi-factor model produces positive correlations between the domain-specific factors. These correlations are all zero in the equivalent parameterization as an S model. On the other hand, the restricted S-parameterization defines a whole factor that completely disappears in the equivalent S-1c-parameterization.

The reported simulation study examines the relationship between the S and S-1c variants beyond cases in which the different variants provide strictly equivalent solutions. This simulation shows that fit indices are very limited in distinguishing the S and S-1c models because the variants can imitate each other well. It indicates that the S-1c variant is even more flexible than the S variant. The simulation highlights the importance of the analytical results on the nesting structure: the parameter estimates when swapping from S to S-1c generally change in the same direction as in the special case of the higher-order factor model. When swapping from S-1c to S, the simulation shows that a “ghost” factor for the reference domain is created and that its meaning depends on the correlations between the specific factors of the initial S-1c model.

In sum, the mutual imitation of bi-factor model variants can produce several variations in the estimates of corresponding parameters. There is a highly misleading switch between either obtaining substantial estimates of the specific factor of the reference domain or obtaining substantial (positive) estimates of the correlations between the specific factors of the further domains. In some sense, this makes bi-factor model parameter interpretation almost impossible: the statement that there is a positive relationship between the specific factors is not generally true if a model that fixes them to zero can explain the data just as well. Similarly, the statement that there is shared specific variance on any particular domain is not generally true, if a model that omits the specific factor of this domain can explain the data just as well. This is an excellent reminder that being able to define a latent variable with a positive variance does not prove the existence of anything in particular.

Whereas the article clearly maps the problem of excessive model flexibility, it is hard to come up with guidelines for model selection and parameter interpretation in applied research. From a statistical point of view, there is no clear preference for one of the models if they fit the data equally well and make almost identical predictions. Researchers can either make no model selection decision and report estimates in both variants, or use other selection criteria. Eid et al. (2017) recommend to use S-1, S-1c, or S*I-1 models to obtain a better interpretable general factor whenever the specific domains are not randomly sampled. This recommendation serves to clearly define the involved variables as random veriables on an explicated set of outcomes (Eid et al., 2017), but it does not provide a clear interpretation of all its parameters. The current work uncovers that the meaning of parameters in such a flexible model is fundamentally fuzzy, because patterns in the data can be mapped on different parts of the model. To understand the consequences of model selection on the parameter estimates, the article provides a detailed analysis of the relationships between the parameters of the two variants and shows that all parameters of the model change in value systematically when swapping between variants.

This work has an important connection to the weak factor problem discussed in the first article of this dissertation. Because the size of the parameter estimates varies substantially depending on the choice of model variant, there can be relevant differences in statistical power of the respective hypothesis tests. This is especially problematic if the true population value of the parameter is small. Therefore, one of the variants can be superior to the others in generating a full set of significant specific factor loading estimates. As noted in this second article, this explains such a finding in the simulation study by Geiser et al. (2015), which has previously been attributed to the model's supposedly inferior representation of the sampling process. However, the simulation from the first article could not show a general superiority of one of the variants across a large variety of cases. In addition, the way the S and S-1c model mutually imitate each other offers an alternative interpretation of weak domain factors: potentially, they are methodological artifacts of the choice against correlations between the further specific factors.

3.3 Schematic restrictions

Petras, N. (2024b). Building hierarchically structured factor models with systematically selected residual correlations [manuscript submitted for publication]. Department of Psychology, University of Mannheim.

Bi-factor models are usually used schematically, starting with one of the model variants. Yet, there is no statistical or empirical necessity to only consider models in which all observed variables load on a specific factor or belong to a specifically chosen reference domain. Only if the measure is designed to generate this data structure this default makes sense. The highly attractive hierarchically structured modeling approach of the bi-factor model does not need to be limited in this way. In the third article, I propose an alternative four-step approach: (1) choose a baseline model, (2) establish a hierarchy of relevance among potential residual correlations, (3) choose a number of residual correlations, and (4) estimate the final model on a new sample. The suggested approach systematically solves several problems that plague previous research. In the first step, the default baseline model is a single-factor model, and only those specific factors that are clearly and demonstrably relevant, are added. These factors should be specified based on prior knowledge or key assumptions of the study design to provide a clear interpretation. They should be tested to not only improve the overall model fit when included but also to produce substantial and interpretable factor loading estimates. This avoids specifying factors merely for the sake of the completeness of the model structure. In the second step, the hierarchy of relevance among residual correlations is obtained using Bayesian lasso regularization (Pan et al., 2017; see also Park & Casella, 2008), which is free of several problems of alternative methods, such as the iterative nature of modification indices. The genius of the Bayesian lasso regularization for residual correlations as introduced by Pan et al. (2017) lies in the possibility of estimating all of them at the same time, which is strictly impossible in the frequentist framework, because a model with that many parameters is not identified there. In the third step, the parsimony of the final model is optimized by selecting only the relevant residual correlations (cf. Pan et al., 2017; L. Zhang, Pan, & Ip, 2021). The third article examines the reproducibility of this procedure in a multiverse analysis, showing that it generally yields statistically meaningful results. In the fourth step, the final model is then estimated on a new (sub-)sample, which combines the advantages of confirmatory and exploratory analysis. Other than specific factors in the standard bi-factor model, item pairs with correlated residuals can be overlapping (similar to cross-loadings), which may better represent the complexity of typical questionnaire items. The application example presented in the third article shows that this flexibility offers not only a more plausible model of the data but can also outperform the standard bi-factor approach in both parsimony and model fit simultaneously. The general factor of the final model can be interpreted in the same way that a general factor of a bi-factor model is interpreted: as a measure of the target construct that is clear of domain-specific or item-specific influences. The resulting model has a hierarchical structure, in which the residual correlations can be interpreted as specific variance portions – similar to specific factors in the bi-factor model.

This approach is an important alternative to the inclusion of weak specific factors, as discussed in the first article. It discourages researchers from specifying factors for the sake of completeness and encourages them to come up with the interpretation of potential specific factors before specifying any. In addition, the systematic selection of residual correlations can easily replace weak specific factors that are merely glorified residual correlations with very small factor loadings on all but two of their items. The approach also offers a new, flexible twist on the selection of bi-factor model variants. Especially for applications that aim to optimize the statistical representation of a given measure (instead of optimizing the measure), the suggested procedure provides a pathway to find and test a well-fitting, parsimonious model with a well-interpretable structure.

4 Improving bi-factor model applications

The bi-factor model is a promising approach to psychological measurement and theory testing because it neatly distinguishes the important parts of psychological constructs. The current work highlights three key challenges regarding the bi-factor model that are likely to impede theory testing if the bi-factor model is applied without considering them.

4.1 Recommendations

Researchers can improve the measurement of psychological constructs by considering the expected strength of specific factors a priori, with a sum of squared (standardized) loadings greater than one as a good benchmark for usable specific factors (Petras & Meiser, 2024). Compared to current practice, this means that a lot of measures need to be extended or revised if researchers want to test theories regarding specific factors, such as content domains in personality scales. In Petras and Meiser (2024), it was also shown that a switch to the S-1 model does little to nothing to address the problem of specific factors being weak and should only be considered for reasons of interpretability. The frequent occurrence of weak specific factors can not only be seen as a limitation of current measures. It also is a hint that theories including them may need to be questioned. This adds another concern resulting from the lack of formalized definitions. Measures of the same construct frequently define different subdomains and sample different item content (e.g., Fried, 2017), a symptom of inconsistent theory and terminology (“jingle-fallacy,” Flake & Fried, 2020). This problem is especially visible on the subscale level: measures of (supposedly) the same construct define a jungle of facets (e.g. of the Big Five personality traits) with inconsistent terminology and coverage across measures. Therefore, additional theoretical work may improve the statistical properties of psychological measures by weeding out inconsistent facets.

Researchers can improve the measurement of psychological constructs by providing theoretical arguments for the use of a particular bi-factor model variant (as in Eid et al., 2017). When in doubt, it is prudent to estimate multiple variants to check the sensitivity of conclusions to variations in the modeling approach. These two points are relevant, as the S and S-1c variants of the bi-factor model can imitate each other exceptionally well and the flexibility of the bi-factor model means that its fit is a bad indicator of the model structure reflecting the data-generating process in the population (Petras, 2024a). All seemingly equivalent parameters systematically change their values when switching between the model variants, which means that all hypothesis tests are potentially sensitive to the choice of variant.

Researchers can improve the measurement of psychological constructs by carefully considering the structure of the construct (and measure) at the specific level of a hierarchically structured factor model. Petras (2024b) offers a comprehensive new approach to model specification that flexibly builds on the schematic bi-factor model variants. This approach avoids nonsensical "hypothesis" tests regarding factors that were specified merely for completeness' sake and whose meaning is unclear. Furthermore, it offers a data-driven approach to identify relevant pairwise relationships between observed variables. To select a fitting modeling approach, researchers need to realize what their main goal is: to obtain a model (or measure) that closely resembles a theorized structure, or to provide a well-fitting account of the data. The proposed approach strikes a balance by explicitly accommodating both. It meaningfully represents the construct by exclusively extracting theoretically relevant factors in a rigid a priori model structure (baseline model). It also assures that the final model fits the data closely (yet parsimoniously), using a flexible, data-driven selection of residual correlations. In the empirical example, the proposed approach yielded a final model that fitted the data better than the respective full bi-factor model and at the same time was more parsimonious. This shows the potential for improvement when modeling the domain-specific level of psychological constructs with this approach. Alternatively, this finding can be understood to show the potential to improve the measure towards a more meaningful subscale structure.

The proposed new approach can be useful in the test of theories on the general trait, even if the loading pattern on the general factor may not change compared to the standard bi-factor model. Accurately modeling the specific relationships between the observed variables makes it possible to judge the importance and meaning of otherwise unexplored influences that lead to an unexplained bad fit in simpler accounts of the data and contribute to ill-defined specific factors in the standard bi-factor model. Ideally, only a few easily interpretable residual correlations are found so that the general trait can be interpreted with confidence. Exploratory Structural Equation Modeling (ESEM, Asparouhov & Muthén, 2009; Marsh et al., 2010, 2014) offers some of the same advantages. In comparison, the suggested approach in Petras (2024b) leads to a much more parsimonious model by drawing a line between meaningfully strong relationships that should be included as parameters in the structure of the model, and parameters that would only catch noise and are therefore excluded.

For the selection of models, the previous bi-factor literature lists several schematic variants of the model as options (S, S-1, S-1c, higher-order model) and suggests that the interpretation resulting from the sampling structure (randomly selected versus picked domains) is the major reason for selecting between variants (Eid et al., 2017). The current work takes a different perspective: it focuses on an accurate and sparse description of a measure's content beyond the general trait. An important first step is to abandon the exclusive use of complete bi-factor models in cases where the inclusion of some of the specific factors (or the allocation of some of the observed variables to the factors) is highly questionable (Petras & Meiser, 2024). Secondly, a sparse selection of residual correlations can be used instead of – or in combination with – specific factors to represent specific variance efficiently and flexibly (Petras, 2024b). If the design of the measure and the data merit the use of a schematic variant of the bi-factor model, researchers should be aware of the consequences of their choice. Whereas the S-1 and S-1c models do offer a straightforward interpretation (Eid et al., 2017), their statistical relationship to the S model is complex (Petras, 2024a).

Beyond model fit indices, the judgment of the fit of a bi-factor model (or hierarchically structured factor model) to the data should include an interpretation of the parameter estimates. A good model fit is only as informative as the restrictions of the model. Due to the flexible, relatively unrestricted nature of the bi-factor model, drawing conclusions about the true data-generating process is very limited (Petras, 2024a), especially when traditional cut-offs for a "good" model fit (Hu & Bentler, 1999) are used. It is important to keep in mind that the more flexible a model is, the less informative its fit to the data is about the data-generating process in the population (Roberts & Pashler, 2000). An excessively flexible model can imitate almost any data-generating process. Specifically, the S-1c and S bifactor models imitate each other almost perfectly (Petras, 2024a), despite leading to different conclusions about the existence and interrelationships of specific factors. Even less informative is model fit in a Bayesian lasso-informed model as proposed in (Petras, 2024b). The process that leads to the selection of the final model almost guarantees its good fit to the data. To understand if there is a misfit between the structure of the model and the data-generating process, it is then more relevant to examine the estimates themselves. An extreme example of that would be a bi-factor model that fits the data well but produces near-zero estimates on the general factor loadings for one or more subscales (Bornovalova et al., 2020, fig. 1B). The model fit may tempt researchers to accept that the bi-factor structure represents the population well, but the estimates indicate that one or more subscales do not measure the target construct's common core at all, which is usually an uninterpretable result regarding the to-be-tested theory. Similarly, model fit indices do not indicate if specific content (such as a pair of correlated residuals) is desired to be in the measure. it merely indicates how well the model fits the data if the specific content is accounted for. To summarize, the inherent flexibility of the proposed bi-factor approaches can result in a close fit of the model to a very badly designed measure. Therefore, in all the discussed models, researchers should closely inspect the model parameter estimates to judge if the model structure represents the data well (see also Watts et al., 2019). The proposed approach in Petras (2024b) avoids at least some misleading conclusions based on model fit by avoiding the specification of superfluous specific factors or residual correlations.

4.2 Limitations and outlook

The current work leaves some open questions and sparks new ones. Although Petras (2024a) provides a better understanding of the relationship between the bi-factor model variants, documenting their inherent flexibility to imitate each other doesn't fully resolve the issue. In light of the growing popularity of the bi-factor model and even more flexible approaches, such as exploratory structural equation modeling (ESEM, Asparouhov & Muthén, 2009; Marsh et al., 2010, 2014), it seems relevant to reconsider and build on the work by Roberts and Pashler (2000). Roberts and Pashler (2000) raise the concern that the structure of very flexible models may not be as meaningful as researchers think – at least not due to their excellent model fit, because it is known a priori that such flexible models will fit the data well. A key focus of future research may be the identification of relevant model restrictions within the discussed models that can be empirically tested to establish a better understanding of the construct and to improve measures. Furthermore, Petras and Meiser (2024) and Petras (2024a) exclusively focussed on idealized simulated data with continuously (normally) distributed traits and errors. It remains open, how well the conclusions generalize to other cases, such as item bi-factor models of categorical data.

The current work identified several underlying, difficult, and unresolved statistical problems. The work on the reliability and bias in bi-factor scores (Petras, 2024a; Petras & Meiser, 2024) has uncovered that scores are systematically biased relative to one another. It remains unclear if this generalizes to alternatives of the computed regression factor scores, such as plausible values (Wu, 2005). This finding adds to the list of easily overlooked challenges regarding factor scores (for an overview, see Lechner et al., 2021). Furthermore, the problem of setting an inclusion criterion for residual correlations using the Bayesian lasso has previously been tackled using simulation studies with idealized data (L. Zhang, Pan, & Ip, 2021). The replication attempts in the current work show that any cut-off will likely produce limited replicability of inclusions. The study highlights the need for a principled criterion instead of a conventional rule of thumb (Petras, 2024b). Finally, the usefulness of fit indices that account for parsimony was very limited and riddled with contradictions when comparing a large number of models with different numbers of residual correlations Petras (2024b). This showcases that the current practice of using multiple indices in parallel is an insufficient band-aid fix in the unsolved problem of balancing fit and parsimony. This issue complicates not only the proposed modeling approach.

Given the discussed issues with bi-factor models, researchers need to report their studies in appropriate detail. It is important to provide at least the covariance matrix of the observed variables, or better, the full raw data. This is not only necessary for meta-research like the meta-analysis in Petras and Meiser (2024). Every study that does not report the covariance matrix or the data forces researchers to collect new data for every new statistical method or robustness check, slowing down applied research, research on scale validation, and meta-research massively. There have been several massive shifts in the statistical approach in the past (e.g. from a strong preference for group-factor models towards bi-factor models), but the closed-data research culture has prevented researchers from retroactively applying new state-of-the-art methods to many publications. Furthermore, researchers should report the model estimates of bi-factor models clearly in the main article and discuss the size of factor loadings and strength of factors when discussing the measurement model. As discussed above, the interpretability of bi-factor models hinges as much on the pattern of estimates as on the model's fit. When choosing a model variant and allocating variables to factors at the specific level, researchers need to reflect and report their rationale. The current work has shown that it is often problematic to merely adopt the allocation of variables to factors from a group-factor model to a bi-factor model (Petras & Meiser, 2024). Petras (2024b) provides an alternative approach to the use of the classic bi-factor model by allowing a more informed specification at the specific level. In any of these cases, it is crucial to clearly state why the specific level of a hierarchically structured model is defined in a particular way and how the meaning of specific factors is derived.

It is just as important to focus on improving measures as on improving statistical models. Besides fine-tuning the model to a particular version of the measure, the measure can also be fine-tuned itself. Especially if researchers aim to measure domain-specific traits reliably, for example, to test more precise theories about relationships between psychological constructs, improvements and extensions of existing measures are necessary (Petras & Meiser, 2024). The possibility to model almost anything with flexible approaches, such as the approach proposed in (Petras, 2024b), may tempt researchers to prioritize details in existing measures over the more important goal of developing good theories and measures. For this purpose, the hierarchical structure in the bi-factor model and related models is particularly useful, because it allows researchers to identify and isolate wanted and unwanted specific content beyond the core of the target construct. Such specific content may either be purposefully selected for or may be eliminated from a measure during item selection. Therefore, it may be very fruitful to thoroughly consider the potential structure of a measure at the specific level already during the writing of items and use bi-factor models to select items with the desired content.

Future research may consider the question of how strong and reliable a specific factor needs to be to be useful for the structural part of a structural equation model. One major application of the bi-factor model is to disentangle the relationships of the parts of a psychological construct with other variables, capitalizing on its orthogonally defined factors to test differentiated theories of the measured construct. The specific factors' usefulness in such a scenario would be a good indicator to judge how measures need to be designed. Petras and Meiser (2024) shows clearly that this is a concern: bi-factor model applications frequently feature specific factors that are too weak (or too weakly related to some of their items) to be properly interpreted.

Finally, the development and integration of software need more attention. The proposed approach in Petras (2024b) is implemented as a modified version of the custom-code Gibbs-sampler by Pan et al. (2017), which has been further developed independently by others with a focus on generating software-specific MPlus syntax (L. Zhang, Pan, Dubé, et al., 2021). Neither code is integrated with other software, yet, such as the massively popular `lavaan` package (Rosseel, 2012) for SEM in R. `lavaan` automatically computes modification indices for residual correlation selection and is not yet linked to the arguably much superior Bayesian lasso approach by Pan et al. (2017). Similarly, the use of the Wald-test is much simpler for users of `lavaan`, than the use of the LRT. This may contribute to its widespread use: in the meta-analysis reported in Petras and Meiser (2024), there was not a single application example in which the superior (conceptually and performance-wise) LRT was performed, but many papers reported results of the Wald-test of factor variances. Updating software to provide easy access to the discussed and proposed advanced statistical methods to a broad user base is a necessary next step after sharing these ideas conceptually in journal publications.

5 Conclusion

The bi-factor model is popular in psychological measurement for good reasons. It provides a compelling representation of psychological constructs. Its distinction between general and specific variance enables researchers to answer nuanced research questions. On the other hand, the current work identifies three major concerns regarding the bi-factor model and its routine application. In sum, the work shows that a successful test of nuanced theories via hierarchically structured factor models, such as the bi-factor model, requires a more advanced understanding of the statistical oddities of the bi-factor model, as provided in the first two articles (Petras, 2024a; Petras & Meiser, 2024). Furthermore, it became clear that there are practical challenges beyond the purely statistical discussion of the model variants. The current work extends the statistical toolkit and provides a go-to approach (Petras, 2024b) but also highlights that research on the preconditions of meaningful hypothesis tests is necessary. Activities, such as the extension and revision of measures and the development of theoretical arguments that can decide between competing model specifications, are necessary to test meaningful hypotheses about complex, unobservable psychological constructs.

6 Bibliography

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Bader, M., & Moshagen, M. (2022). *No probifactor model fit index bias, but a propensity toward selecting the best model.* *131*(6), 689–695. https://doi.org/10.1037/abn0000685

Bagozzi, R. P. (2007). *On the meaning of formative measurement and how it differs from reflective measurement: Comment on Howell, Breivik, and Wilcox (2007).* https://doi.org/10.1037/1082-989X.12.2.229

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

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*, *88*(1), 18–27. https://doi.org/10.1016/j.biopsych.2020.01.013

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221. https://doi.org/10.1037/a0023350

Canivez, G. L., Grieder, S., & Buenger, A. (2021). Construct validity of the german Wechsler intelligence scale for children–fifth edition: Exploratory and confirmatory factor analyses of the 15 primary and secondary subtests. *Assessment*, *28*(2), 327–352. https://doi.org/10.1177/1073191120936330

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. *The Scientific Study of General Intelligence*, 5–21. https://doi.org/10.1016/B978-008043793-4/50036-2

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

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

//doi.org/doi.org/10.1037/met0000083

Eid, M., Koch, T., & Geiser, C. (2022). Multitrait–multimethod models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 349–366). Routledge.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-c (m-1) model. *Psychological Methods*, *8*(1), 38.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press. https://doi.org/10.4324/9781410605269

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 953–986. https://doi.org/10.1002/9781118489772.ch30

Eronen, M. I., & Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory. *Theory & Psychology*, *30*(6), 786–799. https://doi.org/10.1177/0959354320969876

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Gäde, J. C., Schermelleh-Engel, K., & Klein, A. G. (2017). Disentangling the common variance of perfectionistic strivings and perfectionistic concerns: A bifactor model of perfectionism. *Frontiers in Psychology*, *8*, 160. https://doi.org/10.3389/fpsyg.2017.00160

Geiser, C., Bishop, J., & Lockhart, G. (2015). Collapsing factors in multitrait-multimethod models: Examining consequences of a mismatch between measurement design and model. *Frontiers in Psychology*, *6*, 946. https://doi.org/10.3389/fpsyg.2015.00946

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., Waldman, I. D., Cicero, D. C., Conway, C. C., Docherty, A. R., et al. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *Journal of Abnormal Psychology*, *128*(7), 740. https://doi.org/10.1037/abn0000434

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*, 537–549. https://doi.org/10.1007/s11336-011-9218-4

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77*(3), 442–454. https://doi.org/10.1007/s11336-011-9218-4

Kline, R. B. (2023). Structural equation modeling. *New York: Guilford*.

Lechner, C. M., Bhaktha, N., Groskurth, K., & Bluemke, M. (2021). Why ability point estimates can be pointless: A primer on using skill measures from large-scale assessments in secondary analyses. *Measurement Instruments for the Social Sciences*, *3*(1), 1–16. https://doi.org/10.1186/s42409-020-00020-5

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*, 51–69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471. https://doi.org/10.1037/a0019227

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

McGuigan, F. (1953). Formalization of psychological theory. *Psychological Review*, *60*(6), 377. https://doi.org/10.1037/h0059119

Moshagen, M. (2021). *semPower: Power analyses for SEM*. https://CRAN.R-project.org/package=semPower

Moshagen, M., & Bader, M. (2023). semPower: General power analysis for structural equation models. *Behavior Research Methods*, 1–22. https://doi.org/10.3758/s134

28-023-02254-7

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The bayesian lasso. *Psychological Methods*, *22*(4), 687. https://doi.org/10.1037/met0000112

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*(3), 524. https://doi.org/10.1037/0022-3514.81.3.524

Petras, N. (2024a). When factor variance and factor correlations are interchangeable: The relationship between the bi-factor model variants [manuscript submitted for publication]. *Department of Psychology, University of Mannheim.*

Petras, N. (2024b). Building hierarchically structured factor models with systematically selected residual correlations [manuscript submitted for publication]. *Department of Psychology, University of Mannheim.*

Petras, N., & Meiser, T. (2024). Problems of domain factors with small factor loadings in bi-factor models. *Multivariate Behavioral Research*, *59*(1), 123–147. https://doi.org/10.1080/00273171.2023.2228757

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer New York. https://doi.org/10.1007/978-0-387-89976-3

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358. https://doi.org/10.1037/0033-295X.107.2.358

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137. https://doi.org/10.1037/met0000045

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Scheel, A. M. (2022). Why most psychological research findings are not even wrong.

Infant and Child Development, *31*(1), e2295.

Soto, C. J., & John, O. P. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117. https://doi.org/10.1037/pspp0000096

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*(6), 1285–1303. https://doi.org/10.1177/2167702619855035

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2-3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128. https://doi.org/10.1007/BF02294531

Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2021). Using bifactor models to examine the predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational Research Methods*, *24*(3), 530–571. https://doi.org/10.1177/1094428120915522

Zhang, L., Pan, J., Dubé, L., & Ip, E. H. (2021). Blcfa: An r package for bayesian model modification in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 649–658. https://doi.org/10.1080/10705511.2020.1867862

Zhang, L., Pan, J., & Ip, E. H. (2021). Criteria for parameter identification in bayesian lasso methods for covariance analysis: Comparing rules for thresholding, p-value, and credible interval. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 941–950. https://doi.org/10.1080/10705511.2021.1945456

A Acknowledgements

"Through others we become ourselves."

(attributed to) Lev S. Vygotsky

Most of all I thank my parents Doris and Henning, and my larger family, for supporting me – and not least my academic life – from the very beginning. I wish all of you could have seen me complete this dissertation. A further special thanks goes to my supervisor Thorsten Meiser for his patient support, expert guidance, and inspiration.

Throughout my life I have been blessed with inspiring, tireless teachers and lecturers. I thank my school teachers, especially Mr. Muschick, Mr. Fritze, and Mr. Werner, and the many inspiring lecturers at Uni Konstanz, including Tobias Flaisch and Michael Dantlgraber, with whom I worked most closely. During my PhD years, I was lucky to be part of the graduate training program "Statistical Modeling in Psychology", which not only brought my knowledge on statistics to the next level, but also provided me with more inspiring colleagues and seminars than I could list here. I hope I could do the workshop on academic writing some justice. Thanks also to the SMiP-coach Silke, who had some success in keeping me sane.

I thank my co-workers at the psychological methods chair at Uni Mannheim, and all the other close colleagues, for their companionship and kindness. You make me feel at home. I thank the Open Science Meetup, the Reproducibilitea crews, and all our guests, for repeatedly fostering my hope that the science community keeps changing for the better. Last but definitely not least, I thank Fabienne, Sarah, Rebekka, and Juli, as well as many other friends and close colleagues, for their friendships that carried me through the last five years and more.

B Copies of Articles

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2023.2228757

Problems of Domain Factors with Small Factor Loadings in Bi-Factor Models

Nils Petras and Thorsten Meiser

University of Mannheim

ABSTRACT
Many measurement designs produce domain factors with small variances and factor loadings. The current study investigates the cause, prevalence, and problematic consequences of such domain factors. We collected a meta-analytic sample of empirical applications, conducted a simulation study on statistical power and estimation precision, and provide a reanalysis of an empirical example. The meta-analysis shows that about a quarter of all standardized domain factor loadings is in the range of $-.2 < \lambda < .2$ and about a third of all domains is measured by five or fewer indicators, resulting in small factor variances. The simulation study examines the associated difficulties concerning statistical power, trait recovery, irregular estimates, and estimation precision for a range of such realistic cases. The empirical example illustrates the challenge to develop measures that produce clearly interpretable domain factors. Study planning and interpretation need to take the (expected) sum of squared factor loadings per domain factor into account. This is relevant even if influences of domain factors are desired to be small, and equally applies to different model variants. We propose several strategies for how researchers may better unlock the bifactor model's full potential and clarify its interpretation.

KEYWORDS
bi-factor model; statistical power; specific factors; bi-factor(S-1) model

Introduction

Bi-factor models (Holzinger & Swineford, 1937) have become increasingly popular in psychological research over the past years (Reise, 2012; Zhang et al., 2021). One major reason is their ability to distinguish domain-specific variation in item responses from a general trait. Other than traditional models with a set of correlated factors, bi-factor models include an overall trait across different content domains, raters, tasks, or otherwise grouped indicators. This trait is of focal interest in many studies, for example as a general measure of quality of life (Chen et al., 2006), intelligence (Beaujean, 2015; Gignac & Watkins, 2013; Keith & Reynolds, 2018), or psychopathology ("p-factor," Caspi et al., 2014; Lahey et al., 2012; Patalay et al., 2015; Watts et al., 2019).

Domain factors capture additional, domain-specific variation. Critically, many common study designs entail weak domain factors (small factor variance). In the following, we consider domain factors to be "weak" to the degree that appropriate statistical tests for their detection have low power, they provide unreliable trait estimates, or their related estimates are small and therefore difficult to interpret. Weak domain factors are abundant in the literature. A review of articles from 2013 and 2014 found non-significant factor loadings or non-significant domain factor variances ("collapsing factors") in 47 of 82 articles (57%, Eid et al., 2017).

Whereas some studies merely account for domain-specific variation to obtain a "clean" measure of the general trait, others are concerned with the domain factors themselves. In validation studies, the presence of certain domain factors indicates a valid measurement design. Domain factor loadings indicate if indicators are valid exemplars of their assigned domain. In substantive research, the unique association of the general factor and the domain factors with third variables can be independently studied. In this way, structural equation models (SEM) can test increasingly differentiated theories within complex nomological nets (Eid et al., 2018; Zhang et al., 2021). Finally, practitioners may be interested in domain-specific individual scores (DeMars, 2013; Reise et al., 2013). The distinction between a general factor and domain

CONTACT Nils Petras nils.petras@uni-mannheim.de Department of Psychology, School of Social Sciences, University of Mannheim, L13, 15, 68161 Mannheim, Germany
Supplemental data for this article can be accessed online at https://doi.org/10.1080/00273171.2023.2228757.

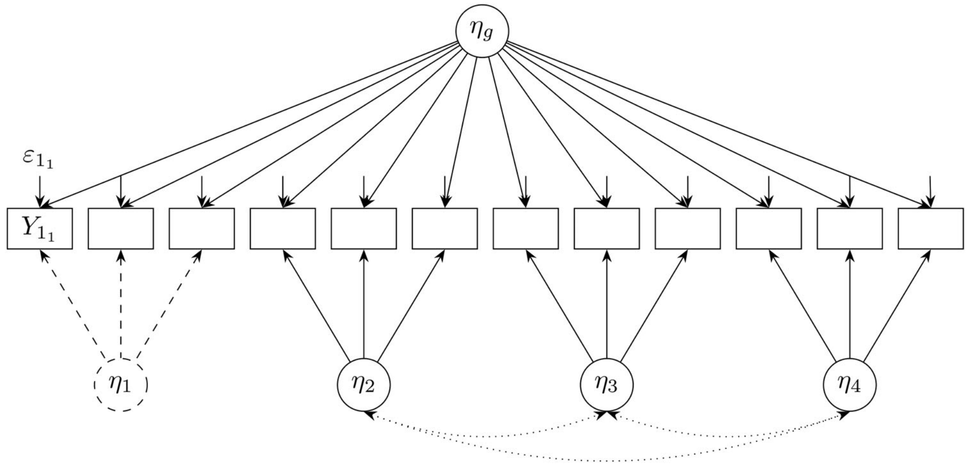


Figure 1. Bi-factor model path diagram with a general trait η_g and four domain traits (η_{1-4}, S model); only if some items exclusively load on the general factor (e.g. omitted dashed η_1, S-1 model), freely estimating correlations between domains is a reasonable option (dotted double-headed arrows, S-1c model).

factors offers a whole new perspective on psychological constructs and their relationships.

In the following section, we introduce the bi-factor model and its notation. After that follows an investigation of the causes, prevalence, and consequences of weak domain factors. The role of statistical power and the strength of domain factors in confirmatory bi-factor models has not yet been addressed in the literature. Although there are results on the recovery of loading matrixes in exploratory bi-factor analysis (Giordano & Waller, 2020), to our knowledge, the problem of weak domain factors has not been targetedly researched in the bi-factor EFA literature, either. Therefore, this study aims to assess which conditions are necessary to reliably detect and estimate domain factors and their loadings and compare these to real studies. It will be discussed how awareness of potentially weak domain factors when designing, choosing, or interpreting measures can drastically improve the utility of bi-factor model applications.

Bi-factor models

Bi-factor models use a general factor across all indicators and a set of domain factors for sets of related indicators (Figure 1). In the symmetrical model variant (S), every indicator loads on both a general factor η_g and one domain factor η_s.

The S bi-factor model of the response vector $\mathbf{Y}_i$ of case i is shown in Equation (1). $\mathbf{\Lambda}$ is the matrix of factor loadings and $\boldsymbol{\eta}_i$ the vector of latent trait values of case i. The error values in the vector $\boldsymbol{\varepsilon}_i$ are assumed to be independently and normally distributed for each indicator variable Y.

$$\mathbf{Y}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

Equation (2) shows the characteristic loading pattern of bi-factor models: all indicators load on the general factor η_g (first column of $\mathbf{\Lambda}$) and on one of the k domain factors (further columns). So λ_{s_js} is the loading of the j'th item of domain s on the domain factor η_s and λ_{s_jg} its loading on the general factor η_g.

$$\mathbf{Y}_i = \begin{pmatrix} \lambda_{1_1g} & \lambda_{1_11} & 0 & \cdots & 0 \\ \lambda_{1_2g} & \lambda_{1_21} & 0 & \cdots & 0 \\ \lambda_{1_3g} & \lambda_{1_31} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ \lambda_{2_1g} & 0 & \lambda_{2_12} & \cdots & 0 \\ \lambda_{2_2g} & 0 & \lambda_{2_22} & \cdots & 0 \\ \lambda_{2_3g} & 0 & \lambda_{2_32} & \cdots & 0 \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ \lambda_{k_mg} & 0 & 0 & \cdots & \lambda_{k_mk} \end{pmatrix} \begin{pmatrix} \eta_{ig} \\ \eta_{i1} \\ \eta_{i2} \\ \cdots \\ \eta_{ik} \end{pmatrix} + \boldsymbol{\varepsilon}_i \tag{2}$$

Since all factors of the model are orthogonal in the S variant, the variance-covariance matrix $\mathbf{\Phi}$ of its factors is a diagonal matrix. In the S-1 bi-factor model variant proposed by Eid et al. (2017), one domain factor is omitted (cf. Figure 1). The presence of the reference domain, whose indicators exclusively load on η_g, enables a proper variant of the bi-factor model in which the remaining domain factors may correlate freely (S-1c).[1] In the S-1 and S-1c models, η_g is

[1] For a discussion of problems regarding the estimation of correlated domain factors in the S model see Markon (2019). Conceptually, a full set of positively correlated domain factors (= correlations between all indicators) and the general factor are to some degree redundant, leading to problems in both estimation and interpretation.

interpreted as the common trait as assessed with the reference domain. In the terminology of classical test theory (CTT, Novick, 1966), in S-1 models the general trait combines the common true score of all domains and the true score specific to the reference domain (Eid et al., 2017). Compared to the S model, the S-1 and S-1c models therefore provide improved clarity in the interpretation of η_g if domains are not randomly sampled. If the domains are a meaningful selection, as in most multifaceted psychological measures, "defining the latent variables of the [S bi-factor and second-order] models […] as random variables on a well explicated set of possible outcomes" (Eid et al., 2017, p. 548) could not be achieved.

To examine problems with weak domain factors, a measure of their strength is needed. In the following, we use the sum of squared loadings SS_λ in the fully standardized bi-factor model with indicator and trait variances equal to one (Equation (3)).

$$SS_\lambda(\eta_s) = \sum_{j=1}^{m} \lambda_{s_j s}^2 \tag{3}$$

This quantity measures the total share of indicator variance of the factor. $SS_\lambda = 1$ means that the factor explains a total indicator variance equal to the variance of one indicator.[2]

To better understand the influence of each factor, the variance of each indicator Y_{s_j} can be decomposed into three components: consistency, specificity, and error. Note that the simplified Equation (4) assumes a fully standardized model. The first term $\lambda_{s_j g}^2$ is the consistency of the indicator: the proportion of variance due to the general trait η_g. The second term $\lambda_{s_j s}^2$ is the specificity of the indicator: the proportion of variance due to the domain-specific trait η_s. The remaining error is assumed to be independently, randomly, and normally distributed with a variance of $\sigma_{\varepsilon s_j}^2$. The reliability ($Rel$) of an indicator is the proportion of its variance that is explained by the latent variables.

$$Rel(Y_{s_j}) = \lambda_{s_j g}^2 + \lambda_{s_j s}^2 = 1 - \sigma_{\varepsilon s_j}^2 \tag{4}$$

Weak and "anomalous" domain factors

Weak and anomalous domain factors are a consequence of the structure of bi-factor models and the typical construction process of psychological measures. There are several reasons why weak domain factors—desired or not—should be expected in practical applications:

First, the measurement of domain factors and the general factor compete for each indicator; indicator reliability is split into consistency and specificity (Equation (4)). Standardized factor loadings for both are typically lower compared to models with indicators relating to only one factor each. In indicators with high consistency, the ratio of domain factor variance to error variance can be small, even though the reliability is high. It is a frequent intention to use reliable total scores as the main criterion when applying measures in practice (e.g., conscientiousness and general intelligence—rather than their facets—in personnel selection). Even if another purpose of a measure is to discern different parts of a construct (e.g., different facets of a personality trait or different aspects of intelligence), a likely concern is that the indicator still measures the overall construct (e.g., the personality trait or general intelligence). A key challenge is that factor loadings in correlated-factor models confound the relationship of indicators to general variance (shared among all domains) and domain-specific variance. Therefore, indicators without domain-specific variance are not automatically disqualified. The goal conflict between measuring a general trait and domain-specific variance may be more or less likely to occur and more or less easy to solve depending on the nature of the construct and the other desired properties of the measure.

Second, each domain consists of a fraction of the indicators of the overall measure. Measures based on a correlated-factor model had their number of indicators chosen based on stronger factors, which include a substantial portion of the general trait from the bi-factor model. Factors in the correlated-factor model contribute to the general factor in the bi-factor model to the extent of their intercorrelation. The leftover domain-specific variance can be tiny. Especially problematic are short measures which were reduced to a barely acceptable length. They may measure a general trait or a set of correlated factors efficiently (shortened as much as possible without their reliability falling below a target value) but fail to produce reliable domain-specific factors in bi-factor models. As we will show in more detail below, if researchers choose the desired length of a measure without explicitly considering the consequences for domain factor measurement, they are in danger of choosing too few indicators to properly recover them.

[2]In Exploratory Factor Analysis (EFA) and Principal Component Analysis (PCA), the eigenvalues of the covariance matrix are used as a decision criterion for the number of factors to include. In PCA, these are the SS_λ values of the unrotated components and in EFA this relationship holds approximately. Therefore, the effective inclusion criterion is usually near $SS_\lambda = 1$.

For these reasons, one should expect a substantial portion of domain factors to have few and small factor loadings ($\lambda_s < .2$) and therefore little variance—even before considering the substantive research context. Given that the surge in popularity of the bi-factor model (Reise, 2012; Zhang et al., 2021) is in large parts based on reanalyzes of older measures, this disconnect of the listed particularities of the bi-factor model from the development process of the measures should be expected to lead to weak domain factors and small domain-factor loadings. Whereas some research areas may welcome such outcomes—potentially, because they adequately reflect the trait of interest—we argue that obtaining weak domain factors should not be an accident. Researchers should be aware of this issue before conducting their research.

Indeed, Eid et al. (2017) showed an abundance of problematic empirical examples. Not only were there many domain factor loadings that did not significantly differ from zero. Multiple domain factors "collapsed" entirely, showing non-significant variance estimates or a set of non-significant factor loadings. Some extreme cases had negative factor variance estimates.[3] This led many researchers to question or modify their application of the bi-factor model (see also Watts et al., 2019) and Eid et al. (2017) to speak of "anomalous results". The prevalence of studies with at least one anomaly was 61% in their sample of articles that used a bi-factor model and were published in 2013 or 2014. This number might have been even higher if there were unpublished studies or researchers quietly switched to another model.

Problematic results were one reason why Eid et al. (2017) questioned the use of the symmetrical bi-factor model (S). They criticized its use in cases where domains are specifically selected (single-level sampling structure) as opposed to randomly sampled (two-level sampling structure). They base their argument on Stochastic Measurement Theory (SMT, Steyer, 1989):

> From the perspective of SMT, the latent variables in traditional bifactor and related G-factor models cannot be defined as random variables on a well explicated random experiment when only a single-level sampling design is considered. […] From the scope of SMT many of the anomalous results encountered in empirical applications in fact have to be expected when domains are not randomly selected or when they cannot be considered interchangeable. (Eid et al., 2017, p. 555)

They consequently introduced the S-1 and S-1c variants[4] as sound alternatives from the perspective of SMT (Eid et al., 2017, p. 550ff). They did not discuss the effect of small domain strength, insufficient statistical power, or the rate at which anomalous results occur in S-1 models. Because they classified all non-significant estimates of factor loadings and factor variances as "anomalous" results due to badly specified models, we consider the current work a crucial extension to their work, because it inquires into alternative explanations. If "anomalous" results are equally frequent in S and S-1 models, the consideration of the sampling structure would be irrelevant to problems with weak domain factors.

Statistical power, effect size, and estimation precision

In the context of our simulation study, we consider domain factors to be weak if they cause a problem: a) if their associated null hypothesis cannot be rejected (the model without the domain factor fits the data equally well, given a finite, reasonable sample size) or b) if they produce (comparatively) unreliable trait estimates, meaning that the trait recovery (R^2) is half as good as for the general factor (or worse). One purpose of the simulation study is to provide a range of benchmark values for applied researchers to compare empirical results to. To understand the surprisingly high prevalence of null results in the literature, statistical power needs to be taken into account. For power analysis, the size of the effect needs to be known: how large are estimates of domain factor loadings and domain factor strengths in empirical applications? Moreover, for many applications, it is not enough to show that certain parameters in the model differ significantly from 0. Sufficient model parameter estimation precision and trait recovery precision are crucial for interpretation. Especially studies that use domain factors to predict other variables or use domain-specific scores rely on unbiased trait estimates and sufficient precision.

The presence of domain-specific variance may be a mere nuisance to the measure of the general factor for some purposes or areas of research. In that sense, weak or completely absent domain factors are desirable, as long as they do not produce irregular estimates. The corresponding ideal case is a model with a single general factor explaining all systematic variance of the indicators. This is especially true for

[3] Setting the factor variance instead of the first loading to 1 for model identification would prevent that, but most likely shift the problem to other parameters. Therefore, we considered this to be a problematic phenomenon.

[4] The S*I-1 variant is not discussed here.

applications that assign specific factors to different raters or alternative methods of measurement (e.g., Frey et al., 2017; Scholz et al., 2022). These factors do not necessarily have a useful substantive meaning. Instead, they are influences that should be controlled for. In such scenarios, researchers may want to avoid strong domain factors. Nevertheless, judging their strength and impact may be the focal point of a study. A research question could be if two measures (or two types of raters) can be treated as interchangeable or if biases are introduced by choosing one over the other. For this purpose the ability to judge the statistical power to detect undesired domain-specific influences and the precision with which they are captured by the model is relevant.

The current study

To identify the necessary conditions to reliably detect and properly estimate domain factors and their loadings, we conducted a simulation study. We compare its results to the conditions in a meta-analytic sample of empirical applications. The meta-analysis uses the reported factor loading matrixes of the studies listed by Eid et al. (2017). It tests our arguments on why weak domain factors should be expected in practice: How large are domain factor loadings and general factor loadings typically? How many indicators per domain are used? How prevalent are reliable indicators with low specificity ($\lambda_g > .5$ and $\lambda_s < .2$)? Do null results happen in small samples ($n \leq 300$) only?

In the simulation study, the measurement design was varied to answer the following questions: What is the strength of a detectable domain factor under realistic conditions? Which measurement designs provide a relatively adequate recovery of domain trait scores? What are the core influences on the precision of domain factor loading estimates? Under which conditions occur unacceptable "anomalous" results (negative domain factor variance estimates, non-convergence)? Can the newly proposed model variants (S-1 or S-1c) reduce the number of irregular results or null results?

After presenting the meta-analysis and the simulation study, we finally reuse open data to provide an empirical example to facilitate the discussion. The following discussion combines the meta-analysis results and simulation results to examine the origins and consequences of the outlined practical challenges. We propose several steps to maximize the utility of bi-factor applications and outline limitations.

Meta-analysis

Methods

For the analysis of factor loadings and SS_λ of domain factors in the literature, we chose to adopt the list of empirical examples in Eid et al. (2017) to enable comparison with their work. These studies were originally sampled from PsycInfo using the terms "bifactor" and "bi-factor" (all fields), and include publications from 2013 or 2014. They were coded to contain either a non-significant domain factor variance estimate (Eid et al., 2017, Table 1) or a non-significant domain factor loading estimate (Eid et al., 2017, Table 2). We searched the 47 articles for S bi-factor loading matrixes ($\boldsymbol{\Lambda}$). Only one set of estimates per sample was included to not bias the overall result by repetition. Two articles reported two bi-factor studies on unique samples, which were both included. 21 articles were excluded from subsequent analysis: incomplete report of estimates (1), IRT model (1), exploratory model (1), free estimation of domain factor correlations (5), no consideration of S model variant (4), exclusive report of adapted models (7), outlier[5] (1). We reconstructed one unreported S bi-factor model based on the reported correlation matrix.[6] Reversely keyed indicators and domain factors were recoded for the current analysis so that all factor loadings are expected to be positive. An indicator or domain was considered reversely keyed if the factor loadings were expected to be negative based on the study design and theory.

28 models from 26 articles were included in the final sample (a reference list can be found in the Appendix). Two were coded by Eid et al. (2017) as including a non-significant domain factor variance estimate. The other 26 were coded as including (at least one) non-significant domain factor loading estimate. The sample of models includes 3 ability tests, 21 self-report scales, and 4 other-report scales. Table 1 shows the large variety of constructs encountered in the sampled articles (see also Eid et al., 2017 Tables 1 + 2). We sorted the constructs into three broad categories: *Clinical/health* constructs include mental and physical health related outcomes and behaviors. *Personality* constructs include non-clinical, relatively stable interindividual differences. *Education* constructs are specific to the education context. Of the 28 models in our analysis, 18 dealt with *clinical/health* constructs,

[5] 8 of 15 indicator reliabilities exceeded 0.948, model fit was almost perfect (TLI = 1.00, CFI = 1.00, RMSEA = 0.010), despite the diverse indicator content (Blanco et al., 2014), Table 3).

[6] Another one had to be omitted due to irregular estimates. The error variance of an indicator variable was estimated to be impossibly large and negative, leading to uninterpretable results.

Table 1. Constructs in the meta-analysis sample.

area	construct
clinical / health	cognitive abilities (a), depression, ADHD / ODD (2, o), anxiety disorder, risk of developing a mental disorder (o), depression / anxiety / stress, ADHD, loneliness, emotional distress, anxiety / depression, burnout, sun protection behavior, fatigue, medically unexplained symptoms, seasonal depression, health
personality	anxiety (2), callous-unemotional traits, dark triad, susceptibility to emotional contagion, disgust sensitivity, ethnic identity
education	EFL listening proficiency (a), responsive teaching (o), academic skills (a), teacher self-efficacy in inclusive classrooms

Note. Numbers indicate frequencies; other codes: (o) = other-report; (a) = ability test; unmarked = self-report;" /" indicates the combination of multiple constructs in the same model without a superordinate term; ADHD = attention deficit hyperactivity disorder; ODD = oppositional defiant disorder; EFL = english as a foreign language.

Table 2. Simulation design.

parameter	values
n	200, 300, 500, 1000, 2000
λ_g	.5, .7
λ_s	.2, .3, .4, .5, .6
m	3, 6
model variant	S, S-1, S-1c

Note. Fully crossed design with $5 \times 2 \times 5 \times 2 \times 3 = 300$ conditions. n = sample size, m = number of indicators per domain.

6 with *personality* constructs, and 4 with *education* constructs. A full table linking articles to constructs can be found on the osf page of this article.

Results

Figure 2 shows the combined distribution of factor loadings on the general factor (λ_g) and the domain factor (λ_s) for each indicator variable.[7] Indicator reliabilities ($\text{Rel} = \lambda_g^2 + \lambda_s^2 = 1 - \sigma_\varepsilon^2$) show a large variability ($M = 0.54$, SD $= 0.19$). This may reflect differences in the breadth of constructs as well as differences in the quality of the selected indicators. The sizes of λ_s and λ_g are limited by each other: $\lambda_s \leq \sqrt{1 - \lambda_g^2}$ and $\lambda_g \leq \sqrt{1 - \lambda_s^2}$. But the impact of this negative dependency is counteracted by variation in the indicator reliability: Low values of λ_s and λ_g coincide in indicators with a large variance of the measurement error. The resulting correlation between the factor loadings is $r_{\lambda_g \lambda_s} = -0.35$ ($t = -9.02$, $p < 0.001$, 95% CI [-0.42, -0.28]). This suggests competition in the measurement of the traits. For each indicator with $\lambda_s > \lambda_g$, there are 4.30 indicators with $\lambda_s < \lambda_g$. 24.79% of all indicators have a very small domain factor loading ($-.2 < \lambda_s < .2$), but also a reasonably high factor loading on the general factor ($\lambda_g > .5$). This likely reflects indicator selection procedures that focus on the measurement of the general trait or maximize the internal consistency of the whole measure. 17 of 28 models include at least one negative factor loading estimate. Note that negatively keyed factors and indicators were recoded before plotting, so these are unexpected results. Figure 3 shows the number of indicators per domain. 31.25% of all domains were measured by 5 or less indicators.

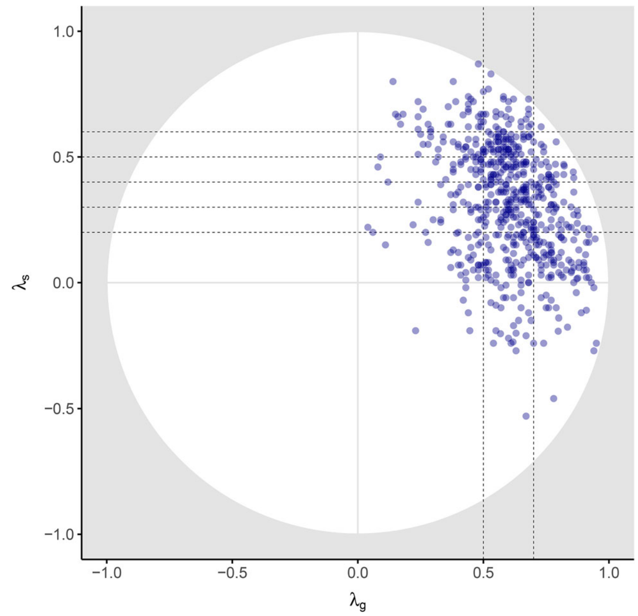


Figure 2. Fully standardized factor loadings of individual indicator variables from 28 S bi-factor models; dashed lines indicate simulation conditions.

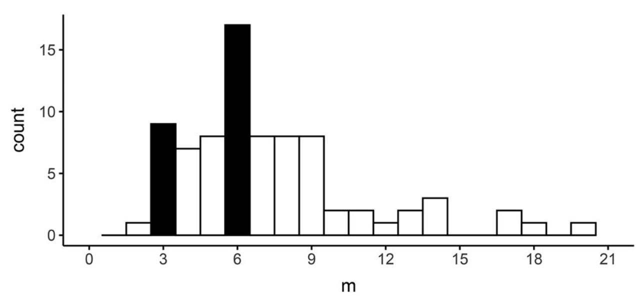


Figure 3. m = number of indicators per domain from 28 S bi-factor models; filled bars mark simulation conditions; for some indicators it was unclear if their loadings were fixed or estimated at 0.00.

What is the resulting strength of the domain factors? Figure 4 shows the combined distribution of domain

[7]One indicator was excluded from analysis due to an impossible combination of reported standardized factor loadings ($\lambda_s = 0.98$, $\lambda_g = 0.59$).

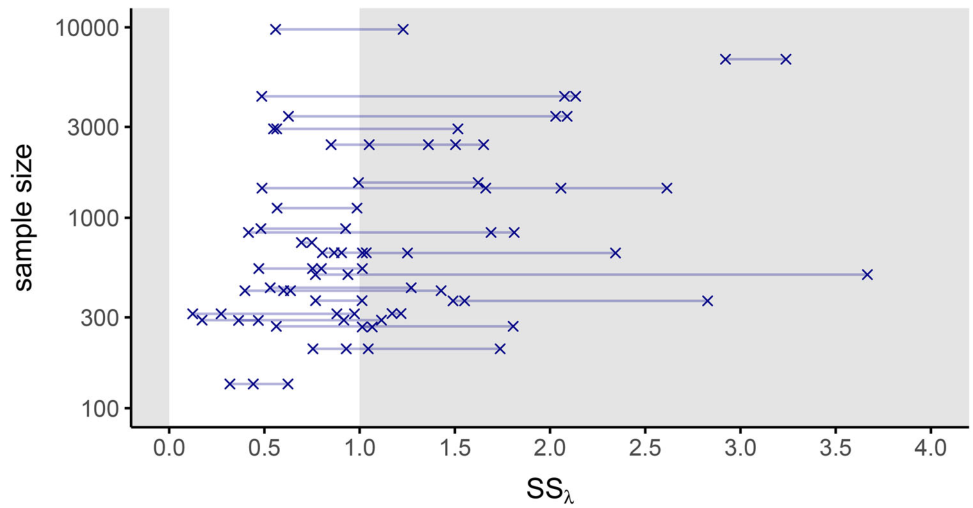


Figure 4. Sum of squared loadings and sample sizes of domain factors from 28 S bi-factor models; domain factors of the same model are connected by a line.

factor SS_λ and sample sizes. 52.50% of factors have $SS_\lambda < 1$, and 16.25% have $SS_\lambda < 0.5$. 26 of 28 models were included because of a non-significant domain factor loading, but most of them show at least one whole weak domain factor (if judged by $SS_\lambda < 1$, more detailed discussion below). Weak domain factors could be the product of noise in small samples even if the true underlying factor is strong in the population. Figure 4 shows that a lack of power due to insufficient sample size alone cannot explain weak domain factors: they occur across all sample sizes. In conclusion, the presence of at least one weak domain factor ($SS_\lambda < 1$) is the norm in the sampled bi-factor models, not the exception.

Simulation study

Methods

In the simulation study, random data for bi-factor models of the three model variants were generated. In S-1 and S-1c models, the first of four domain factors was omitted. For S-1c data generation the correlation between the second and the third domain was set to $r_{23} = .5$ and all other correlations were set to zero. Conditions relevant to statistical power and estimation precision were systematically varied (Table 2).[8]

To vary the strength of the domain factors, the factor loading size λ_s and the number of indicators per domain were varied. Factor loadings were held constant across all indicators and invariant during data generation, which greatly simplifies interpretation. We only included domain factor loadings that are positive and at least $\lambda_s = .2$, so it can be checked if sampling variation of truly admissible values explains the occurence of negative or zero factor loadings in practice (Figure 2). For both the sample size and λ_s, realistic values and values in a problematic range were included (down to $n = 200$ and $\lambda_s = 0.2$). The domain factor loadings lie in a range that was frequently observed in the reviewed empirical example studies ($.2 \leq \lambda_s \leq .6$, dashed lines in Figure 2). Given these fixed values for λ_s, the reliability of the indicators was varied using two different values for λ_g. This design produces reliabilities between 0.29 and 0.85 across conditions. All factor loadings are fully standardized because random error variance was added to all indicators to reach $\sigma_Y^2 = 1$ and traits were sampled with a variance of one. Since S-1 and S-1c models have no variance attributed to the first domain factor, and λ_g was held constant, they have a higher proportion of error variance on indicators of the first domain. Only continuous data with multivariately normally distributed trait values and error terms were considered. Although contamination with other types of errors is frequent in practice (Micceri, 1989), and the true distribution of latent traits is debatable, normally distributed traits and errors are prototypical for this model class and frequently assumed in practice. The fully crossed design resulted in 300 simulation conditions with 1008 replications per condition.

[8]The correlation between domain factors in the S-1c model also affects the statistical power and estimation precision (Yuan et al., 2010), but was not varied beyond the distinction between S-1 and S-1c models. Higher correlations were shown to lead to both increases and decreases in standard errors for both loadings and factor variances in correlated-factor models depending on the other model parameters (Yuan et al., 2010, Table 3). It is unclear if such differences are substantial in bi-factor models and how they would proliferate to other factors in the model. As seen below, the difference in statistical power between the S-1 and S-1c model, which is essentially a large variation of a domain factor correlation (0 vs. 0.5), proved to be relatively inconsistent and unimportant in comparison to other factors.

The simulation study was conducted using the software `R` (Version 4.0.2 and 4.0.3, R Core Team, 2020) and the package `SimDesign` (Version 2.0.1, Chalmers & Adkins, 2020). `mvtnorm` (Version 1.1-1, Genz et al., 2020) was used to randomly sample trait and error values from a multivariate normal distribution. Models were estimated using `lavaan` (Version 0.6-7, Rosseel, 2012).

For each sample dataset, all model variants were estimated using maximum likelihood (ML) estimation with the default settings of `lavaan` (Version 0.6-7, Rosseel, 2012). The fixation of the first factor loading to one for identification made negative estimates of the domain factor variance possible. In S-1c models, all correlations between domain factors were freely estimated. This results in a fully crossed design regarding data-generating model variant and estimation model variant. To analyze anomalous results, improper solutions (e.g. negative variance estimate) were retained. In the following, converged solutions are those, for which lavaan indicated convergence and standard errors of estimates were obtained. If not specified otherwise, the presented results refer to correctly specified models only, meaning the data-generating model and the estimated model variant are the same. Results on domains are presented as a summary (mean) for domains two, three, and four, even for the S-1c models. The distinction between the uncorrelated fourth domain and the other domains in the S-1c model did not prove relevant in any of the analyses.

The statistical power to detect domain factors was measured in three different ways. First, the proportion of significant variance estimates of the domain factor was calculated based on the Wald-Test against zero with $\alpha = .05$. This test corresponds to the "anomalous results" in Eid et al. (2017) (non-significant domain factor variance estimates). Results for this test are part of the default summary output of `lavaan`. Note, that this is a test against the boundary of the parameter space ($H_0 : Var(\eta_s) = 0$). For this reason, the distributional assumption is violated and results are conservatively biased (Molenberghs & Verbeke, 2007; Stoel et al., 2006). The uncorrected version is used to represent what plausibly was the general practice in the sample of studies above. Second, the proportion of significant likelihood-ratio-tests (LRT) comparing the model with and the model without the first domain factor was calculated. The LRT tests the difference in model misfit $\Delta\chi^2 \sim \chi^2(df_{ModH_0} - df_{ModH_1})$ between the correctly specified model H_1 (which includes the domain factor's variance and its loadings) and the incorrectly specified model H_0 (which by omitting the domain factor essentially fixes the latent variances and all related factor loadings at 0 and is therefore nested within the first model) against 0. This is a more adequate test to decide if the domain in question should be part of the model. The LRT is based on all estimated parameters related to the domain in question, whereas the Wald-Test is based solely on the variance estimate. Therefore, differences in the results can be expected. Note that non-converged models were counted as false negatives, so the reported values for statistical power can never exceed the convergence rate. Omitting non-converged cases would be biased in conditions with a low convergence rate. Furthermore, researchers planning a study are likely most interested in the probability of a successful study than in the conditional probability given convergence. Third, the theoretical power of this LRT was computed for all simulation conditions, testing the correctly specified model variant against the same model without the domain factor in question. The model without the domain factor was fit to the theoretical variance-covariance matrix under the true model with the domain factor present. The misfit between the resulting model implied variance-covariance matrix and the true variance-covariance matrix was then used to compute the statistical power of the LRT using the semPower R package (Moshagen (2021)).

For individual indicators, the average number of significant indicators per domain was calculated under each condition. Significance was judged based on the Wald-Test of the factor loadings with $\alpha = .05$.

To assess the quality of estimated trait values the squared correlations between the true and the estimated trait values (R^2) were calculated. This is the proportion of variance of the estimated trait values that is determined by the true trait. Trait values were estimated using regression factor scores (DiStefano et al., 2009), as implemented in lavaan (Version 0.6-7, Rosseel, 2012). To assess the precision of factor loading estimates, root mean square errors (RMSEs) were calculated for each repetition. They were computed based on the differences between the estimates and the true factor loadings in the population, given by the simulation condition.

To complete the list of potential "anomalous" results discussed by Eid et al. (2017), the proportion of cases with at least one negative domain factor variance estimate was computed. The simulation did not replace (or tweak the estimation of) cases that did not converge. Instead, convergence rates are analyzed below.

To assess the importance of the simulation conditions (Table 2) for each outcome, we estimate general linear[9] models. Because these models merely serve to indicate the relevance of the conditions, we use a simple baseline model without interactions. For the parameters with multiple conditions on a metric scale (n and λ_s), we also include a quadratic term to allow for non-linear effects. To assess the importance of a given parameter, we compare this baseline model (Equation (5)) to the model without the term(s) relating to this parameter. In Equation (5), *outcome* refers to all the individually analyzed outcomes (statistical power, estimation precision, …) and *variant* is is a dummy-coded factor with three levels. For brevity, we only report the p-value of the F-Test for model comparison, as well as the difference in adjusted R^2. We describe any predictor with a $\Delta R^2 < .01$ (equivalent to $r < 0.1$) as irrelevant, regardless of its statistical significance.

$$Outcome = variant + n + n^2 + m + \lambda_g + \lambda_s + \lambda_s^2 + \varepsilon \tag{5}$$

Results

Domain factor detection

In general, the power of the LRT tends to exceed that of the Wald-test (McCulloch & Searle, 2004, p. 150). In the current simulation results, the LRT for model comparison consistently shows superior statistical power to the Wald-Test of the factor variance. There is no condition with a meaningful advantage of the Wald-test. A substantial advantage of the LRT shows under many conditions: Under conditions where at least one test has a power estimate below 1 (not all replications significant) the mean difference in statistical power is 0.26 in favor of the LRT (additional figure in supplementary materials).

Figure 5 presents an overview of the power of the LRT depending on SS_λ and sample size. The simulated values (including non-converged cases as false negatives) are connected *via* vertical lines with the theoretical values. The model variant is irrelevant to the statistical power of the LRT to detect domain factors ($p = 0.62$, $\Delta R^2 = 0.00$). Sample size ($p = 0.00$, $\Delta R^2 = 0.10$), number of indicators per domain ($p = 0.00$, $\Delta R^2 = 0.05$), loading on the general factor ($p = 0.00$, $\Delta R^2 = 0.02$), and the size of the domain factor loadings ($p = 0.00$, $\Delta R^2 = 0.52$), all contribute uniquely to the prediction of statistical power. Consider a domain factor with $SS_\lambda = 0.75$, based on three standardized domain factor loadings of 0.5: The LRT easily detects the presence of the domain, even in samples of $n = 200$ ($1 - \beta = 0.98$). For smaller effects, there is a steep drop in statistical power. Judging by the relationship between SS_λ and the statistical power (Figure 5), adding a single indicator with $\lambda_s \geq .4$ ($\Delta_{SS_\lambda} \geq 0.16$) can improve power drastically. Realistic variations in the reliability of indicators beyond their loading on the domain factor ($\lambda_g = .5$ (circles) vs. $\lambda_g = .7$ (triangles)) result in large differences in statistical power (up to $\Delta_{1-\beta} = 0.46$). The blindness of the theoretical analysis to non-convergence is a major cause for the difference between the theoretical and simulated power under challenging conditions. Table 3 compares the cumulative results of the LRT by model variant for correctly specified models. Across conditions, the model variant barely influences convergence or power, S-1 models converge slightly more often. This explains a slight increase in the proportion of non-significant results because convergence is most often an issue in low power conditions. In case of misspecification there are much larger differences (see section on convergence).

Figure 6 presents an overview of the statistical power of the test of domain factor loadings. The model variant is irrelevant to the statistical power of the test of domain factor loadings ($p = 0.41$, $\Delta R^2 = 0.00$). Sample size ($p = 0.00$, $\Delta R^2 = 0.08$), number of indicators per domain ($p = 0.00$, $\Delta R^2 = 0.04$), loading on the general factor ($p = 0.00$, $\Delta R^2 = 0.02$), and the size of the domain factor loadings themselves ($p = 0.00$, $\Delta R^2 = 0.56$), all contribute uniquely to the prediction of statistical power. The more indicators a domain factor has, and the less error variance its indicators have (higher λ_g), the more precisely its loadings are estimated (for an analytical approach, see Yuan et al. (2010)). Under favorable circumstances ($\lambda_g = .7$, $m = 6$) a sample size of $n = 300$ is more than sufficient for $\lambda_s = .3$ (in the population) to be detected with high power ($1 - \beta = 0.99$). The power is much higher compared to realistic, but much less favorable conditions ($\lambda_g = .5$, $m = 3$, $1 - \beta = 0.51$). To compensate for this, the sample size would have to be increased to $n > 1000$ ($1 - \beta > 0.95$).

[9]For outcomes on a scale of 0 to 1, we considered linear models to be sufficient, because they detect the presence of monotonous effects, and their easily interpretable determination coefficient is able to roughly order them by importance. Binomial regression would not have offered an easy to interpret determination coefficient and a logit transform would have led to many infinity values due to observed relative frequencies of exactly 1.

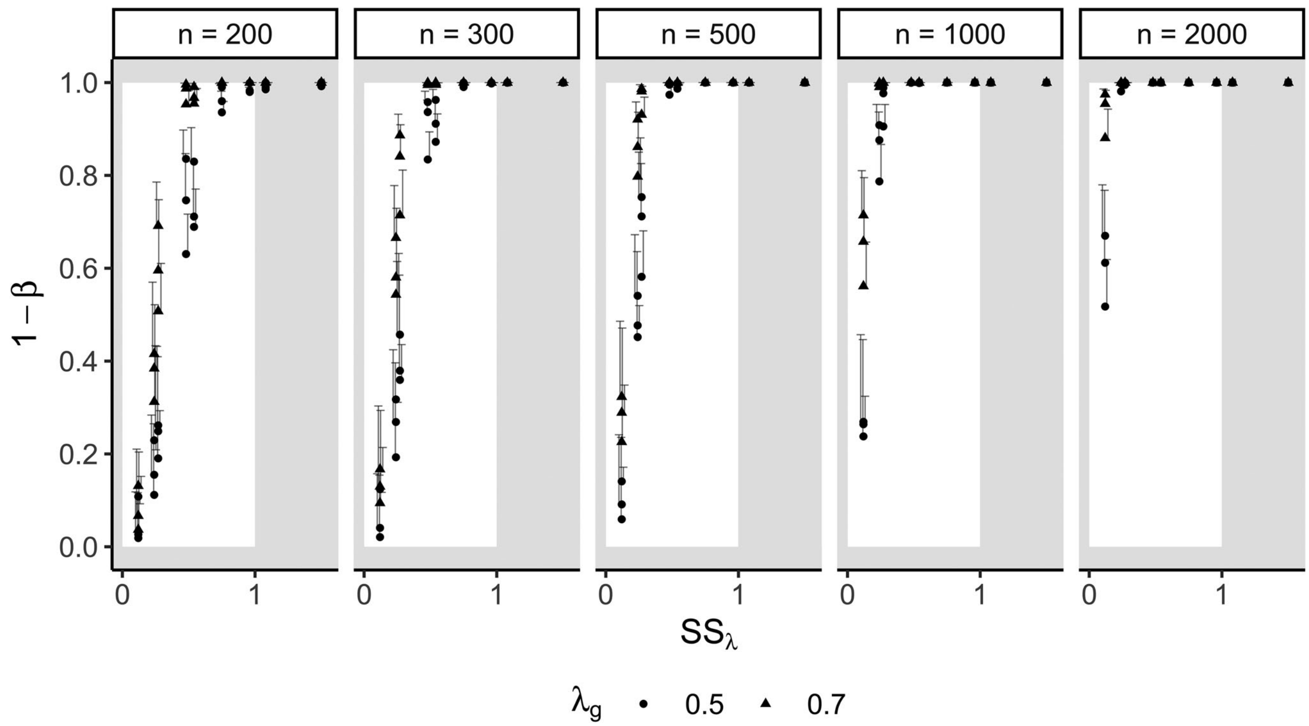


Figure 5. Power to detect domain factors by Likelihood-Ratio-Test. Only correctly specified models are shown. Each symbol represents one simulation condition. Vertical lines show the discrepancy between simulated power (symbol) and theoretical power (arrow tail; small horizontal offset for readability).

Table 3. Likelihood Ratio Test outcomes (percent) by model variant.

	significant	not significant	not converged
S-1	87.76	4.30	7.93
S-1c	85.72	4.05	10.23
S	86.13	4.04	9.83

Note. Correctly specified models only.

Parameter recovery

The distribution of the RMSE of domain factor loading estimates is heavily skewed and includes outliers from irregular estimates. Therefore, Figure 7 shows the median of the RMSE distribution across replications. Note, that for a small proportion of replications, the RMSE was substantially higher.[10] The model variant is irrelevant to median estimation precision of domain factor loadings ($p = 0.38$, $\Delta R^2 = 0.00$). Sample size ($p = 0.00$, $\Delta R^2 = 0.16$), number of indicators per domain ($p = 0.00$, $\Delta R^2 = 0.05$), loading on the general factor ($p = 0.00$, $\Delta R^2 = 0.05$), and the size of the domain factor loading itself ($p = 0.00$, $\Delta R^2 = 0.34$), all contribute uniquely to the prediction of the estimation precision. Domain factor loadings that are relatively small in the population are estimated with less precision than larger ones.[11] Higher overall indicator reliability (higher λ_g) and more indicators per domain increase precision. The problem case that a domain factor loading is truly substantial but estimated near zero can only be expected under a combination of multiple adverse conditions. For example: Assuming a normal distribution of estimates and $RMSE = 0.05$ (dashed line), only 2.28% of $\lambda_s = .3$ are estimated at 0.2 or lower. Only very few near-zero loadings can be explained by estimation uncertainty (cf. Figure 2). This could also be understood from the estimated standard errors and confidence intervals of the loading estimates in empirical studies reporting negative or near-zero estimates.

Figure 8 shows that domain trait recovery barely improves with increased sample size and improves much slower with increased effect size than statistical power. The model variant ($p = 0.00$, $\Delta R^2 = 0.00$) and sample size ($p = 0.00$, $\Delta R^2 = 0.01$), are irrelevant to domain trait recovery. The number of indicators per domain ($p = 0.00$, $\Delta R^2 = 0.07$), loading on the general factor ($p = 0.00$, $\Delta R^2 = 0.04$), and the size of the

[10]The same plot, but with 0.95 quantiles (instead of medians) of the RMSE distributions is included in the supplementary materials.

[11]The same absolute difference on the scale of λ_s is larger on the scale of λ_s^2 (indicator variance) for larger values of λ_s. The truth of this claim, therefore, depends on the scale.

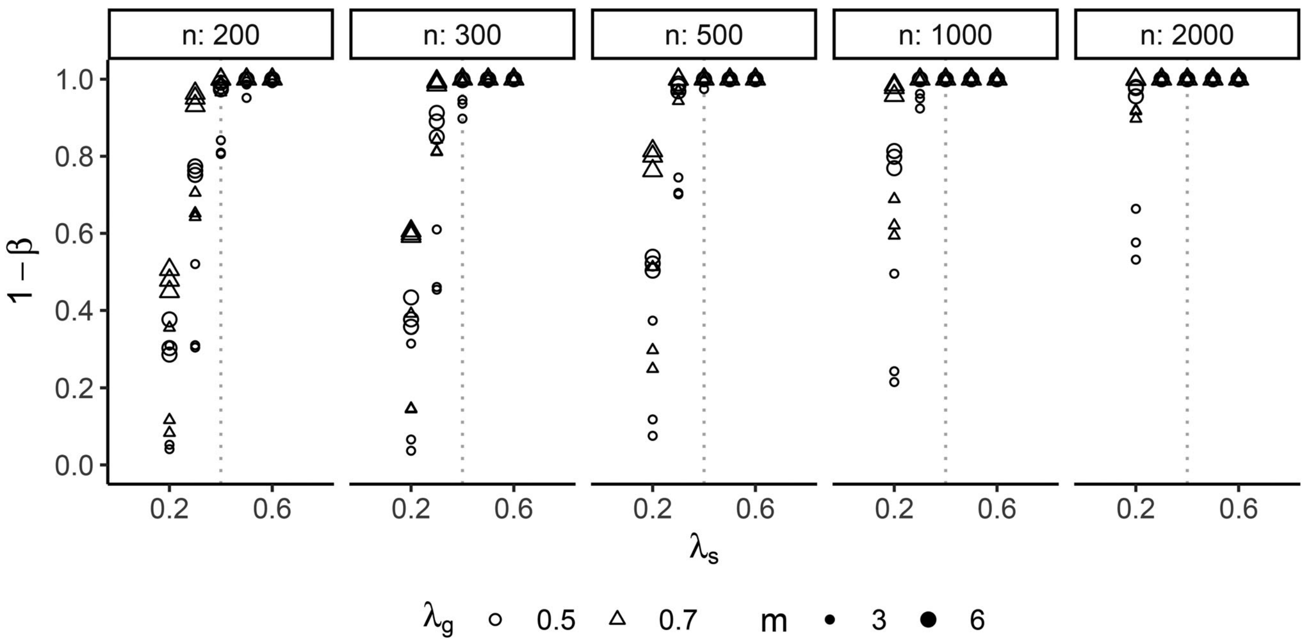


Figure 6. Power to detect domain factor loadings by Wald-Test. Each symbol represents one simulation condition. m = number of indicators.

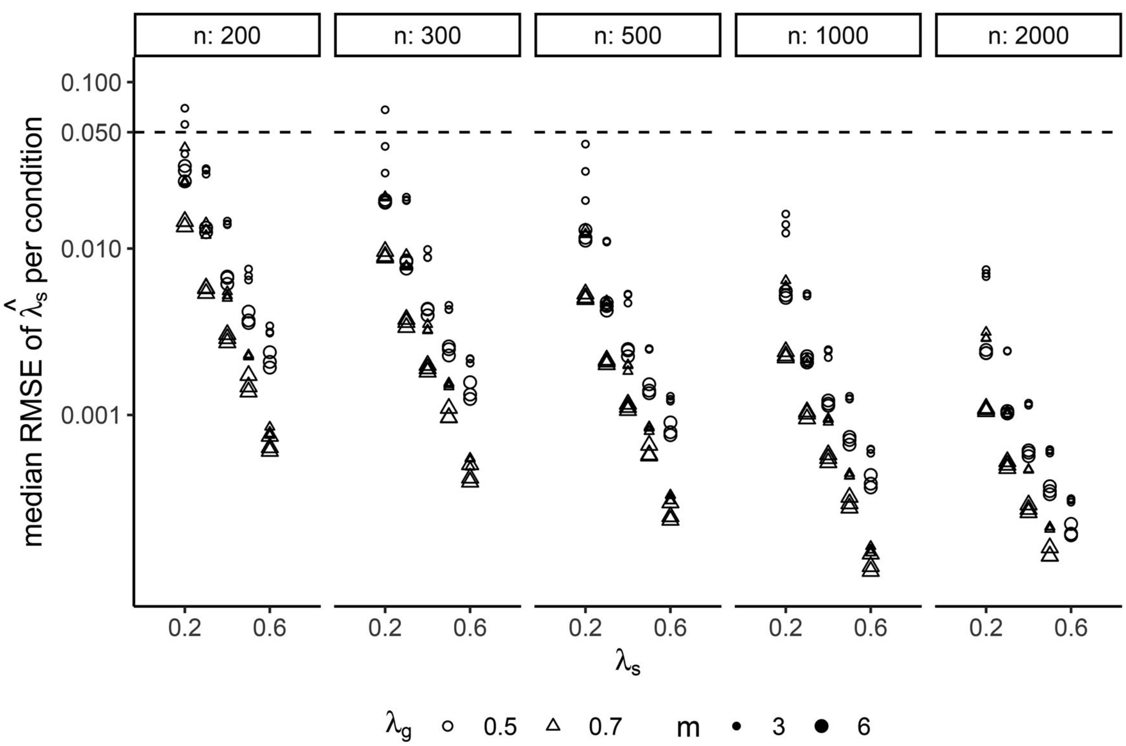


Figure 7. Estimation Precision of domain factor loadings. Each symbol represents one simulation condition. The logarithmic y-axis scale is cut at 0.0001. Only correctly specified models are displayed. m = number of indicators.

domain factor loadings ($p = 0.00$, $\Delta R^2 = 0.88$), all contribute uniquely to the prediction of trait recovery. At $SS_\lambda = 1$, even in large samples only about 50-70% of the variance of the factor score is determined by the true trait. Below $SS_\lambda = 1$, this value quickly declines even further, falling below half the typical value of the general trait ($\approx .7$ to 0.95, see below).

Figure 9 shows the influence of the domain traits on the recovery of η_g. The sample size ($p = 0.00$, $\Delta R^2 = 0.00$), is irrelevant to general trait recovery.

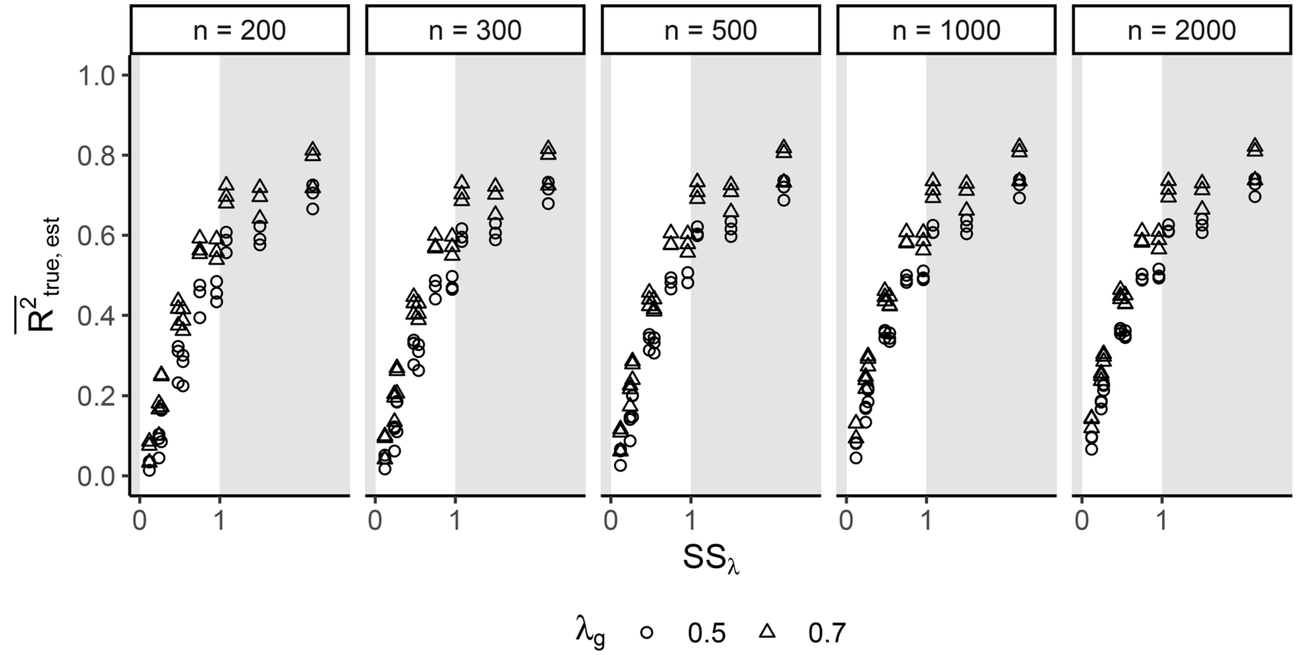


Figure 8. Average squared correlation between true domain trait values and estimated factor scores. Each symbol represents one simulation condition.

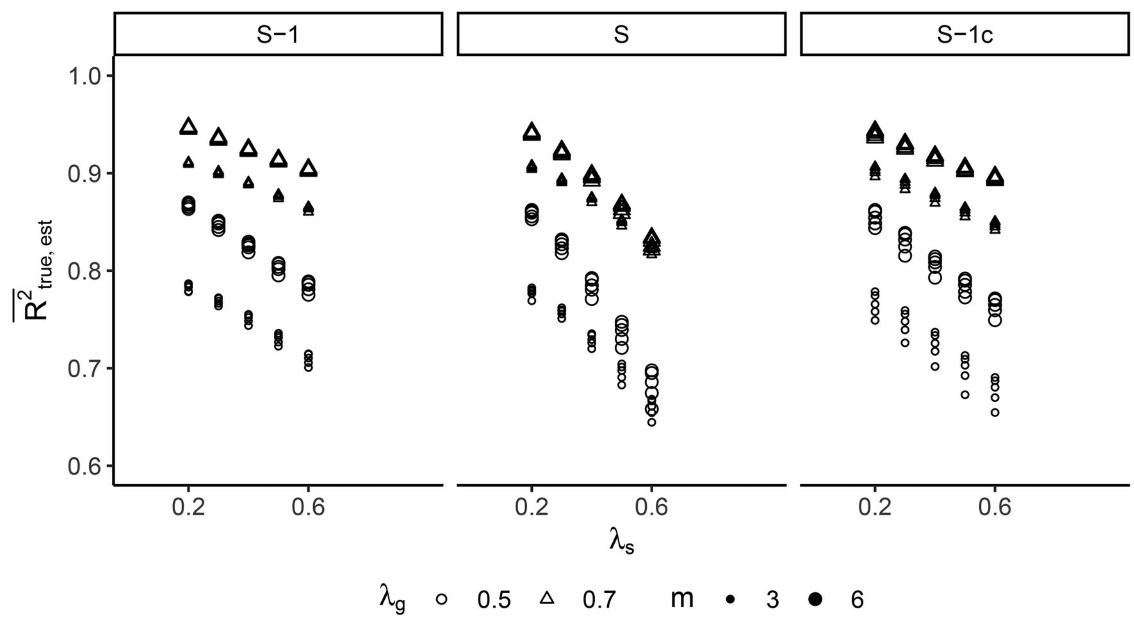


Figure 9. Average squared correlation between true general trait values and estimated factor scores. Each symbol represents one simulation condition. The variation between identical symbols is due to sample size (200 to 2000).

The model variant ($p = 0.00$, $\Delta R^2 = 0.03$), the number of indicators per domain ($p = 0.00$, $\Delta R^2 = 0.12$), the loading on the general factor ($p = 0.00$, $\Delta R^2 = 0.67$), and the size of the domain factor loadings ($p = 0.00$, $\Delta R^2 = 0.15$), all contribute uniquely to the prediction of general trait recovery. Importantly, the recovery of η_g gets worse the higher the domain factor loadings are (for constant general factor loadings). This may be counter-intuitive because it means that less reliable indicators (lower λ_s^2 and higher σ_ε^2) produce more reliable factor scores of η_g. That this effect seems strongest for the S model is probably a consequence of the additional domain factor. The model variant in Figure 9 refers to both data generation and estimation. If instead

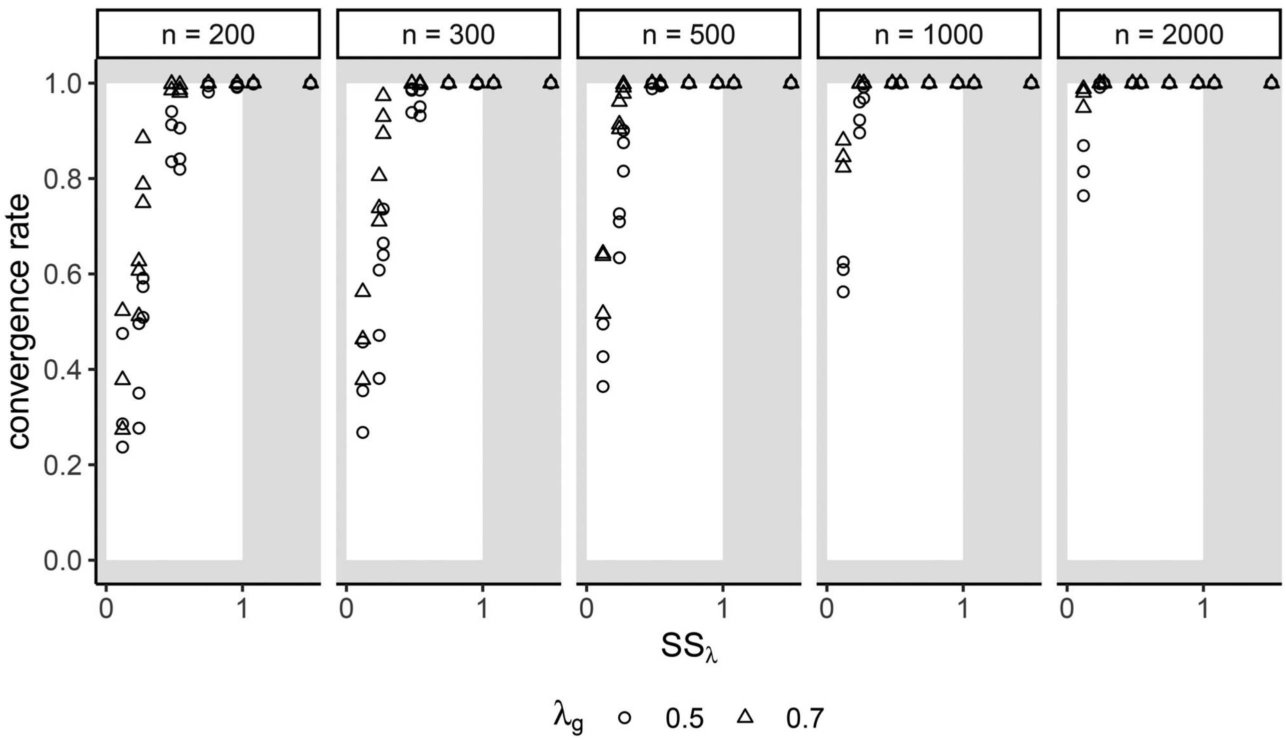


Figure 10. Convergence rate. Each symbol represents one simulation condition. Only correctly specified models are shown.

the S-1 or S-1c model is estimated on S data, the recovery of η_g is worse[12] (additional figure in supplementary materials).

Anomalous results

The main contributors to convergence problems of correctly specified models (Figure 10) are weak domain factors. The model variant ($p = 0.49$, $\Delta R^2 = 0.00$) is irrelevant for the rate of convergence. The sample size ($p = 0.00$, $\Delta R^2 = 0.10$), the number of indicators per domain ($p = 0.00$, $\Delta R^2 = 0.03$), the loading on the general factor ($p = 0.00$, $\Delta R^2 = 0.02$), and the size of the domain factor loadings ($p = 0.00$, $\Delta R^2 = 0.49$), all contribute uniquely to the prediction of convergence rates. Selective non-convergence in the presence of small factor loadings has also been observed in several other studies (for a discussion of those results, see Yuan & Bentler, 2017). Beyond that, small sample sizes and weaker loadings on η_g increase the risk of convergence problems. For model variants, the picture is less clear. The S variant tends to perform worst under otherwise problematic conditions, which may be related to the additional weak domain. In cases with misspecification (not shown), the combination of S-1 data and the estimation of the S model produces particularly bad results: the S model has convergence rates below 0.7 under all conditions. This problem is less frequent if the data-generating model is S-1c instead of S-1. According to the present simulation results, convergence problems originate from specifying factors for domains with no specific variance, not from the S model variant per sé: S model estimation works fine if the reference domain has a specific variance. Negative domain factor variance estimates are most prevalent if the true variance is small ($SS_\lambda \leq .27$, figure in supplementary materials). Without misspecification, there is no principled advantage of S-1 models over S models regarding anomalies.

Empirical example

To illustrate the potential for difficulties with weak domain factors in practice, we reanalyzed the open data shared by Dueber and Toland (2023) (https://doi.org/10.17605/OSF.IO/3QT5S). The Scoliosis Quality of Life Index (SQLI) questionnaire features 20[13] indicators measuring four subdomains with five indicators each: self-esteem (SE, indicators 1-5), back pain (BP,

[12]Given that the S-1 model was proposed along with a change in the interpretation of η_g, one could also understand this as the consequence of a change in the meaning of η_g. The current work can only demonstrate the recovery of the original data-generating trait η_g, not the interpretability or reliability of the resulting factor score if the S-1 model is estimated.

[13]The dataset provided by Dueber and Toland (2023) omits two indicators refering to satisfaction with management.

indicators 6-10), physical activity (PA, indicators 11-15), and moods and feelings (MF, indicators 16-20). The data comprise $n = 2322$ cases of adolescent idiopathic scoliosis patients.

As stated in the introduction, the approach to indicator selection plays a key role in the emergence of weak domain factors. The SQLI was developed as an adaptation of an existing questionnaire (Asher et al., 2000; Haher et al., 1999) without a repeated analysis of its covariance structure (Feise et al., 2005). The original indicator selection of the original questionnaire included an exploratory factor analysis (EFA) with varimax rotation. In a major overhaul of this original instrument, many indicators were exchanged or changed, effectively reducing the number of dimensions from seven to five (Asher et al., 2000). None of the authors report an effort to prioritize or balance generality (measuring quality of life) and discrimination of subdomains (covering distinct features of the chosen dimensions). Assumably, the resulting domain factor variance is largely a by-product of other design choices (desired total length of the scale, conceptualization and choice of domains, subscale reliability standards).

To understand the structure of the SQLI, the dissection of the indicator variances into general factor variance, domain factor variance (including the 95% confidence interval of the estimates), and unique indicator variance in a S bi-factor model[14] ($CFI = 0.95$, $RMSEA = 0.056$, $srmr = 0.051$) is displayed in Figure 11. All but one factor loading reach significance and one domain factor loading is estimated to be significantly negative ($\hat{\lambda}_{SQLI\_10,BP} = -.13$). Plotting variance proportions makes it immediately obvious that there are several indicators which barely contribute to their domain factors. This may be surprising to researchers, even if they knew the correlated-factor model of the same data ($CFI = 0.91$, $RMSEA = 0.067$, $srmr = 0.065$), in which all indicators load substantially on their respective factors ($\hat{\lambda} \geq .35$, for example $\hat{\lambda}_{SQLI\_12,PA} = 0.55$, 95% CI $[0.52, 0.58]$). Because confidence intervals are depicted in Figure 11, it is clearly visible that the near-zero estimates of some domain factor loadings in the bi-factor model are hard to explain as random underestimations. Some indicators just contribute less to the estimation of factors overall (such as SQLI_5), but importantly, there are meaningful differences in the specificity of equally reliable indicators (such as SQLI_2 and SQLI_8). The presence of near-zero domain factor loadings results in two domain factors with $SS_\lambda < 1$: $SS_\lambda^{BP} = 0.71$, $SS_\lambda^{MF} = 1.30$, $SS_\lambda^{PA} = 0.77$, $SS_\lambda^{SE} = 1.51$. For this reason, researchers might expect the domain factor scores of these factors to be substantially less reliable than those of the others (cf. Figure 8). But in turn, the domains with a higher SS_λ have smaller average factor loadings on the general SQLI factor ($M(\hat{\lambda}_{BP}) = 0.64$, $M(\hat{\lambda}_{MF}) = 0.46$, $M(\hat{\lambda}_{PA}) = 0.65$, $M(\hat{\lambda}_{SE}) = 0.39$), which also limits the precision of their factor scores. When comparing to the most favorable conditions in Figure 7, it becomes clear that domain trait recovery could be slightly increased if the indicator selection would be optimized for the measurement of the domain traits by selecting for high λ_s (which may or may not be a relevant goal). At the same time, the domain factor detection is trivial in a sample of $n > 2000$ cases (Figure 5). All p-values of the LRTs comparing the full bi-factor model to models excluding individual domain factors were below $p < 10^{-102}$ (Wald-tests of domain factor variances: all $p < 10^{-8}$).

This example shows how easily small factor loadings can appear when using a bi-factor model on a measure developed with a correlated-factor model. In this case the main problem is the limited interpretability of domains due to some of the domain factor loadings unexpectedly being close to zero.

Discussion

The aim of the meta-analysis and simulation was to identify the necessary conditions to reliably detect and estimate domain factors and their loadings, and compare these to real studies. The meta-analysis shows that many domain factor loadings are small ($\lambda_s < .2$) in practice (Figure 2) and mostly smaller than the loadings on the general factor. There is an abundance of indicators that contribute barely anything to their domain factor ($|\lambda_s| < .2$) but have reasonable loadings ($\lambda_g > .5$) on the general factor. On the one hand, this may be desired because it provides a relatively pure general factor. On the other hand, given that many domains are measured by six or less indicators (Figure 3), this results in low domain strengths (SS_λ; Figure 4). The diverse nature of the sampled constructs (Table 1), in combination with the extremely high prevalence of models having at least one domain factor for which $SS_\lambda < 1$ (Figure 4) shows that weak domain factors can be found in many research contexts.

The simulation, which covers a realistic range of factor loading values, provides an overview of the consequences of small domain factor variances

[14]There are notable differences in the factor loading estimates between this model and the values reported by the original authors (Dueber & Toland, 2023, Figure 6) because they estimated a model for categorical data, as can be seen in their open code.

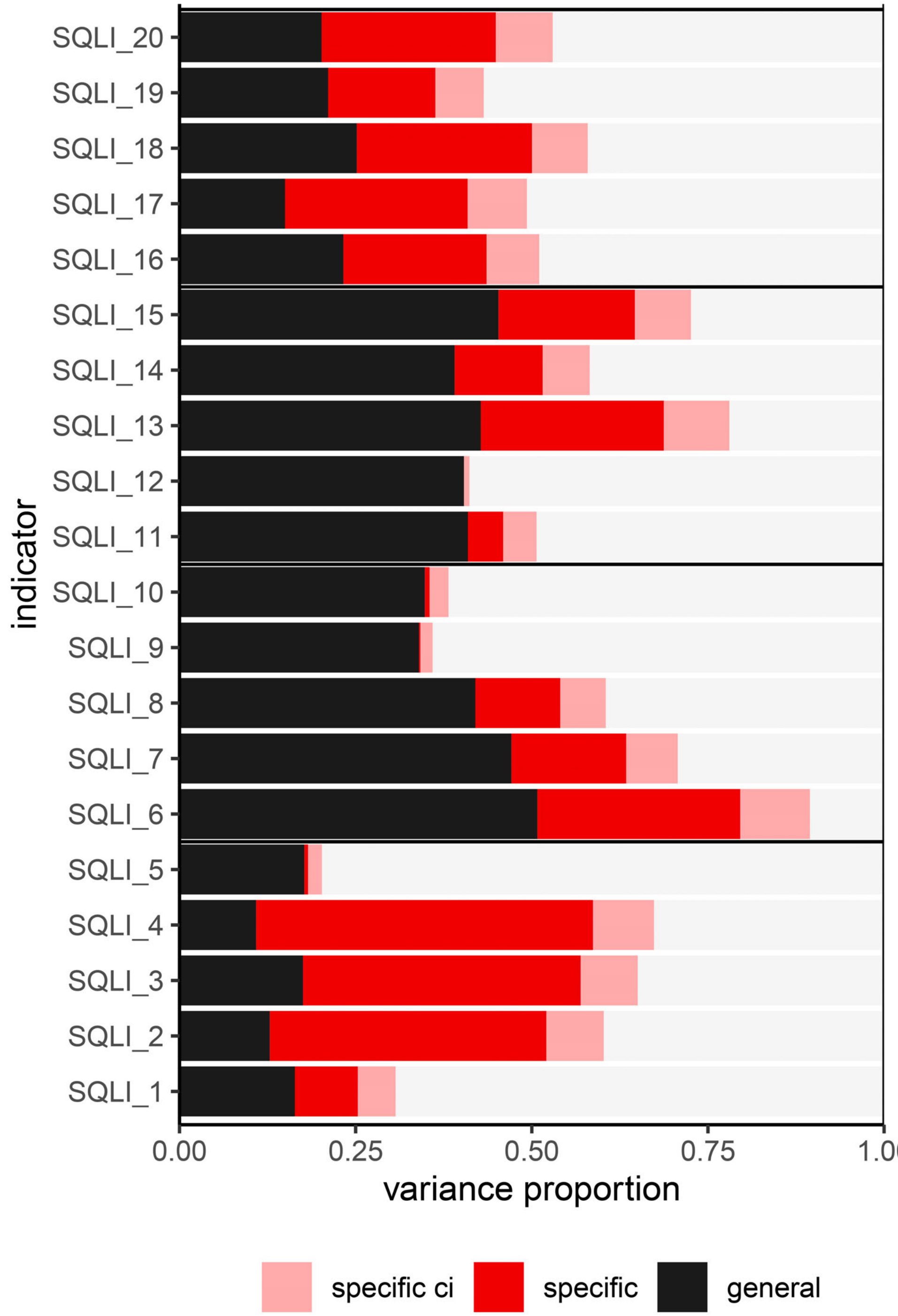


Figure 11. Variance proportions of the Scoliosis Quality of Life Index (SQLI) questionnaire explained by general and specific factors; open data by Dueber and Toland (2023); bars are non-overlapping; specific = squared lower limit of domain factor loading estimate, variance attributable to the domain factor with relative certainty; specific ci = complete 95% confidence interval of the squared estimate of the domain factor loading (lower limit to upper limit), variance potentially attributable to the domain factor; gray areas indicate leftover (error) variance if the upper limit of the specific ci were true; thick horizontal lines separate domains; the factor loadings of the indicators SQLI_10 and SQLI_12 on the respective domain factors were estimated to be negative.

(especially in the range of $SS_\lambda < 1$). The presence of domain factors is best detected by a likelihood-ratio-test (LRT) that compares the model with to the model without the domain factor. This way, domain factors with $SS_\lambda \geq 1$ will almost always be detected. In large samples and with high overall indicator reliability, much smaller effects are reliably detectable (Figure 5). Larger samples however do not meaningfully improve the precision of the estimation of domain factor scores (Figure 8) or general factor scores (Figure 9).

There was almost no difference between model variants for any of the results, meaning that "anomalous" results and the occurence of weak domain factors are not avoided by using the S-1 or S-1c variant. Judging the degree to which the prediction of other variables is affected by domain size is beyond the current simulation study (for a discussion of such models, see Zhang et al. (2021)).

How to avoid problems with weak domain factors?

Before conducting a bi-factor study, it is important to specify its goal: Should domain factors or their scores be used? Is the only consideration to obtain the best possible measure of η_g? Are all domains equally relevant? If those questions are answered at the time of the design of the study (or ideally: the measure), appropriate decisions can be made.

Expected SS_λ of domain factors

We recommend aiming for domain factor strengths of $SS_\lambda > 1$ regardless of sample size if domains should be measured. Null results of the LRT and non-convergence are unlikely for domain factors of strength $SS_\lambda > .75$. But researchers may overestimate the precision with which such domain factors are measured. About half of the domain factors from the meta-analysis are so small ($SS_\lambda < 1$) that their scores can be expected to contain $\lesssim 60\%$ true trait variance (see Figure 8). This makes the use of subscale scores highly questionable (see also Reise et al., 2013). Domain factor variance estimates below zero occur almost exclusively if the true effect size of the domain is tiny ($SS_\lambda \leq .27$). From a theoretical standpoint, factors with $SS_\lambda > 1$ are more meaningful because they represent more variance than any single indicator. In exploratory factor analysis, factors with $SS_\lambda \leq 1$ are almost always omitted, because they cannot be distinguished from random noise (parallel analysis, e.g. Hayton et al., 2004). If a study is merely concerned with measuring η_g, $SS_\lambda < 1$ can easily be tolerated (see below). If the measure's design goal is to provide valid and reliable scores of a specific domain, selecting a set of indicators with $SS_\lambda < 1$ is suboptimal, so more or better (higher λ_s) indicators need to be selected.

Number of indicators per domain

The desirable number of indicators depends on their specificity, but three to four indicators per domain are too few under most conditions. Few indicators result in small domain factor variances ($SS_\lambda < 1$). Randomly sampling six indicators from those observed in practice (Figure 2) results in $SS_\lambda < 1$ in 61.12% of cases. Adding indicators or selecting a longer measure improves the estimation precision of each individual factor loading. If domains contain very few indicators (or very few indicators with substantial loadings), including correlated error terms may be more appropriate than specifying a domain factor. The importance of increasing the number of indicators per factor to improve the recovery of the factor structure has previously been noted for EFA (Mundfrom et al., 2005; Preacher & MacCallum, 2002). For confirmatory bi-factor models, it is especially important to consider that the same number of indicators usually represents smaller SS_λ compared to other models, meaning that more indicators are needed to reliably measure domain factors compared to factors of other models (e.g., correlated-factor models).

Indicator specificity

Selecting indicators based on their specificity implies that measures are developed or revised using bi-factor models because other models do not assess indicator specificity.[15] In many cases that is not feasible for the purpose of a specific application. But it is feasible to consider the specificity of the indicators to choose realistic study goals. Low specificity is a major contributor to weak domain factors, as showcased in the empirical example (Figure 11). On the other hand, low specificity is desirable for the estimation of η_g scores. Factor loadings can themselves be of interest, for example in validation studies. Null results for domain factor loadings occur frequently for true factor loadings of 0.2 and in relatively small samples ($n \leq 500$) for loadings of 0.3 (Figure 6). In addition, small factor loadings are estimated much less precisely than larger ones (Figure 7). For the abundance of estimated loadings smaller than 0.3 in the literature (Figure 2) it is therefore difficult to judge if they are truly reflecting the domain. Indicators with low specificity are somewhat less problematic if their reliability is good (high λ_g). In the empirical example, there seemed to be a strong tradeoff between λ_g and λ_s, which we also observed more generally in the meta-analysis ($r_{\lambda_g \lambda_s} = -0.35$ ($t = -9.02$, $p < 0.001$, 95% CI [-0.42, -0.28]). This tradeoff does not exist for other models. Whereas the literature on factor structure

[15]The unique proportion of lower-order factor variance (disturbance) in higher-order factor models is not indicator-specific.

recovery in EFA considers the number and communality (i.e. reliability) of indicators (Mundfrom et al., 2005), we suggest to use SS_λ for orientation in confirmatory bi-factor analysis instead. From the results of our simulation it is clear that the size of domain factors—not the reliability of indicators—is the most important influence on statistical power and trait recovery regarding domain factors.

A priori power analysis and estimation of domain trait recovery

To estimate the statistical power to detect a domain factor, the results of this study can be used as a guideline. Alternatively, the semPower R package (Moshagen, 2021) can be used to compute the theoretical power. A simple example script is provided in the supplementary materials and can be adapted to the application at hand. The script first shows how to specify the population model and estimation model syntax to obtain the true and the model-implied variance-covariance matrixes. In the next step, the degrees of freedom for power analysis *via* semPower are set to the difference in the degrees of freedom of the two models. This is different from a standard power analysis for model misspecification. Here, the correctly specified model is the alternative option during model selection, instead of being treated as the unknown truth. The script further demonstrates how to obtain an estimate of the trait recovery for the hypothesized model. Its code is based solely on the expected standardized factor loadings (and domain factor correlations for S-1c models). It needs minimal computational resources (no simulations). If the a priori expectation for the model parameters is very uncertain, a conservative case with relatively low factor loadings should be checked. The distribution from the current meta-analysis (Figure 2) may serve as a reference. It is important to realize that theoretical power does not consider the issue of non-convergence and can therefore vastly overestimate the chance to obtain a significant result (Figure 5).

Measurement of the general factor

The most efficient way to improve the measurement of η_g is to use more indicators with higher factor loadings on η_g (Figure 9). Non-convergence becomes an issue in cases with weak domains ($SS_\lambda \leq .27$ Figure 10) or when trying to estimate non-existent domains. However, in cases that do converge, strong domain factor loadings ($\lambda_s \geq .5$, see Figure 9) are an issue. For the estimation of η_g factor scores, indicators preferably contain random error instead of domain-specific variance—even if the domain factors are included in the model. The measurement of η_g does improve with sample size, but extremely inefficiently ($\Delta R^2 < .01$). Even a tenfold increase in sample size rarely compensates for an otherwise suboptimal design.

Omission of domain factors or domain factor loadings

It is prudent to consider a set of plausible models for model selection and robustness checks. The popularity of the S bi-factor model may suggest that all indicators should be allocated to a domain, but this serves no statistical purpose. Indicators that do not belong to a domain do not invalidate the model. The current meta-analysis found a large proportion of indicators with low specificity—likely due to indicator selection based on other models. In the empirical example, the bi-factor model of the SQLI included several indicators with little to no contribution to their domain factor, which could have easily gone unnoticed during the development of the measure, even if a correlated-factor model would have been considered. Indicator allocation to domains should be reconsidered in these cases. For this purpose, exploratory bi-factor analysis techniques (Jennrich & Bentler, 2011, 2012) and bi-factor exploratory structural equation models (Morin et al., 2016) were developed. Instead, what does lead to all kinds of problems are domain factors without specific variance in the population. Such null results for domain factors can be perfectly acceptable, for example, if domains represent converging measurement methods. But importantly, the respective factors then have to be omitted from the model.

Troubleshooting non-convergence

If a bi-factor model does not converge, one should try to omit the domain factor that is expected to be the weakest. Non-convergence is not an issue given a correctly specified model and reasonably large domain factors (Figure 10). In practice, however, "all models are wrong" (Box, 1976, p. 792). So with inevitable misspecification, non-convergence may occur more frequently—possibly most frequently for the S model variant. Convergence is worst for the S model on S-1 data, or if domain factors ($SS_\lambda \leq .27$) and sample sizes are small (See Figure 10). The main problem seems to be the specification of superfluous or very weak domain factors, which should be avoided. For a detailed analysis of convergence problems in structural equation modeling and some other potential solutions, see Yuan and Bentler (2017).

How to interpret weak domain factors and weak domain factor loadings?

In the interpretation of bi-factor models, statistical power and the precision of estimates needs to be taken into account more thoroughly. For this, it is useful to compute the SS_λ of domain factors. Our simulation provides a general reference for statistical power and parameter recovery[16] given a range of realistic cases. The example script (supplementary materials) can be used to examine a specific case. Domain factors can include a surprisingly small amount of systematic variance (Figure 8, see also Reise et al., 2013) and may have multiple indicators whose attribution to them is unclear (Figures 2 and 11). If domains are used to predict third variables, this may explain their failure to do so. They could be just as weak as domains that result from random allocation of indicators to domains: Bi-factor models tend to fit almost any pattern in the data (Bonifay & Cai, 2017).

Taking a closer look at the factor loadings is often crucial. The large variation in loadings on the domain factor (Figures 2 and 11) means there is a very uneven mixture of the contribution of indicators to domains (e.g. Watts et al., 2019). To communicate factor composition clearly, figures of factor loadings (e.g. bar charts, such as Figure 11) can be useful. In addition to the variation in the factor loading estimates, there is substantial variation in their estimation precision (Figure 7). They should be interpreted more carefully when they are small ($\lambda_s \leq .3$), overall indicator reliability is far from perfect ($Rel < 0.5$), or the sample size is small ($n \leq 300$). Point estimates are most misleading for the most relevant loadings: small loadings that are often hard to interpret. It would be useful to always report (and interpret) standard errors and confidence intervals of factor loadings to make this visible, as we did in Figure 11. However, the fact that many domain factor loadings are estimated near or below zero (Figure 2) cannot be explained by sampling variation alone (Figure 7), certainly not in the empirical example.

[16]An alternative is the index H for construct reliability (Hancock, 2001; see also Rodriguez et al., 2016) which is more straightforward but does not take the impact of η_g into account (cf. Figure 8).

Are S-1 models and models with a null result on a domain factor the same?

Models with omitted domain factors should not all be interpreted the same. If a domain factor is omitted because it is too weak, the resulting model is structurally equivalent to an S-1 model. However, the domain in question may not necessarily be interpreted as a natural reference domain, especially if it has small loadings on the general factor. For the interpretation of the remaining estimates, it does not matter if the absence of the domain was defined or estimated, so the interpretation of η_g does not need to change. A priori S-1 models on the other hand were proposed irrespective of the size of the unique variance of the reference domain and should therefore be interpreted differently (see Eid et al., 2017). Their reference domain clarifies the meaning of η_g, which is especially relevant if the reference domain has a unique variance that could be attributed to it.

Are small domain factor loadings an empirical fact or a technical artifact?

Looking at the distribution of factor loadings in Figure 2, researchers may come to the conclusion that the many small domain-factor loadings ($\lambda_s < .2$) are a valid empirical finding, rather than indicating a statistical or measurement issue. If they reflect the nature of the construct accurately, it would be undesirable to try to find indicators with higher domain-factor loadings. Such an effort could even challenge the validity of the measure. For this reason, it is important to consider the multiple ways in which these factor loadings are influenced. Firstly, indicators may be selected based on their factor loadings—irrespective of their content—usually prefering those with higher reliabilities. This strategy is based on the idea that there are better and worse constructed, and more or less relevant indicators, and the better, more relevant ones should be chosen. Secondly, indicators may be selected for reflecting a certain domain based on their content, in a try to best capture the essence of the domain (e.g., extraversion indicators that most clearly describe prototypical social boldness behaviors). In both cases, near-zero domain-factor loadings would indicate a failure to construct or select appropriate indicators. Thirdly, indicators may be selected, because they are considered to measure an important, irreplacable part of the target construct, irrespective of dimensionality (e.g., symptoms in clinical assessment or criterion-relevant tasks in a performance test). To the degree that these indicators are properly designed, small factor loadings or SS_λ values of domain factors are then a relevant empirical finding. In these cases, researchers need to deal with the resulting domain factor and accept interpretational difficulties. Overall, we consider the results of our meta-analysis to be a mixture of these different scenarios. The current study should help researchers to avoid obtainining such results by accident, that is without having strong arguments to interpret small factor loadings as a relevant empirical finding.

Limitations and future directions

High estimation precision does not guarantee interpretability. We agree with Eid et al. (2017) that the interpretability of bi-factor models needs more careful attention and should guide model selection. S-1 models were introduced to improve interpretability in cases with a fixed set of domains (in which domains are not randomly sampled). Eid et al. (2017) demonstrated a straightforward interpretation of S-1 models for this common case. They warned that S models lack a clear interpretation of the general factor in cases with a fixed set of domains. Although the current simulation showed that anomalous results occur in all model variants, this does not mean that S and S-1 models are equally interpretable. On top of that, the S variant is prone to identification problems when used as a measurement model in SEM (Zhang et al., 2021).

In the current simulation, factor loadings were fixed to be equal and constant within and across domains. This very selective set of scenarios greatly simplified the design and interpretation of the simulation. Most probably, problems with the estimation of a particular domain factor or domain factor loading are less severe if the rest of the model consists of more reliable indicators. Vice versa, the estimation of one part of the model may become more problematic if the rest of the model consists of less reliable indicators. For this reason, we suggest interpreting the results of the simulation with the whole model in mind. When in doubt one should check the specific case. Furthermore, we omitted imperfections (cross loadings, correlated errors) in the simulated data, which are frequently encounterd in practice (Morin et al., 2016). Such added complexity could both hamper efforts to detect and estimate domain factors and produce spurious or inflated factors.

The current simulation assumes continuous, normally distributed error terms (and latent traits). In practice, this assumption is usually violated (Micceri, 1989) and robust methods should be considered (see e.g., Yuan & Bentler, 2007). Furthermore, data analyzed in Confirmatory Factor Analysis (CFA) are frequently categorical (i.e., measured on Likert-scales). In principle, categorical data are better analyzed using Item Response Theory (IRT) models. The estimation of parameters, χ^2 values, and fit indexes in CFA can be—but is not necessarily—biased by the categorization of data (DiStefano, 2002; Finney & DiStefano, 2006). Despite these issues, many researchers make use of CFA models on categorical data. If bi-factor CFA models are used to analyze categorical or decidedly non-normal data, it is especially important to consider the current results to be an optimistic upper limit of the to-be-expected statistical power, trait recovery, and parameter estimation precision. Future research may show if bi-factor IRT models also tend to produce weak domain traits on typical data.

The current study did not examine how weak domain factors affect estimates in the structural part of SEMs. This topic is only partly touched by the simulation data of Zhang et al. (2021) who demonstrated a strong influence of the model variant on SEM estimates. Further research is needed to explore the influence of domain strength on relationships with other variables. Domain factors with $SS_\lambda < 1$ might show estimates of latent relationships that are imprecise and biased toward zero, because they are measured with less precision. To corroborate the empirical result of our meta-analysis that many measures do not produce a full set of interpretable domain-specific factors, assessing the prevalence of weak or vanishing domain factors using exploratory models (Jennrich & Bentler, 2011, 2012; Morin et al., 2016) on a representative sample of studies would be useful. This is especially relevant, because results of bi-factor CFA might be biased in cases with substantial cross-loadings, which can realistically be expected in many applications (Morin et al., 2016). Finally, several models are structurally similar to the bi-factor model (multitrait-multimethod models, longitudinal models, latent state-trait models, e.g. Koch et al., 2018). Future research may show to what degree these involve similar challenges.

Conclusion

The role and prevalence of study designs that produce small domain factor strengths—which lead to null results or uninterpretable results—are underappreciated in the literature. Study planning and interpretation need to take the (expected) strength of domain factors and domain factor loadings into account. The outlined strategies aim to enable researchers to fully unlock the model's potential. The bi-factor model does not generally produce problematic results, but it needs appropriate data. The crucial step is to select or design measures for the use of bi-factor models. If that is not possible, the results have to be interpreted with caution and alternative models should be considered. Moreover, the current study provides further explanations for the results that Eid et al. (2017) termed "anomalous". It shows that they occur in the S-1 and S-1c variants with roughly the same frequency if there is no misspecification involved.

Many of the above suggestions imply that existing measures need to be revised or new measures need to be developed to meet common study goals. This is both a challenge and a chance. There are many reasons why current measurement practices are considered suboptimal (Flake & Fried, 2020). Bi-factor models offer new opportunities to create improved measures, especially if the underlying construct is multifaceted by definition. The measurement of domain traits may be a practical challenge, but with it comes an opportunity to refine psychological research.

Author note

The authors made the following contributions. Nils Petras: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing; Thorsten Meiser: Conceptualization, Supervision, Writing - Review & Editing.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2277 "Statistical Modeling in Psychology".

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like thank Celine Kumpf and Alicia Gernand for their help on collecting the meta-analysis data. We thank Celine Kumpf and Alicia Gernand for their help on collecting the meta-analysis data. We thank Marie Mundt for assisting in the literature search for the empirical example with open data. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution or the German Research Foundation is not intended and should not be inferred.

ORCID

Nils Petras http://orcid.org/0000-0001-9528-2298
Thorsten Meiser http://orcid.org/0000-0001-6004-9787

Data availability statement

All code and data of the meta-analysis and simulation study, as well as the code generating the manuscript are available here: https://osf.io/qys8u/.

References

Asher, M. A., Lai, S. M., & Burton, D. C. (2000). Further development and validation of the scoliosis research society (SRS) outcomes instrument. *Spine*, *25*(18), 2381–2386. https://doi.org/10.1097/00007632-200009150-00018

Beaujean, A. A. (2015). John carroll's views on intelligence: Bi-factor vs. Higher-order models. *Journal of Intelligence*, *3*(4), 121–136. https://doi.org/10.3390/jintelligence3040121

Blanco, C., Rubio, J. M., Wall, M., Secades-Villa, R., Beesdo-Baum, K., & Wang, S. (2014). The latent structure and comorbidity patterns of generalized anxiety disorder and major depressive disorder: A national study. *Depression and Anxiety*, *31*(3), 214–222. https://doi.org/10.1002/da.22139

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science: A Journal of the Association for Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

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. https://doi.org/10.1207/s15327906mbr4102_5

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(3), 327–346. https://doi.org/10.1207/S15328007SEM0903_2

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20. https://doi.org/10.7275/da8t-4g52

Dueber, D. M., & Toland, M. D. (2023). A bifactor approach to subscore assessment. *Psychological Methods*, *28*(1), 222–241. https://doi.org/10.1037/met0000459

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence*, *6*(3), 42. https://doi.org/10.3390/jintelligence6030042

Feise, R. J., Donaldson, S., Crowther, E. R., Menke, J. M., & Wright, J. G. (2005). Construction and validation of the scoliosis quality of life index in adolescent idiopathic scoliosis. *Spine*, *30*(11), 1310–1315. https://doi.org/10.1097/01.brs.0000163885.12834.ca

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural Equation Modeling: A Second Course*, *10*(6), 269–314.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*(10), e1701381. https://doi.org/10.1126/sciadv.1701381

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T. (2020). *mvtnorm: Multivariate normal and t distributions*. https://CRAN.R-project.org/package=mvtnorm

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. https://doi.org/10.1080/00273171.2013.804398

Giordano, C., & Waller, N. G. (2020). Recovering bifactor models: A comparison of seven methods. *Psychological Methods*, *25*(2), 143–156. https://doi.org/10.1037/met0000227

Haher, T. R., Gorup, J. M., Shin, T. M., Homel, P., Merola, A. A., Grogan, D. P., Pugh, L., Lowe, T. G., & Murray, M. (1999). Results of the scoliosis research society instrument for evaluation of surgical outcome in adolescent idiopathic scoliosis: A multicenter study of 244 patients. *Spine*, *24*(14), 1435. https://doi.org/10.1097/00007632-199907150-00008

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*(3), 373–388. https://doi.org/10.1007/BF02294440

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods 7*(2), 191–205. https://doi.org/10.1177/1094428104263675

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*(4), 537–549. https://doi.org/10.1007/s11336-011-9218-4

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika*, *77*(3), 442–454. https://doi.org/10.1007/s11336-012-9269-1

Keith, T. Z., & Reynolds, M. R. (Eds.) (2018). *Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests*. The Guilford Press.

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychological Methods*, *23*(3), 505–523. https://doi.org/10.1037/met0000146

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), 971–977. https://doi.org/10.1037/a0028355

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*, 51–69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

McCulloch, C. E., & Searle, S. R. (2004). *Generalized, linear, and mixed models*. John Wiley & Sons. https://doi.org/10.1002/0470011815.b2a10021

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician*, *61*(1), 22–27. https://doi.org/10.1198/000313007X171322

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. https://doi.org/10.1080/10705511.2014.961800

Moshagen, M. (2021). *semPower: Power analyses for SEM*. https://CRAN.R-project.org/package=semPower

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

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: The Journal of Mental Science*, *207*(1), 15–22. https://doi.org/10.1192/bjp.bp.114.149591

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics*, *32*(2), 153–161. https://doi.org/10.1023/A:1015210025234

R Core Team. (2020). *R A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/ https://doi.org/10.18637/jss.v048.i02

Scholz, D. D., Hilbig, B. E., Thielmann, I., Moshagen, M., & Zettler, I. (2022). Beyond (low) agreeableness: Toward a more comprehensive understanding of antagonistic psychopathology. *Journal of Personality*, *90*(6), 956–970. https://doi.org/10.1111/jopy.12708

Steyer, R. (1989). *Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability*. Methodika.

Stoel, R. D., Garre, F. G., Dolan, C., & Van Den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*(4), 439–455. https://doi.org/10.1037/1082-989X.11.4.439

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*(6), 1285–1303. https://doi.org/10.1177/2167702619855035

Yuan, K.-H., & Bentler, P. M. (2007). Robust procedures in structural equation modeling. In *Handbook of latent variable and related models* (pp. 367–397). Elsevier. https://doi.org/10.1016/S1871-0301(06)01017-1

Yuan, K.-H., & Bentler, P. M. (2017). Improving the convergence rate and speed of fisher-scoring algorithm: Ridge and anti-ridge methods in structural equation modeling. *Annals of the Institute of Statistical Mathematics*, *69*(3), 571–597. https://doi.org/10.1007/s10463-016-0552-2

Yuan, K.-H., Cheng, Y., & Zhang, W. (2010). Determinants of standard errors of MLEs in confirmatory factor analysis. *Psychometrika*, *75*(4), 633–648. https://doi.org/10.1007/s11336-010-9169-1

Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2021). Using bifactor models to examine the predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational Research Methods*, *24*(3), 530–571. https://doi.org/10.1177/1094428120915522

Appendix A. List of articles included in the meta-analysis sample:

Balsamo, M., Romanelli, R., Innamorati, M., Ciccarese, G., Carlucci, L., & Saggino, A. (2013). The State-Trait Anxiety Inventory: Shadows and Lights on its Construct Validity. *Journal of Psychopathology and Behavioral Assessment, 35*, 475–486. https://doi.org/10.1007/s10862-013-9354-5

Booth, T., Bastin, M. E., Penke, L., Maniega, S. M., Murray, C., Royle, N. A., Gow, A. J., Corley, J., Henderson, R. D., Hernández, M. del C. V., Starr, J. M., Wardlaw, J. M., & Deary, I. J. (2013). Brain white matter tract integrity and cognitive abilities in community-dwelling older people: The Lothian Birth Cohort, 1936. *Neuropsychology, 27*(5), 595–607. https://doi.org/10.1037/a0033354

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychological Assessment, 25* 1, 136–145. https://doi.org/10.1037/a0029228

Burns, G. L., de Moura, M. A., Beauchaine, T. P., & McBurnett, K. (2014). Bifactor latent structure of ADHD/ODD symptoms: predictions of dual-pathway/trait-impulsivity etiological models of ADHD. *Journal of Child Psychology and Psychiatry, 55*(4), 393–401. https://doi.org/10.1111/jcpp.12165

Byrd, A. L., Kahn, R. E., & Pardini, D. A. (2013). A validation of the inventory of callous-unemotional traits in a community sample of young adult males. *Journal of Psychopathology and Behavioral Assessment, 35*(1), 20–34. https://doi.org/10.1007/s10862-012-9315-4

Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing, 30*(2), 177–199. https://doi.org/10.1177/0265532212456833

DeSousa, D. A., Zibetti, M. R., Trentini, C. M., Koller, S. H., Manfro, G. G., & Salum, G. A. (2014). Screen for Child Anxiety Related Emotional Disorders: Are subscale scores reliable? A bifactor model analysis. *Journal of Anxiety Disorders, 28*(8), 966–970. https://doi.org/10.1016/j.janxdis.2014.10.002

DiStefano, C., Greer, F. W., & Kamphaus, R. W. (2013). Multifactor modeling of emotional and behavioral risk of preschool-age children. *Psychological Assessment, 25*(2), 467–476. https://doi.org/10.1037/a0031393

Gomez, R. (2013). Depression Anxiety Stress Scales: Factor structure and differential item functioning across women and men. *Personality and Individual Differences, 54*(6), 687–691. https://doi.org/10.1016/j.paid.2012.11.025

Gomez, R., Kyriakides, C., & Devlin, E. (2014). Attention-Deficit/Hyperactivity Disorder symptoms in an adult sample: Associations with Rothbart's temperament dimensions. *Personality and Individual Differences, 60*, 73–78. https://doi.org/10.1016/j.paid.2013.12.023

Grygiel, P., Humenny, G., Rebisz, S., Świtaj, P., & Sikorska, J. (2013). Validating the Polish adaptation of the 11-item De Jong Gierveld Loneliness Scale. *European Journal of Psychological Assessment, 29*(2), 129–139. https://doi.org/10.1027/1015-5759/a000130

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for General and Domain-Specific Elements of Teacher-Child Interactions: Associations With Preschool Children's Development. *Child Development, 85*(3), 1257–1274. https://doi.org/10.1111/cdev.12184

Hyland, P., Shevlin, M., Adamson, G., & Boduszek, D. (2013). The factor structure and composite reliability of the Profile of Emotional Distress. *The Cognitive Behaviour Therapist, 6*, 1–12. https://doi.org/10.1017/S1754470X13000214

Jonason, P. K., Kaufman, S. B., Webster, G. D., & Geher, G. (2013). What Lies Beneath the Dark Triad Dirty Dozen: Varied Relations with the Big Five. *Individual Differences Research, 11*(2), 81–90.

Lo Coco, A., Ingoglia, S., & Lundqvist, L.-O. (2014). The Assessment of Susceptibility to Emotional Contagion: A Contribution to the Italian Adaptation of the Emotional Contagion Scale. *Journal of Nonverbal Behavior, 38*(1), 67–87. https://doi.org/10.1007/s10919-013-0166-9

Luciano, J. V., Barrada, J. R., Aguado, J., Osma, J., & García-Campayo, J. (2014). Bifactor analysis and construct validity of the HADS: A cross-sectional and longitudinal study in fibromyalgia patients. *Psychological Assessment, 26*(2), 395–406. https://doi.org/10.1037/a0035284

Mészáros, V., Ádám, Sz., Szabó, M., Szigeti, R., & Urbán, R. (2014). The Bifactor Model of the Maslach Burnout Inventory-Human Services Survey (MBI-HSS)-An Alternative Measurement Model of Burnout. *Stress & Health: Journal of the International Society for the Investigation of Stress, 30*(1), 82–88. https://doi.org/10.1002/smi.2481

Norwalk, K. E., DiPerna, J. C., & Lei, P.-W. (2014). Confirmatory factor analysis of the Early Arithmetic, Reading, and Learning Indicators (EARLI). *Journal of School Psychology, 52*(1), 83–96. https://doi.org/10.1016/j.jsp.2013.11.006

Olatunji, B. O., Ebesutani, C., Haidt, J., & Sawchuk, C. N. (2014). Specificity of Disgust Domains in the Prediction of Contamination Anxiety and Avoidance: A Multimodal Examination. *Behavior Therapy, 45*(4), 469–481. https://doi.org/10.1016/j.beth.2014.02.006

Park, M.-H., Dimitrov, D. M., Das, A., & Gichuru, M. (2016). The teacher efficacy for inclusive practices (TEIP) scale: dimensionality and factor structure. *Journal of Research in Special Educational Needs, 16*(1), 2–12. https://doi.org/10.1111/1471-3802.12047

Tripp, M. K., Diamond, P. M., Vernon, S. W., Swank, P. R., Mullen, P. D., & Gritz, E. R. (2013). Measures of parents' self-efficacy and perceived barriers to children's sun protection: construct validity and reliability in melanoma survivors. *Health Education Research, 28*(5), 828–842. https://doi.org/10.1093/her/cys114

Varni, J. W., Beaujean, A. A., & Limbers, C. A. (2013). Factorial invariance of pediatric patient self-reported fatigue across age and gender: a multigroup confirmatory factor analysis approach utilizing the PedsQL™ Multidimensional Fatigue Scale. *Quality of Life Research, 22*(9), 2581–2594. https://doi.org/10.1007/s11136-013-0370-4

Witthöft, M., Hiller, W., Loch, N., & Jasper, F. (2013). The Latent Structure of Medically Unexplained Symptoms and Its Relation to Functional Somatic Syndromes. *International Journal of Behavioral Medicine, 20*(2), 172–183. https://doi.org/10.1007/s12529-012-9237-2

Yap, S. C. Y., Donnellan, M. B., Schwartz, S. J., Kim, S. Y., Castillo, L. G., Zamboanga, B. L., Weisskirch, R. S., Lee, R. M., Park, I. J. K., Whitbourne, S. K., & Vazsonyi, A. T. (2014). Investigating the Structure and Measurement Invariance of the Multigroup Ethnic Identity Measure in a Multiethnic Sample of College Students. *Journal of Counseling Psychology, 61*(3), 437–446. https://doi.org/10.1037/a0036253

Young, M. A., Hutman, P., Enggasser, J. L., & Meesters, Y. (2015). Assessing Usual Seasonal Depression Symptoms: The Seasonality Assessment Form. *Journal of Psychopathology and Behavioral Assessment, 37*(1), 112–121. https://doi.org/10.1007/s10862-014-9440-3

Zheng, Y., Chang, C.-H., & Chang, H.-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research, 22*(3), 491–499. https://doi.org/10.1007/s11136-012-0179-6

When Factor Variance and Factor Correlations are Interchangeable: The Relationship Between the Bi-Factor Model Variants

Nils Petras[1]

[1] University of Mannheim
School of Social Sciences
Department of Psychology

Author Note

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2277 "Statistical Modeling in Psychology"

The Author has no further competing interests to declare that are relevant to the content of this article.

The datasets and code generated and analyzed during the current study are available in the OSF repository https://osf.io/e9c6f/, including this reproducible manuscript.

I thank Thorsten Meiser for his very useful comments on an earlier draft of this manuscript.

Correspondence concerning this article should be addressed to Nils Petras, L13, 15, 68161 Mannheim, Germany. E-mail: nils.petras@uni-mannheim.de

Abstract

Despite the bi-factor model's recent rise in popularity, the mathematical relationship between its variants is not yet understood. The additions of free parameters that characterize the variants can be – but are not necessarily – mere reparameterizations. The current work demonstrates this analytically and through simulations. It is newly established that the higher-order factor model is nested in one of the novel bi-factor model variants, not only the symmetric one. More generally, the simulations show how the bi-factor model variants not nested within each other can closely fit the other variants' data. The mutual imitation between model variants leads to a complex pattern of differences in parameter estimates and factor score estimates, so the validity of many claims is conditional on the model variant. For instance, it is possible that the omission of a specific factor can be perfectly compensated by the addition of freely estimated correlations between the remaining specific factors. These equivalent models suggest very different interpretations. Moreover, swapping between model variants affects all parameter estimates systematically. The current study uncovers these patterns. The potential for similar patterns in multi-trait multimethod model variants is discussed.

Keywords: bi-factor model, confirmatory factor analysis, S-1 bi-factor model, higher-order factor model

Word count: 6257

When Factor Variance and Factor Correlations are Interchangeable: The Relationship Between the Bi-Factor Model Variants

Introduction

Bi-factor models (Holzinger & Swineford, 1937) have become increasingly popular in psychological research over the past years (Reise, 2012; Zhang et al., 2020). They account for linear relations among a set of indicators (observed variables, items) by defining a general factor and several specific domain factors across different content domains, raters, tasks, or otherwise grouped indicators. For decades, the standard variant of this model was fully symmetrical, meaning that every indicator variable loads on the general and one specific factor. More recently, several variants of the model were proposed (Eid et al., 2017), but their mathematical relationship is not yet understood.

All variants introduce at least one indicator that exclusively loads on the general factor. One of the variants includes a reduced set of correlated (instead of orthogonal) domain factors. Critically, bi-factor models can either include a full set of domain factors (Holzinger & Swineford, 1937) or freely estimate correlations between domain factors (Eid et al., 2017) – but not both (Markon, 2019). This means there is no proper superordinate model for comparison, in which parameters are freely estimated. For this reason, the relationship between the variants that are not nested within each other needs to be examined directly.

This article analyzes the relationship between the bi-factor model variants regarding two major open questions. 1) How well can the different variants imitate each other? The standard bi-factor model is known to be very flexible (Bonifay & Cai, 2017). Past work has shown that the bi-factor model can account for data from other models much better than vice versa (Bader & Moshagen, 2022; Greene et al., 2019). So far, this has only been shown for the comparison between the bi-factor model and other models. Here, the different variants of the model are compared. 2) How are the parameters of the model variants related to each other? This question is relevant for multiple reasons: a) meaningfully estimating all parameters freely at the same time is not possible, b) the variants can be reparameterizations of each other, c) even if not, the variants often fit the same data almost equally well, and d) the variants share a lot of parameters whose meaning subtly changes with the variant.

To answer these two questions, it is first shown analytically that the higher-order factor model is a special case of multiple variants, meaning that these variants can be mere reparameterizations of each other. For those cases in which they are not, the relationship between the parameters of the variants is analyzed in a simulation study. In the simulation, all bi-factor model variants are estimated on data generated by all the variants. Regarding the question of mutual imitation, the model fit is analyzed. Regarding the relationships between parameters, both the model parameter estimates and the estimated trait values are compared between the variants. Implications for the validity of claims based on bi-factor models will be discussed at the end, as well as the limitations of the current study. It first follows the introduction of the notation.

Bi-factor model variants

In the original, symmetric variant of the bi-factor model (S, Figure 1), every indicator Y loads on both a general factor η_g and one domain factor η_s. The response matrix $\mathbf{Y}$ (Equation (1)) is the sum of the product of the matrix of factor loadings ($\mathbf{\Lambda}$) and the matrix of latent trait values ($\boldsymbol{\eta}$), and the matrix of error values $\boldsymbol{\varepsilon}$, which are independently and normally distributed for each indicator. Characteristic of a bi-factor model with z domain factors is a loading matrix $\mathbf{\Lambda}$ with $z+1$ columns, in which there are two non-zero entries per row: one in the first column, pertaining to η_g, and one in one of the further z columns, pertaining to some η_s.

$$\mathbf{Y} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{1}$$

In the S variant, all factors are orthogonal, so their covariance matrix $\mathbf{\Phi}$ is a diagonal matrix. In the S-1 variant (Eid et al., 2017), one domain factor is omitted (Figure 1).[1] The domain without a specific factor then becomes the reference domain, which defines the meaning of the general factor. In the S-1c variant, the correlations of the remaining domain factors are estimated freely. It is inadvisable to both include domain factors for all indicators and freely estimate domain factor correlations. A full set of positively correlated domain factors would be partially redundant with the general factor, causing problems in estimation and interpretation (Markon, 2019).

The S-1 variant is nested in both the S and the S-1c variant. Either the S or the S-1c variant can be more parsimonious, depending on the number of domains (z) and indicators of the reference domain (m). The difference in the degrees of freedom is shown in Equation (2). For example, in a model with 5 domains and 6 indicators of the reference domain, the degrees of freedom of the S and S-1c variants are equal.

$$df_S - df_{S-1c} = \frac{(z-1)(z-2)}{2} - m \tag{2}$$

According to Eid et al. (2017), when the domains are not randomly sampled, the interpretation of the S-1 and S-1c variants is more straightforward than that of the S variant. They suggest that the reference domain (or item) without a specific factor defines the meaning of the general factor. For the S variant, there is no clear interpretation based on stochastic measurement theory (Steyer, 1989) without assuming randomly sampled domains (Eid et al., 2017). The selection of a specific set of domains is typical for measures with subdomains (= facets), such as personality inventories or intelligence tests. Eid et al. (2017) advise researchers to prefer the S-1 or S-1c variant a priori in these cases, irrespective of model fit. It remains unclear what exactly that means for the interpretation of trait and

[1] Eid et al. (2017) also introduced a S*I-1 variant, in which the reference consists of a single indicator. For simplicity, I omit this special case here.

parameter estimates, because estimated S-1c and S models can be – but are not necessarily – reparameterizations of each other.

Nesting structure

Figure 2 shows the nesting structure of the bi-factor model variants and the related group-factor and higher-order factor models[2]. A model is nested within another model if it is a restricted version of that model. Although there are several nesting relationships, the hierarchy is incomplete.

Among the different variants of the bifactor model, S-1 is a restricted version of both S and S-1c because it is defined by restricting some of their parameters to 0. The relationship between the higher-order model and the bi-factor model variants is more complicated. Higher-order factor models can be reparameterized as S bi-factor models (Schmid & Leiman, 1957; Yung et al., 1999) by restricting the domain-specific loadings of the indicators to be proportional to their general factor loadings. Moreover, the higher-order factor model is a special case of the S-1c bi-factor model, as shown below. Lastly, the higher-order model is a special case of the group-factor model, if there are at least four first-order factors. This leaves the more complex comparison between the S and S-1c bi-factor model variants, which are not nested and have free parameters in different parts of the model. The S-1c and S variants can each provide unique solutions that cannot be reparameterized as solutions of the respective other model variant.

The higher-order factor model is nested within the S-1c bi-factor model

Be S^* a restricted S bi-factor model, with a constant k_x for each domain x, so that $\lambda_{x_i s} = k_x \lambda_{x_i g}$ for all indicators i of the domain.[3] Be $\forall x \in X : k_x > 0$ by inversion of domain

[2] Here, the term "higher-order factor model" refers exclusively to models with one second-order factor. Higher-order factor models in other contexts may include multiple second-order factors and more than two levels of factors but these extensions are rarely seen in applied studies and not relevant here.

[3] k^2 was termed "explained common variance of the specific factor" (ECV_{SS}, Dueber & Toland, 2021) when referring to a complete domain instead of individual indicators.

factors with loadings contrary to the general factor. Be $\forall i \in I : \lambda_{x_i g} > 0$, so that all factor loadings are nonzero and positive. The loading matrix $\mathbf{\Lambda}$ of such a model (with four domains and three indicators per domain) is shown in Equation (3), including a numerical example with $k_1 = 1/2$, $k_2 = 2/3$, $k_3 = 1/4$, and $k_4 = 5/4$.

$$\mathbf{\Lambda}^* = \begin{pmatrix} \lambda_{1_1 g} & k_1\lambda_{1_1 g} & 0 & 0 & 0 \\ \lambda_{1_2 g} & k_1\lambda_{1_2 g} & 0 & 0 & 0 \\ \lambda_{1_3 g} & k_1\lambda_{1_3 g} & 0 & 0 & 0 \\ \lambda_{2_1 g} & 0 & k_2\lambda_{2_1 g} & 0 & 0 \\ \lambda_{2_2 g} & 0 & k_2\lambda_{2_2 g} & 0 & 0 \\ \lambda_{2_3 g} & 0 & k_2\lambda_{2_3 g} & 0 & 0 \\ \lambda_{3_1 g} & 0 & 0 & k_3\lambda_{3_1 g} & 0 \\ \lambda_{3_2 g} & 0 & 0 & k_3\lambda_{3_2 g} & 0 \\ \lambda_{3_3 g} & 0 & 0 & k_3\lambda_{3_3 g} & 0 \\ \lambda_{4_1 g} & 0 & 0 & 0 & k_4\lambda_{4_1 g} \\ \lambda_{4_2 g} & 0 & 0 & 0 & k_4\lambda_{4_2 g} \\ \lambda_{4_3 g} & 0 & 0 & 0 & k_4\lambda_{4_3 g} \end{pmatrix} ; e.g. \mathbf{\Lambda}^* = \begin{pmatrix} .6 & .3 & 0 & 0 & 0 \\ .8 & .4 & 0 & 0 & 0 \\ .4 & .2 & 0 & 0 & 0 \\ .6 & 0 & .4 & 0 & 0 \\ .4 & 0 & .267 & 0 & 0 \\ .6 & 0 & .4 & 0 & 0 \\ .8 & 0 & 0 & .2 & 0 \\ .6 & 0 & 0 & .15 & 0 \\ .8 & 0 & 0 & .2 & 0 \\ .4 & 0 & 0 & 0 & .5 \\ .4 & 0 & 0 & 0 & .5 \\ .5 & 0 & 0 & 0 & .625 \end{pmatrix} \quad (3)$$

In the following, parameters of the S-1c reparameterization are marked with a check ("ˇ"). The derivation that for all S^* with implied covariance matrix $\mathbf{\Sigma}$, there exists a model S-1c for which $\mathbf{\Sigma} = \check{\mathbf{\Sigma}}$ shows that the S^* model is nested within the S-1c model. Without loss of generality, $\mathbf{\Sigma}$ is first standardized to be a correlation matrix to simplify equations. Be r_{11} any correlation between two indicators of the reference domain ("1"), r_{xx} any correlation between two indicators of an arbitrary other domain x, r_{1x} any correlation between an indicator of the reference domain and an indicator of an arbitrary other domain x, and r_{xy} any correlation between indicators of two arbitrary other domains $x \neq y$. The respective values for k are labeled k_1, k_x, and k_y. The respective general factor loadings are labeled

$\lambda_{1_i g}$, $\lambda_{x_i g}$, and $\lambda_{y_i g}$ for the i'th (or j'th) indicator of the respective domain. The off-diagonal entries of $\mathbf{\Sigma} = \check{\mathbf{\Sigma}}$ are related to $\mathbf{\Lambda}$ and $\check{\mathbf{\Lambda}}$ as follows:

$$r_{1_i 1_j} = \lambda_{1_i g}\lambda_{1_j g} + k_1^2 \lambda_{1_i g}\lambda_{1_j g} = \check{\lambda}_{1_i g}\check{\lambda}_{1_j g}; i \neq j \tag{4}$$

$$r_{1_i x_j} = \lambda_{1_i g}\lambda_{x_j g} = \check{\lambda}_{1_i g}\check{\lambda}_{x_j g} \tag{5}$$

$$r_{x_i x_j} = \lambda_{x_i g}\lambda_{x_j g} + k_x^2 \lambda_{x_i g}\lambda_{x_j g} = \check{\lambda}_{x_i g}\check{\lambda}_{x_j g} + \check{\lambda}_{x_i s}\check{\lambda}_{x_j s}; i \neq j \tag{6}$$

$$r_{x_i y_j} = \lambda_{x_i g}\lambda_{y_j g} = \check{\lambda}_{x_i g}\check{\lambda}_{y_j g} + \check{r}_{\eta_x \eta_y}\check{\lambda}_{x_i s}\check{\lambda}_{y_j s} \tag{7}$$

Solving Equations (4)-(7) for the parameters of the S-1c variant (details see Appendix A) results in Equations (8)-(11), which provide the transformed parameter values of the S-1c variant as functions of the parameter values of the S^* model with the same implied correlation matrix:

$$\check{\lambda}_{1_i g} = \lambda_{1_i g}\sqrt{1 + k_1^2} > \lambda_{1_i g} \tag{8}$$

$$\check{\lambda}_{x_i g} = \frac{\lambda_{x_i g}}{\sqrt{1 + k_1^2}} < \lambda_{x_i g} \tag{9}$$

$$\check{\lambda}_{x_i s} = \lambda_{x_i g}\sqrt{k_x^2 + 1 - \frac{1}{1 + k_1^2}} > \lambda_{x_i s} \tag{10}$$

$$\check{r}_{\eta_x\eta_y} = \frac{1 - \frac{1}{1+k_1^2}}{\sqrt{k_x^2 + 1 - \frac{1}{1+k_1^2}}\sqrt{k_y^2 + 1 - \frac{1}{1+k_1^2}}} > 0 \tag{11}$$

Because $k_1, k_x, k_y > 0$, all parameters of the S-1c model are uniquely identified by this transformation, meaning that for each S^* there exists an S-1c parametrization. Therefore, S^* (the higher-order factor model) is nested in S-1c. This adds to the well-known fact that the higher-order factor model is nested in the S variant by proving the same relationship to the S-1c variant. It also implies a special relationship between the parameters of the S and S-1c variants. This may not only be relevant in exact special cases but more generally to the degree that data resemble the S^* case.

Understanding the relationship between the model variants is crucial for the interpretation of bifactor models because the two different parametrizations of the same S^* model lead to different conclusions about the existence, relationship between, and meaning of domain factors. The S-1c and S parametrization provide different estimates of the same parameters in equivalent solutions, as indicated by the inequalities at the end of Equations (8)-(11). Furthermore, the variants provide different estimates of the general trait. The extent to which the S^* proportionality constraint is violated may be of limited practical importance (Raykov et al., 2022) in many applications. It is unclear how well the S-1c and S variants can generally compensate for their relative restrictions with the additional free parameters in other parts of the model (cf. Figure 1). Therefore, the following simulation study on a general set of cases (beyond S-1 and S^*) 1) checks how well the S and S-1c variants can be distinguished by common fit-indices, and 2) examines the differences in trait and parameter estimates between the S and S-1c variants estimated on the same data.

Simulation

Methods

Using R (R Core Team, 2020) and SimDesign (Version 2.0.1, Chalmers & Adkins, 2020), I generated data from each bi-factor model variant (Figure 1).[4] Each dataset contains four domains (as in Figure 1). The trait values were drawn from a multivariate standard normal distribution (using mvtnorm Version 1.1-1, Genz et al., 2020). Random error was added to reach $\sigma^2_Y = 1$ for each indicator. Therefore, all parameters are fully standardized. In the S-1 and S-1c variants, the first domain trait was omitted and its variance was replaced by additional random error. In the S-1c variant, domain factor correlations were either sampled or set to a specific value (see below). Sample size, factor loadings, and the number of indicators per domain were varied in a range typically observed in psychological measurement (Petras & Meiser, 2023).

Simulation A

In the main simulation, the reliabilities of individual indicators ($Rel(Y)$), the specific proportion of their reliable variance ($\lambda^2_s/Rel(Y)$), and the domain factor correlations in the S-1c variant were drawn from (independent) beta-distributions.[5] The parameters were chosen to cover a realistic range of possible scenarios (Table 1).[6] The distribution of the absolute values of the domain factor correlations had a mean of .3 and a standard deviation of .12 ($r \sim Beta(\frac{163}{40}, \frac{1141}{120})$). The sign of each domain factor correlation was randomized with a 50% chance of being negative. This results in a total of 192 conditions (Table 1) with 1008

[4] The simulation code, results, and the reproducible manuscript are available in the osf.io repository: https://osf.io/e9c6f/?view_only=7e61e52c2a664a32826967045e5cbf34 (this is an anonymized peer review link)

[5] $\lambda^2_s/Rel(Y) = k^2$ is true for the S^* model, but in the simulation this proportion is varied across indicators with a standard deviation of $\sigma(\lambda^2_s/Rel(Y))$.

[6] An overview of the parameters of the beta distributions, a figure of their density functions, and the resulting factor loadings are in the online supplement

iterations per condition.

(Table 1)

Simulation B

To further examine changes in parameter and trait estimates when switching between bi-factor model variants, in an additional simulation, model parameters were set to specific values instead of drawn from distributions. To distinguish the roles of correlated and uncorrelated domains within the same dataset, in the S-1c variant, the correlation between two domain factors was non-zero, while the fourth domain was uncorrelated with the others. There were a total of 120 simulation conditions (Table 1). Within the S-1c variant conditions, the correlation between the second and third domain factors was varied in three steps (.2, .5, .8). This resulted in 1008 iterations for each condition of the S-1c variant and 3024 iterations for the other conditions. Because all indicators share the same factor loading values, there is a constant k so that $\lambda_s = k\lambda_g$ across the whole model. Therefore, the data-generating S model in Simulation B is the S^* model (with $k_1 = k_2 = k_3 = k_4$) and thus nested in the S-1c model.

Analysis

Each model variant was fitted on each generated dataset using `lavaan` (Version 0.6-7, Rosseel, 2012) for Maximum Likelihood (ML) estimation. For model identification, the first loading on each factor was set to one. In the S-1c variant, all domain factor correlations were freely estimated.

For each estimated model, the following fit indices were computed to examine model selection[7]: The Standardized Root Mean Square Residual (SRMR, all fit index definitions in online supplement) statistic measures raw misfit of the model implied correlation matrix independently of sample size or model parsimony. In contrast, the Root Mean Squared Error

[7] These similar indices are computed but not analyzed here to avoid redundancy: model chi-square (χ^2_M), comparative fit index (CFI), Tucker-Lewis index (TLI), normed fit index (NFI)

of Approximation (RMSEA) accounts for parsimony. The Akaike and Bayesian Information Criteria (AIC and BIC) use the likelihood of the data given the estimated model and account for parsimony in slightly different ways.

Based on these fit indices, model selection rates between the three alternative model variants are analyzed below. Since all indices measure misfit, the model with the lowest value is coded as selected. On some iterations, no model is selected, because all models failed to converge[8]. To safeguard against large absolute misfit, the proportion of selected models with $SRMR > .08$ or $RMSEA > .06$ (Hu & Bentler, 1999) was computed.

To examine the consequences of misspecifying the model variant, the mean bias of factor loadings per domain was computed. To judge the recovery of the original traits, the correlation between the data-generating trait and the corresponding trait estimates (Regression factor scores, default in `lavaan`, DiStefano et al., 2009; Rosseel, 2012) was computed. To analyze the composition of estimated factor scores, especially when estimating one model variant on data generated from another, the correlations between all data-generating traits and all estimated factor scores were computed. Even without misspecification, factor scores are contaminated with other traits. Therefore, the results for the true model are presented as a benchmark.

Results

Model fit

A non-converged model can not fit the data, so convergence rates are considered first (Figure 3). When estimating the S variant on data from the S-1 or S-1c variants, convergence was imperfect under all simulation conditions. The average convergence rate of S models was

[8] A model is counted as converged if the `lavaan` package indicated convergence and at least one model parameter significance test was successfully computed. Warnings indicated that some "converged" solutions may not be properly interpretable, which could not be checked, given the $\approx 600,000$ estimated models in total.

56.39% on S-1 data and 69.54% on S-1c data. Vice versa, S-1 models converged in 95.94% and S-1c models in 97.78% of cases on S data. That estimating too many factors is more problematic than estimating too few introduces a bias: obtaining only S-1 or S-1c estimates on data from an S population is more likely than vice versa.

For all fit indices, it is much more likely that the S-1c variant fits the data generated by the S variant better than the true (S) model, than vice versa (Simulation A, Table 2). This discrepancy in model flexibility can be observed for both small ($n = 200$) and large ($n = 2000$) samples, although fewer errors were made on large samples in general. Rewarding parsimony more strongly (AIC vs. BIC) increases this effect. The RMSEA produces the fewest errors, due to ties at $RMSEA = 0$ on the relatively clean simulated data.[9] The best fitting model had a bad fit ($SRMR > .08$ or $RMSEA > .06$) in 0% of the cases.

The error rate when deciding for the model that fits the data best varies across conditions (Figure 4). Whereas 64.06% of conditions with S data had a modest error rate of less than 5%, this rate was 87.30% in the worst condition. Accounting for parsimony explains some of that: In the condition with six indicators per domain, the S-1c variant uses three parameters less than the S variant and therefore is more often preferred by fit indices (top right of Figure 4, compared to bottom left).

A bias exists independent of parsimony. To judge the importance of the conditions, the SRMR[10] advantage of the data-generating model ($\Delta SRMR$) is regressed on the simulation parameters (linear, no interactions). The importance of a predictor is judged by the difference in R^2 between the model including all predictors and the model leaving out the predictor of interest. For claims exclusively concerning S or S-1c data, the models were restricted to the respective subset of data. All reported effects are significant with $p < .001$.

[9] Across all conditions, 13.81% of cases were ties at $RMSEA = 0$. These ties occur because of the $max(..., 0)$ rule (equation in online supplement).

[10] Decisions based on χ^2_M are similarly biased.

Figure 5 shows the distribution of $\Delta SRMR$ across individual repetitions. Gray areas indicate a worse fit of the data-generating variant. For the S-1c variant (red), the advantage in model fit is generally larger ($M_S = .0048$, $M_{S-1c} = .014$, $\Delta R^2 = .137$) (and more likely positive, cf. Table 2). This indicates an overall greater flexibility of the S-1c variant. Differences in the areas under the curves are caused by varying rates of non-convergence, especially of the S model on S-1c data (see Figure 3).

The less similar S data (black) are to the special case S^*, the better the variants can be discriminated. The defining feature of S^* is the zero variance in the size of the specific variance proportion within each domain. The higher this variance is, the larger the model fit advantage of the S variant ($M_{.06} = .0025$ (black lines in rows 1 and 3), $M_{.12} = .0072$ (black lines in rows 2 and 4), $\Delta R^2 = .133$).

Furthermore, increasing the sample size (top vs. bottom half, $\Delta R^2 = .053$), the specific proportion of reliable variance (left vs. right half, $\Delta R^2 = .097$), and the reliability of the indicators (columns 1 and 3 vs. 2 and 4, $\Delta R^2 = .157$) increases discriminability. The increase in the variance of the specific factors amplifies the difference between the variants. The effect of the number of indicators is specific to S data, in which more indicators increase discriminability ($\Delta R^2_S = .057$, $\Delta R^2_{S-1c} = .001$). There was no effect of the variance in indicator reliability ($p = .58$).

In sum, fit indices show that the S and S-1c variants can closely imitate each other, with the S-1c variant being more flexible. Both the fit indices that are sensitive to parsimony and those that are insensitive to parsimony show a bias towards the S-1c variant. Depending on several of the data-generating model parameters, the chance that the data-generating variant fits worse than the other can be quite high and even surpass 50%. The closer the data-generating model parameters are to the special case of the higher order model (S^*), the less distinguishable the S and S-1c variants are on model fit indices.

Differences in estimates of the S and S-1c variants

Model parameters. The S and S-1c bi-factor model variants share many comparable, freely estimated parameters (Figure 1). When estimating the S-1c model on S data (Figure 6), the average factor loading estimates consistently change in the same direction as in the special case S^* (Equations (8) to (11)). The loadings on the reference domain factor are set to zero and therefore decrease. To compensate, the loadings of the reference domain indicators on the general factor increase. On the further domains, the loadings on the domain-specific factors increase, whereas the loadings on the general factor decrease. The domain factor correlations of the S-1c model on S data are consistently and substantially positive (Figure 7).

Figure 8 shows the differences in factor loadings when estimating the S model on S-1c data in Simulation B. On the aggregate level of Simulation A, the average loading on the added reference domain factor is substantially above zero, but the other loadings do not change. Simulation B clarifies that the correlations between domain traits of the data-generating S-1c model systematically affect the estimates of the S variant. In the data-generating S-1c model, only two domain traits (η_2 and η_3) are correlated. Whereas the loadings of the positively correlated domains decrease on their specific factors and increase on the general factor, the opposite can be observed for the orthogonal domain. This confounds the general trait with those domain traits that are correlated in the data-generating model and leads to an underestimation of the variance of these correlated domain traits. The higher the correlation (color-coded in Figure 8) and the higher the domain-factor loadings, the more pronounced this effect is.

Factor scores. Differences in the model parameters inevitably result in differences in estimated factor scores. The baseline pattern in the S model variant shows that the factor score computation cannot disentangle traits in bi-factor models completely (Simulation A, Table 3, top left). The averaged correlations off the main diagonal are consistently and substantially non-zero.

The top right section of Table 3 shows the correlations between the same true S traits and the estimated S-1c factor scores. The unique variance of the reference domain (s1 row), which is set to zero in the S-1c model, affects all factor scores substantially. Nevertheless, interpreting the general factor score (g column) as the true score of the reference domain's indicators is supported by the simulation results: The estimated general factor scores are almost independent of the remaining domain traits η_{2-4}. In addition, some variance of the data-generating general trait is captured by the remaining domain factor scores (first row) instead. This pattern may not be obvious from the model's definition but follows directly from the differences in factor loadings (Figure 6, see also Equations (9) and (10)). The trait composition of the S-1 variant on S data only shows minor quantitative differences from that of the S-1c variant.

The S-1c variant also produces biased factor scores at baseline already (Simulation B, Table 3, bottom right). Compared to the orthogonal domain factor four, the correlation between the domain traits two and three in the population markedly increases the bias in their factor score estimates. The S variant compensates for the unmodeled pattern of correlations between the true domain traits in various ways (Table 3 bottom left): An additional "ghost" domain s1, which does not exist in the data-generating model, systematically draws from the true general trait and the true domain traits. Specifically, it is more strongly related to the correlated true domain traits (s2 and s3) than to the domain trait that is orthogonal to the others (s4). The general factor scores are similarly biased towards the correlated domain traits. In Simulation A, this pattern is canceled out in the averaged results due to averaging across the symmetrical distribution (with mean zero) of the domain trait correlations.

Discussion

The relationship between the bi-factor model variants is only incompletely described by the nesting structure. The S-1 variant and the higher-order model (here: S^*) are both nested in the S and S-1c variants (Figure 2). The relationship between the S and S-1c variants was further analyzed in the simulation study. The S and S-1c variants of the bi-factor model are characterized by free parameters in different parts of the model. Whereas the older S variant includes a domain-specific factor of the reference domain, the S-1c variant estimates correlations between the other domain factors freely (Figure 1). These additions in different parts of the model are equivalent if the ratio between general and specific loadings is constant within domains. This special case S^* is a reparameterization of the higher-order factor model (Yung et al., 1999), meaning that it is nested in both the S and S-1c variants of the bi-factor model. Critically, these different parameterizations superficially suggest a very different structure of the data. Beyond the special case S^*, the S and S-1c variants can compensate for unmodeled complexity from the part of the model that is specific to the other variant. This can be seen in the simulation results on the discriminability using fit indices and the differences in parameter estimates.

Comparing the model fit of the S and S-1c variants uncovers their ability of mutual imitation. Depending on the population parameters, deciding if the S and S-1c model variants underlie the population using common fit indices is uncertain or even impossible (Table 2, Figure 5). In addition, both convergence rates (Figure 3) and standard model fit indices (Table 2) systematically favor the S-1c variant. This suggests that it is more flexible to fit any data – including random sampling variation. Given that bi-factor models already show high flexibility compared to competing models (Bader & Moshagen, 2022), meaning they excel in fitting any data (Bonifay & Cai, 2017), fit-based decisions on the S-1c bi-factor model should be interpreted with caution (Roberts & Pashler, 2000). An excellent fit of the bi-factor model does not exclude the possibility that another model (variant) is true in the population, but the bi-factor model can imitate it. This applies to fit indices accounting for parsimony (AIC, BIC, RMSEA), and measures of absolute misfit (SRMR, χ^2_M).

The relationship between the S and S-1c variants can be seen from two different perspectives. To the degree that the S model estimates conform to the S^* (or S-1) restriction, the S and S-1c variants are interchangeable reparametrizations: none is closer to the truth than the other. To the degree that the S and S-1c variants differ in their model-implied covariance matrix, they can be more true or false. Therefore, differences in parameter values when switching between model variants can be seen either as a shift in perspective or as bias (= error). This is especially important when interpreting the unique parts of the model variants, but affects all parameters (Figures 6 to 8).

Does a domain have unique variance (i.e. there exists a domain-specific trait)? Are the domain factors correlated? These questions can not be answered generally or independently. The answer can depend entirely on the parameterization. The S^* case shows that these questions can be identical: the covariance matrix contains information that can be perfectly represented by allowing either the domain factor correlations or the parameters pertaining to a specific factor of the reference domain to be freely estimated. Strong compensatory changes in parameter estimates when switching between model variants are common on a larger variety of data beyond the S^* special case (e.g., Figure 7). Many claims on these estimates in the literature may therefore reflect implicit choices on which model variants to consider instead of the phenomenon of interest.

Researchers before 2017 decided in favor of the S variant by default. A researcher claiming that there exists a specific trait may then miss that a model without that trait, but including correlated further domain factors, would make the exact same predictions. If researchers decide on principle in favor of the S-1c variant, because the domains are not randomly sampled (Eid et al., 2017), they should be aware of the consequences. For example, a researcher claiming that two domain traits are better understood as being correlated because the S-1c model fits much better than the S-1 model may miss that a third model with orthogonal domain factors (S) would make the exact same predictions. An easily overlooked detail is that any unique variance of the reference domain in the population affects all S-1 and S-1c parameter estimates, including the scores on the remaining domain factors (Table 3, second row). This seems to contradict the interpretation given by (Eid et al., 2017, p. 550): "Such a specific factor represents that part of a domain that is not shared with the reference domain."

In the multitrait-multimethod (MTMM) literature, models that leave out one specific factor have been proposed first (Eid, 2000; Eid et al., 2003, 2008). MTMM models typically comprise multiple traits, but the simplified version with a single trait can be identical to the bi-factor model. Similarly to the current study, Geiser et al. (2015) analyzed the relationship between such models and models with a full set of method factors in a simulation. The UM(unconstrained) and C(M-1) models of that study are identical to S and S-1c bi-factor variants (Geiser et al., 2015, fig. 3). Mathematically, what MTMM analysis calls method factors then are domain-specific factors in bi-factor models, and their single trait is the same as the general factor of a bi-factor model. Geiser et al. (2015) found a tendency of the S variant on S-1c data to produce non-significant method factor loadings. The prevalence of these null results increased with the correlation between the data-generating method factors (Geiser et al., 2015, fig. 5). This perfectly matches the finding of decreased specific factor loadings in S models estimated on S-1c data (Figure 8). Geiser et al. (2015) attributed this finding to the badly modeled interchangeability (= random sampling) of the methods (p. 13). However, their C(M-1) that generated the data and the UM(unconstrained) model that was estimated were both specified for non-interchangeable methods (which in that case meant that different factor loading patterns were allowed across methods). Furthermore, in the current simulation, the issue of decreasing specific factor loadings when estimating the S model variant applies to the correlated specific factors of the data-generating S-1c model, but not its orthogonal factors (Figure 8). Therefore, I argue that the described problem of "collapsing" factors with non-significant loadings appears more frequently if there are more strongly correlated specific traits in the data-generating population. When the S model is estimated on such data, the model compensates for fixing the strong correlation to zero by decreasing the factor loadings of (only) the affected domains, among other changes. In consequence, the general trait is confounded with the specific variance of the correlated methods (or domains). This can be partly understood as a perspective shift and does not necessarily indicate an error by the researcher. Due to connections like these, it would be fruitful for future research to link the literature on MTMM models and bi-factor models more closely. This link becomes clearer when understanding the proposal of S-1 bi-factor models (Eid et al., 2017) as a special case of the method-1 approach to MTMM analysis (Eid et al., 2008).

The current simulation did not consider the relationship between latent variables directly. If the interpretation of the factors in the structural part of a structural equation model follows exactly the pattern of results on factor scores is therefore uncertain. Nevertheless, it seems plausible that the pattern of differences between variants is consistent because the factor scores differ due to systematic differences in the factor loadings. Examining the interpretation of different bi-factor model variants within structural equation models would be relevant to future research because the S-1(c) variants show promising characteristics when the bi-factor model is in the predictive part of a structural equation model (Eid et al., 2018; Zhang et al., 2020).

Conclusion

The relationship between the bi-factor model variants is complex and can be misleading. The variants of the bi-factor model exemplify how parameters in one part of a statistical model can partly – or even fully – account for unmodeled complexity in other parts of the model. The higher-order factor model (S^*) was identified as a special case, which is nested in both the S and S-1c bifactor model variants. On data generated from a S^* population, the additions of another domain-specific factor and the addition of correlations among the other domain-specific factors are interchangeable reparameterizations of the same model. This relationship is retained to varying degrees in all data generated by bi-factor population models. Therefore, the validity of many claims is conditional on the model variant in a subtle way and data-based model selection between variants is severely limited. Beyond providing a more comprehensive basic understanding of the bi-factor model variants, the current work offers a reference on how parameter estimates behave based on the choice of model variant.

References

Bader, M., & Moshagen, M. (2022). No probifactor model fit index bias, but a propensity toward selecting the best model. *Journal of Psychopathology and Clinical Science*, *131*(6), 689–695. https://doi.org/10.1037/abn0000685

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(20). https://doi.org/10.7275/da8t-4g52

Dueber, D. M., & Toland, M. D. (2021). A bifactor approach to subscore assessment. *Psychological Methods*. https://doi.org/10.1037/met0000459

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*(2), 241–261. https://doi.org/10.1007/BF02294377

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562.

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence*, *6*(3), 42. https://doi.org/10.3390/jintelligence6030042

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-c (m-1) model. *Psychological Methods*, *8*(1), 38–60. https://doi.org/10.1037/1082-989X.8.1.38

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*(3), 230–253. https://doi.org/10.1037/a0013219

Geiser, C., Bishop, J., & Lockhart, G. (2015). Collapsing factors in multitrait-multimethod models: Examining consequences of a mismatch between measurement design and model. *Frontiers in Psychology*, *6*, 946. https://doi.org/10.3389/fpsyg.2015.00946

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2020). *mvtnorm: Multivariate normal and t distributions*. https://CRAN.R-project.org/package=mvtnorm

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., Waldman, I. D., Cicero, D. C., Conway, C. C., Docherty, A. R., Fried, E. I., Ivanova, M. Y., Jonas, K. G., Latzman, R. D., Patrick, C. J., Reininghaus, U., Tackett, J. L., Wright, A. G. C., & Kotov, R. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *Journal of Abnormal Psychology*, *1939-1846(Electronic),0021-843X(Print)*, 740–764. https://doi.org/10.1037/abn0000434

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*, 51–69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

Petras, N., & Meiser, T. (2023). Problems of domain factors with small factor loadings in bi-factor models. *Multivariate Behavioral Research*, 1–25.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raykov, T., DiStefano, C., Calvocoressi, L., & Volker, M. (2022). On effect size measures for nested measurement models. *Educational and Psychological Measurement*, *82*(6), 1225–1246. https://doi.org/10.1177/00131644211066845

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358. https://doi.org/10.1037/0033-295X.107.2.358

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61. https://doi.org/10.1007/BF02289209

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25–60.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128. https://doi.org/10.1007/BF02294531

Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2020). Using bifactor models to examine the predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational Research Methods*, 530–571. https://doi.org/10.1177/1094428120915522

Table 1

Simulation design (top: A, bottom: B)

parameter	values	description
n	200, 2000	sample size
$E[Rel(Y)]$	.4, .7	average indicator reliability
$\sigma(Rel(Y))$	.06, .12	indicator reliability standard deviation
$E[\lambda_s^2/Rel(Y)]$	.2, .5	average proportion of domain-specific reliable indicator variance
$\sigma(\lambda_s^2/Rel(Y))$	.06, .12	standard deviation of domain-specific reliable indicator variance
m	3, 6	indicators per domain
model variant	S, S-1, S-1c	
n	200, 2000	sample size
λ_g	.5, .7	general factor loading
λ_s	.2, .4, .6	domain factor loading
m	3, 6	indicators per domain
model variant	S, S-1, S-1c	
$r_{\eta_2\eta_3}$	.2, .5, .8	correlation between domain traits 2 and 3 in S-1c variant

Note. Simulation A: Total of 192 conditions ($2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 3$). Data was generated by drawing from beta distributions defined by the parameters. Simulation B: Total of 120 conditions ($2 \times 2 \times 3 \times 2 \times (2 + 1 \times 3)$).

Table 2

Simulation A: Erroneous decisions between the S and S-1c variants. Numbers indicate the percentage of all cases in which the true model (column) had the higher fit index (= more misfit). These values can not exceed the percentage of cases in which both variants converged (bottom row).

	n = 200		n = 2000	
	S	S-1c	S	S-1c
BIC	34.55	2.37	9.33	0.77
AIC	21.39	4.44	4.38	0.93
SRMR	22.02	4.13	4.87	0.91
RMSEA	11.92	4.19	2.49	0.62
both converged	85.15	55.68	99.11	81.05

Table 3

Average correlations between true trait values (rows) and estimated factor scores (columns); for Simulation B (lower part), only the data with a medium true correlation of .5 between η_2 and η_3 is included for simplicity

	S					S-1c			
	g	s1	s2	s3	s4	g	s2	s3	s4
Simulation A: S data									
η_g	**.901**	.181	.181	.181	.181	**.81**	.376	.376	.376
η_1	.147	**.677**	-.145	-.145	-.145	.419	-.324	-.324	-.324
η_2	.147	-.145	**.678**	-.145	-.145	.044	**.607**	.02	.02
η_3	.147	-.145	-.145	**.678**	-.145	.044	.02	**.608**	.02
η_4	.147	-.145	-.145	-.145	**.678**	.044	.02	.02	**.607**
Simulation B: S-1c data									
η_g	**.891**	.219	.098	.098	.202	**.91**	.179	.179	.156
η_2	.256	-.236	**.538**	.092	-.231	.152	**.586**	.302	-.156
η_3	.256	-.236	.091	**.539**	-.231	.152	.301	**.586**	-.156
η_4	.095	-.085	-.129	-.129	**.614**	.13	-.155	-.155	**.599**

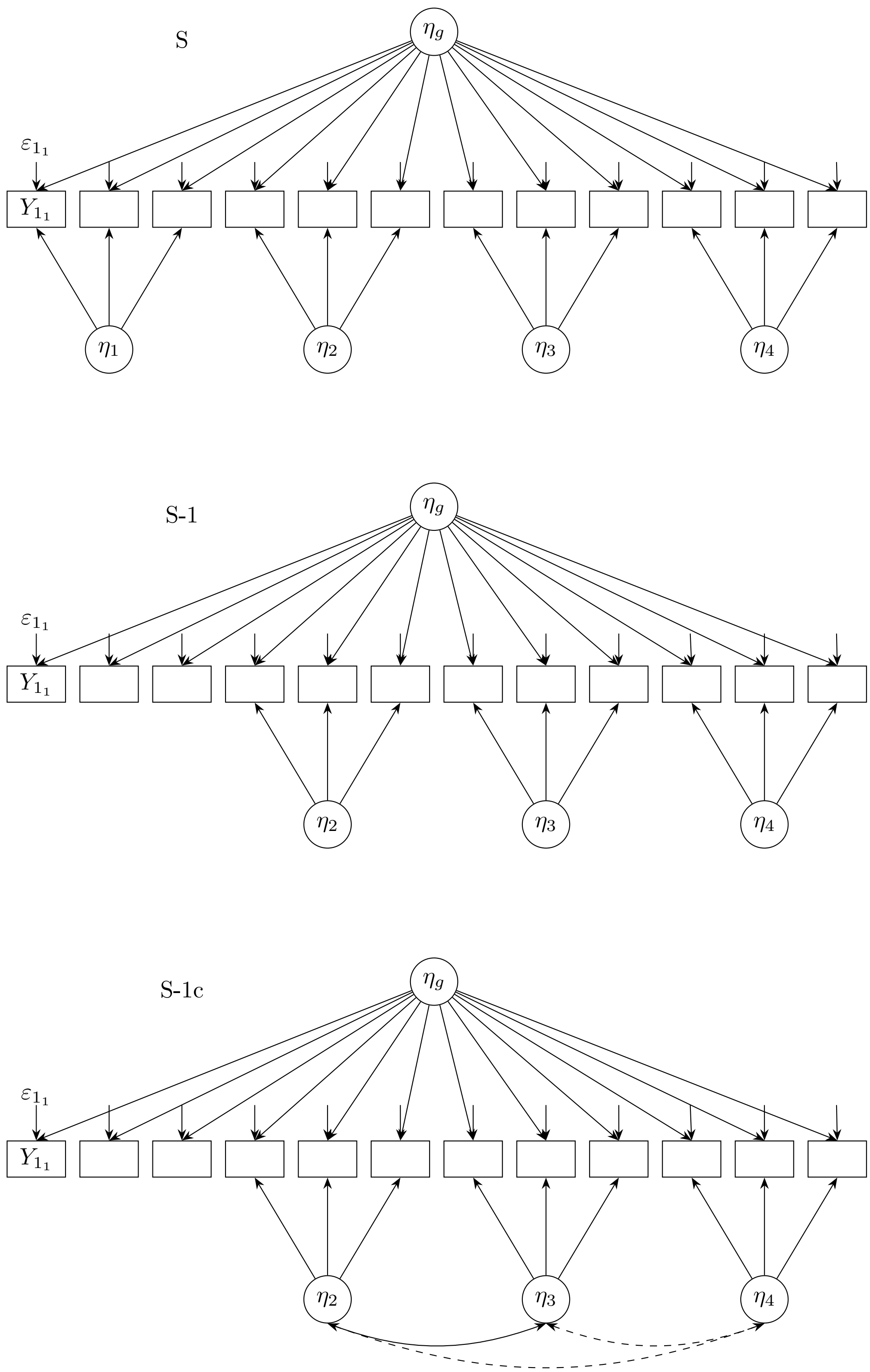


Figure 1

Bi-factor model path diagrams of the S, S-1, and S-1c variants. There are one general factor η_g and – in this example – up to four domain factors (η_{1-4}).

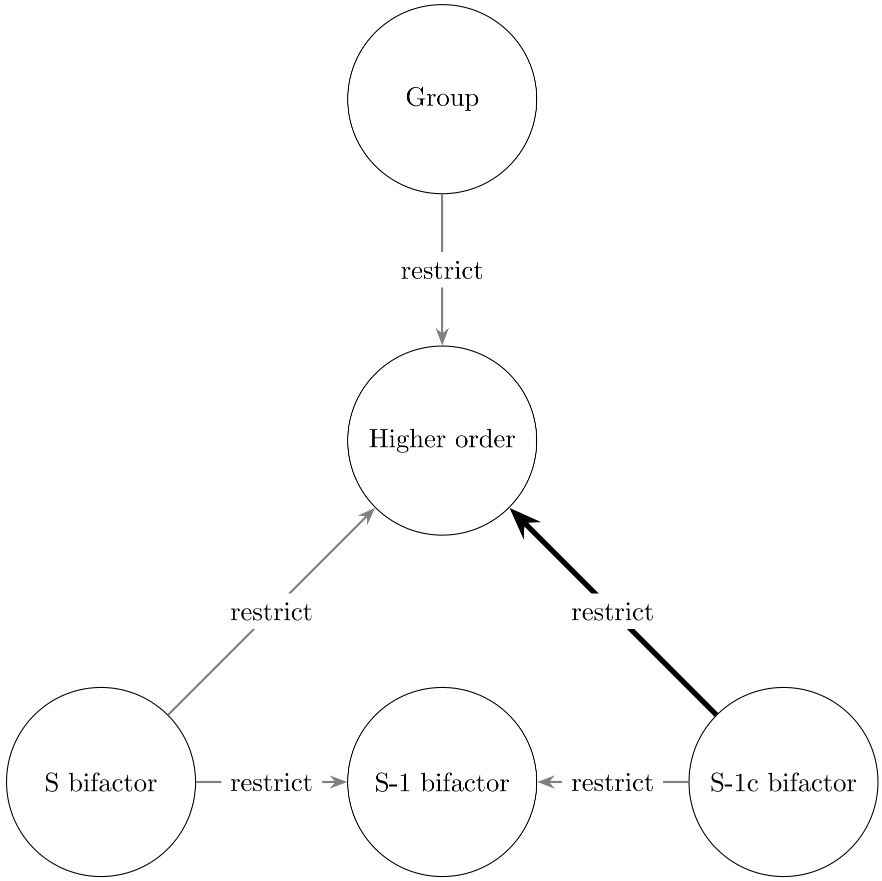


Figure 2

Nesting structure of bi-factor model variants and related models.

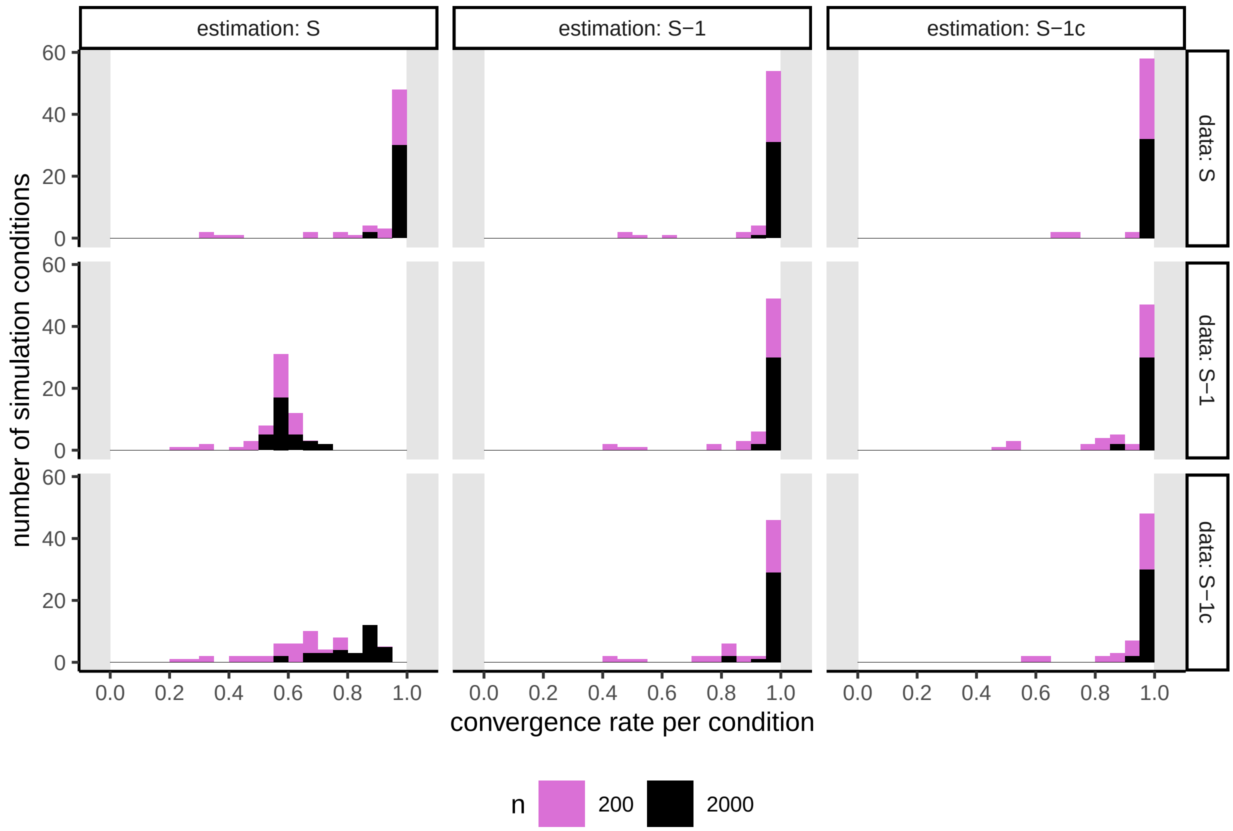


Figure 3

Simulation A: Convergence rate of each simulation condition by model variant combination. Stacked bars indicate the number of simulation conditions with a certain convergence rate (x-axis). Correctly specified models are on the main diagonal, estimates with a mismatch between the data-generating variant and the estimated variant are off the main diagonal.

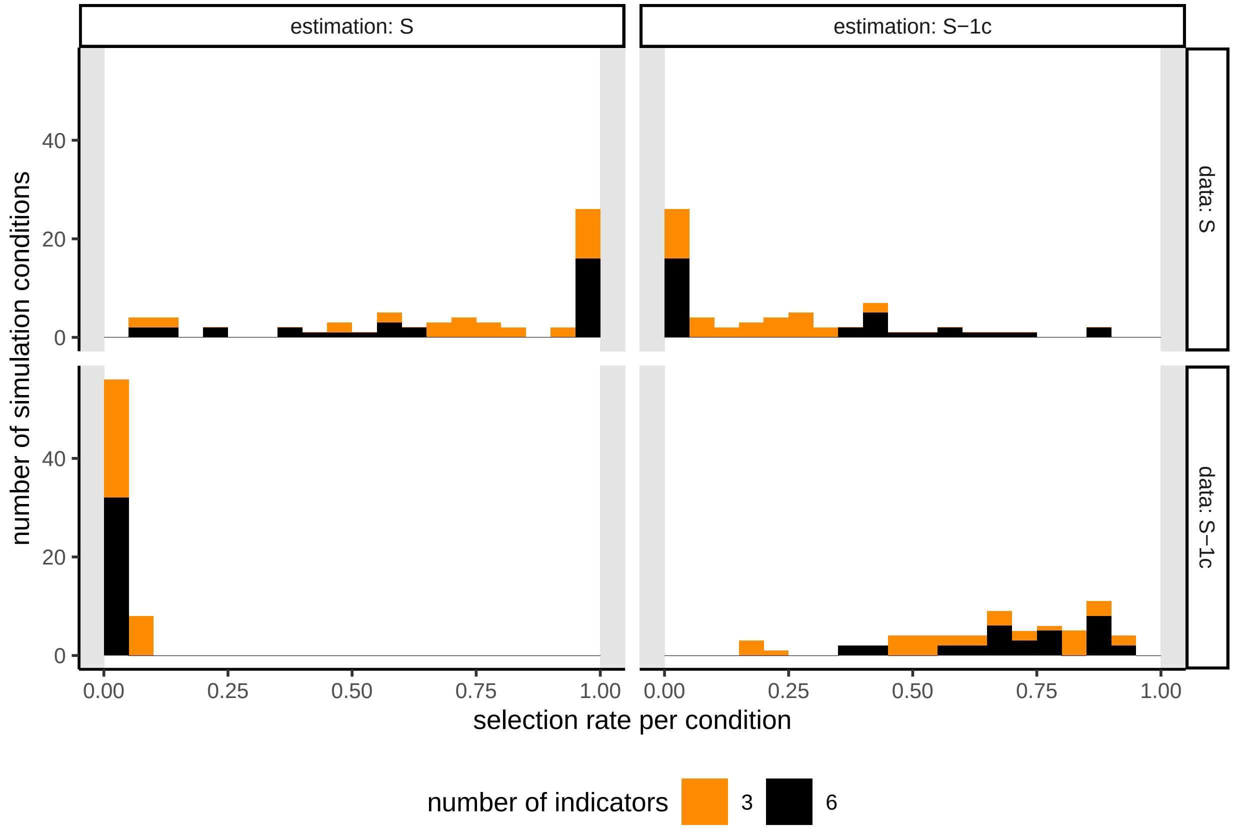


Figure 4

Simulation A: Selection by best model fit (BIC) and model variant combination. Stacked bars indicate the number of simulation conditions with the given selection rate for the estimated model. The estimated variant matches the data-generating variant on the main diagonal. (The S-1 variant is omitted for brevity but considered in the analysis - so that selection rates in the figure do not add up to convergence rates.)

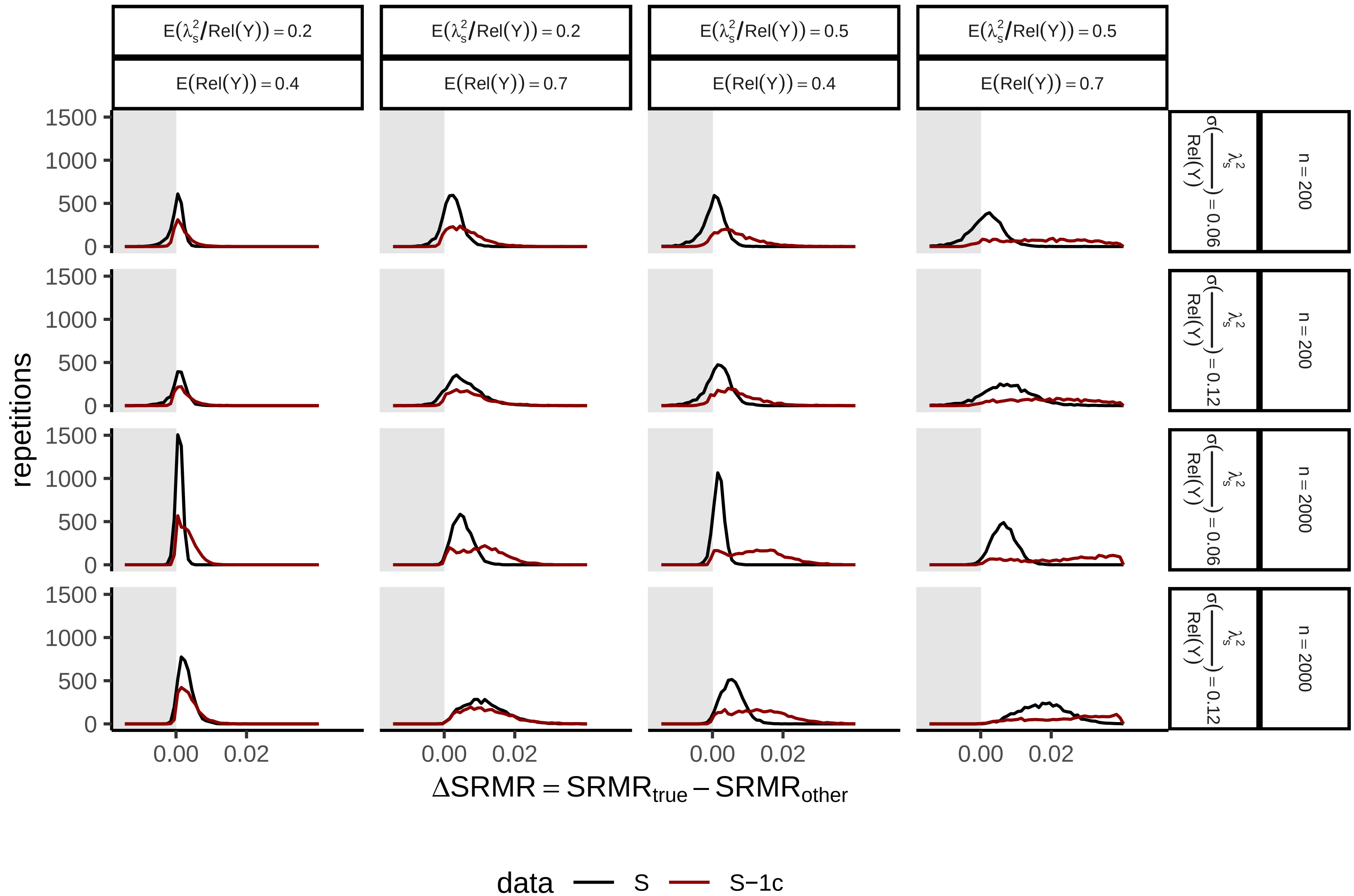


Figure 5

Simulation A: Difference in SRMR between S and S-1c model variant: Gray areas mark cases in which the variant that is true in the population (data-generating model) fits worse.

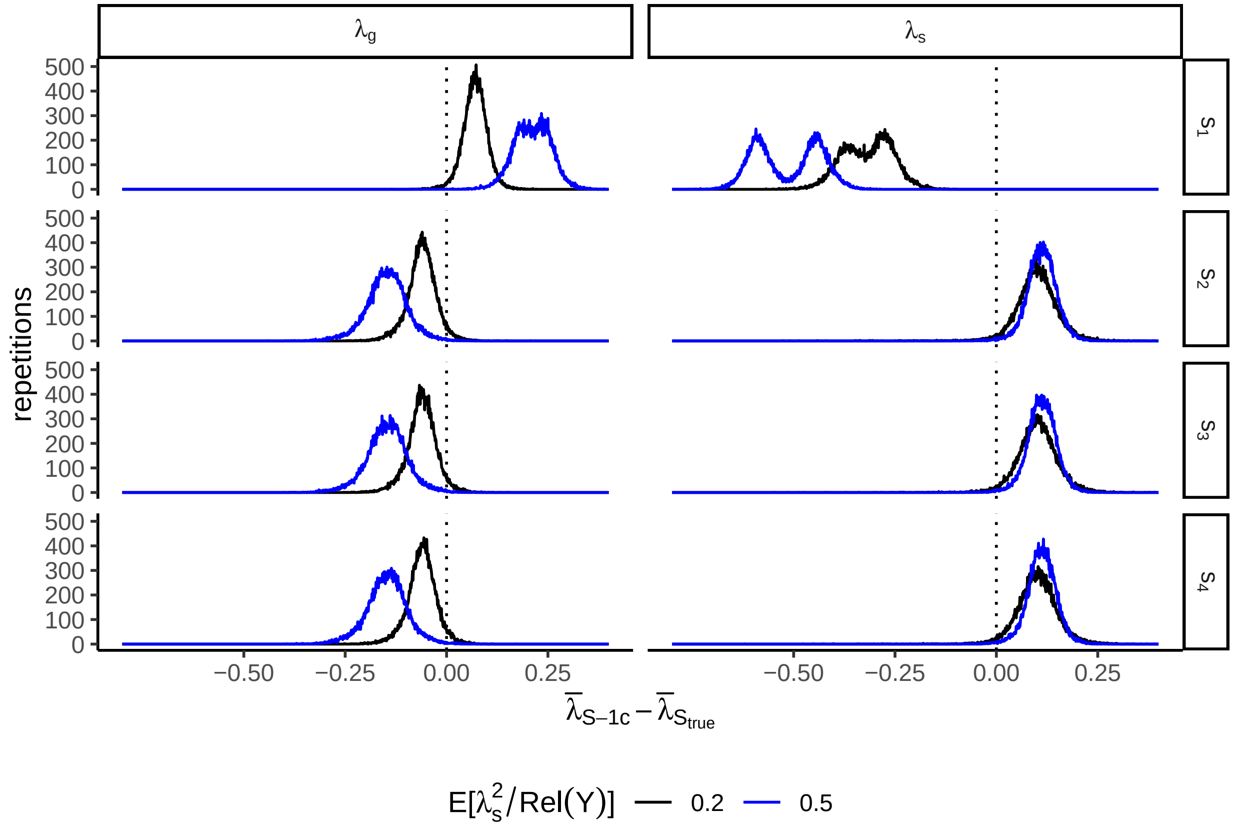


Figure 6

Simulation A: Difference in general (λ_g) and specific (λ_s) factor loading estimates when estimating the S-1c model on S data. Values are averaged across indicators within each domain s_1-s_4 and compared to the true S population values. s_1 is the reference domain. $E[\lambda_s^2/Rel(Y)]$ is the average specific proportion of the reliable variance of all indicators (see Table 1).

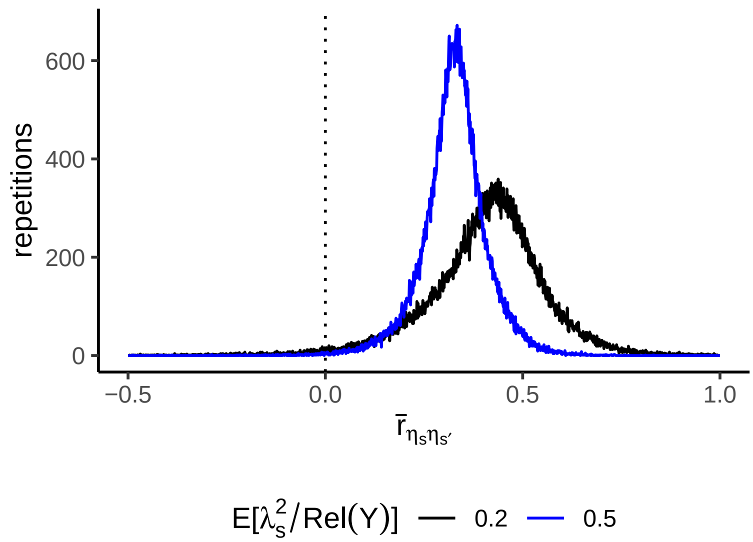


Figure 7

Simulation A: Difference in domain factor correlation estimates when estimating the S-1c model on S data. Since all domain correlations are 0 in the true population model, this difference is equal to the S-1c domain factor correlation estimates $\bar{r}_{\eta_s\eta_{s'}}$. $E[\lambda_s^2/Rel(Y)]$ is the average specific proportion of the reliable variance of all indicators (see Table 1).

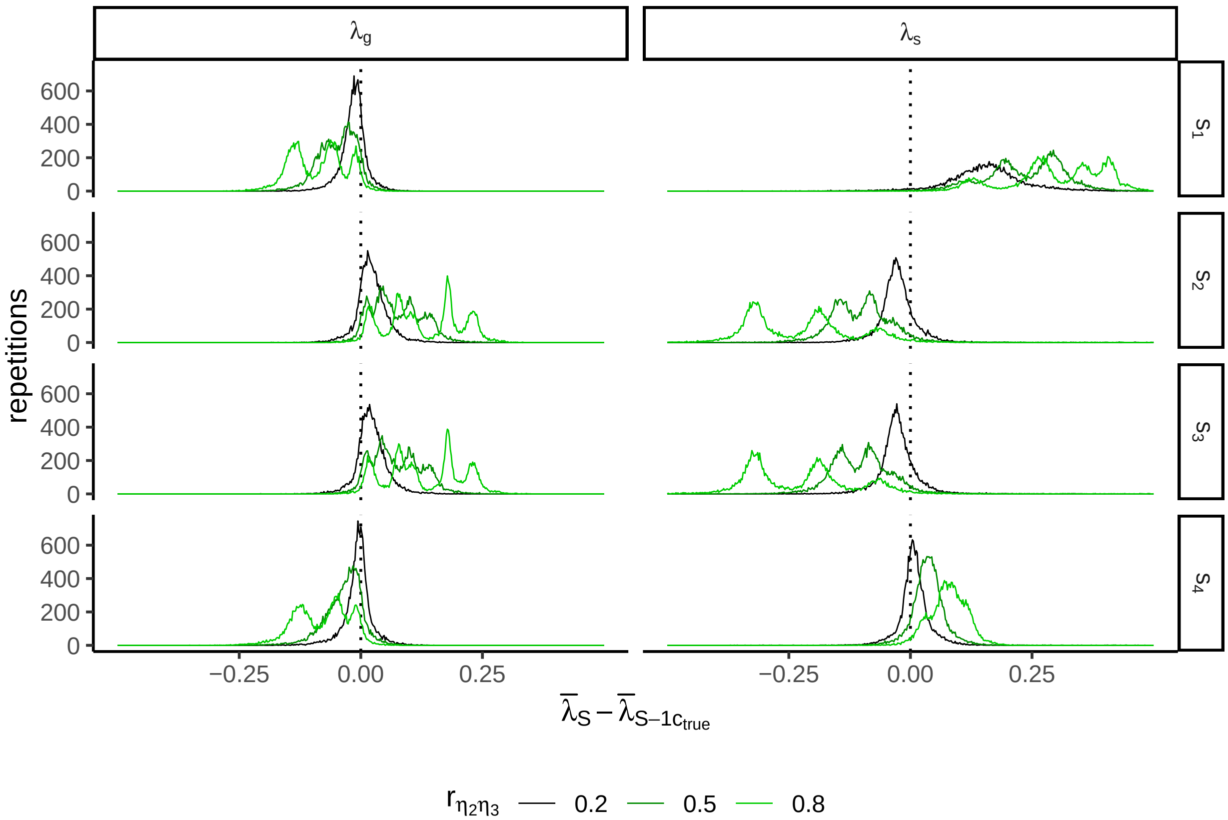


Figure 8

Simulation B: Difference in general (λ_g) and specific (λ_s) factor loading estimates when estimating the S model on S-1c data. Values are averaged across indicators within each domain s_1-s_4 and compared to the true S-1c population values. s_1 is the reference domain. The domain specific trait η_4 of s_4 is orthogonal to the others in the population and $r_{\eta_2\eta_3}$ is the correlation between the other domain traits.

Appendix

Derivation of Equations (8) to (11)

First, Equation (4) on the correlations within the reference domain is used to obtain the general factor loadings of the reference domain in the transformed S-1c model. Restructure Equation (4):

$$r_{1_i1_j} = \lambda_{1_ig}\lambda_{1_jg} + k_1^2\lambda_{1_ig}\lambda_{1_jg} = \check{\lambda}_{1_ig}\check{\lambda}_{1_jg}; i \neq j \tag{A1}$$

$$\check{\lambda}_{1_ig}\check{\lambda}_{1_jg} = \lambda_{1_ig}\lambda_{1_jg}(1 + k_1^2) = \lambda_{1_ig}\lambda_{1_jg}(1 + k_1^2) \tag{A2}$$

Solve for the individual parameters (see also Equation (8)):

$$\begin{aligned} \check{\lambda}_{1_ig} &= \lambda_{1_ig}\sqrt{1 + k_1^2} \\ \check{\lambda}_{1_jg} &= \lambda_{1_jg}\sqrt{1 + k_1^2} \end{aligned} \tag{A3}$$

Next, the general factor loadings in the non-reference domains in the S-1c model are calculated by inserting Equation (A3) into Equation (5), which describes the correlations of the indicators of the reference domain with the other indicators.

Insert Equation (A3) into Equation (5):

$$r_{1_ix_j} = \lambda_{1_ig}\lambda_{x_jg} = \check{\lambda}_{1_ig}\check{\lambda}_{x_jg} \tag{A4}$$

$$\check{\lambda}_{1_ig}\check{\lambda}_{x_jg} = \lambda_{1_ig}\sqrt{1 + k_1^2}\check{\lambda}_{x_jg} = \lambda_{1_ig}\lambda_{x_jg} \tag{A5}$$

Solve for $\check{\lambda}_{x_jg}$ (see also Equation (9)):

$$\check{\lambda}_{x_j g} = \frac{\lambda_{x_j g}}{\sqrt{1 + k_1^2}} \tag{A6}$$

To obtain the specific factor loadings in the S-1c model, Equation (A6) is inserted into Equation (6), which describes the correlations between indicators within non-reference domains.

Insert Equation (A6) into Equation (6):

$$r_{x_i x_j} = \lambda_{x_i g}\lambda_{x_j g} + k_x^2 \lambda_{x_i g}\lambda_{x_j g} = \check{\lambda}_{x_i g}\check{\lambda}_{x_j g} + \check{\lambda}_{x_i s}\check{\lambda}_{x_j s}; i \neq j \tag{A7}$$

$$\lambda_{x_i g}\lambda_{x_j g}(1 + k_x^2) = \frac{\lambda_{x_i g}\lambda_{x_j g}}{1 + k_1^2} + \check{\lambda}_{x_i s}\check{\lambda}_{x_j s} \tag{A8}$$

Solve for $\check{\lambda}_{x_i s}\check{\lambda}_{x_j s}$:

$$\check{\lambda}_{x_i s}\check{\lambda}_{x_j s} = \lambda_{x_i g}\lambda_{x_j g}(1 + k_x^2 - \frac{1}{1 + k_1^2}) \tag{A9}$$

Similar to Equation (A2), Equation (A9) can be solved for the individual parameters like this:

$$\begin{aligned} \check{\lambda}_{x_i s} &= \lambda_{x_i g}\sqrt{1 + k_x^2 - \frac{1}{1 + k_1^2}} \\ \check{\lambda}_{x_j s} &= \lambda_{x_j g}\sqrt{1 + k_x^2 - \frac{1}{1 + k_1^2}} \end{aligned} \tag{A10}$$

To obtain the domain factor correlations in the S-1c model, Equations (A6) and (A10) are inserted into Equation (7), which describes the correlations between indicators of different non-reference domains.

Insert Equations (A6) and (A10) into Equation (7):

$$r_{x_i y_j} = \lambda_{x_i g}\lambda_{y_j g} = \check{\lambda}_{x_i g}\check{\lambda}_{y_j g} + \check{r}_{\eta_x \eta_y}\check{\lambda}_{x_i s}\check{\lambda}_{y_j s} \tag{A11}$$

$$\lambda_{x_i g}\lambda_{y_j g} = \frac{\lambda_{x_i g}\lambda_{y_j g}}{1+k_1^2} + \check{r}_{xy}\lambda_{x_i g}\sqrt{1+k_x^2-\frac{1}{1+k_1^2}}\lambda_{y_j g}\sqrt{1+k_y^2-\frac{1}{1+k_1^2}} \tag{A12}$$

In solving for $\check{r}_{xy}$, the absolute factor loadings are canceled out:

$$\check{r}_{\eta_x \eta_y} = \frac{1-\frac{1}{1+k_1^2}}{\sqrt{k_x^2+1-\frac{1}{1+k_1^2}}\sqrt{k_y^2+1-\frac{1}{1+k_1^2}}} > 0 \tag{A13}$$

Building hierarchically structured factor models with systematically selected residual correlations

Nils Petras[1]

[1] University of Mannheim

School of Social Sciences

Author Note

Declarations:

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2277 “Statistical Modeling in Psychology”.

Availability of data and materials: The open data example analyzed in this manuscript can be found under the following link: http://openpsychometrics.org/_rawdata/MACH_data.zip.

Code availability: The analysis code and reproducible manuscript can be found under the following link: https://osf.io/8s3ez/ (The repository is currently private, reviewers can access it via the following link: https://osf.io/8s3ez/?view_only=5a152407862f4a109e8617770a0f3e6c)

Conflicts of interest/Competing interests, Ethics approval, Consent to participate, Consent for publication: not applicable

I would like to thank Thorsten Meiser and Lesa Hoffman for their helpful comments and suggestions.

Correspondence concerning this article should be addressed to Nils Petras, University of Mannheim, School of Social Sciences, L13, 15, 68161 Mannheim, Germany. E-mail: nils.petras@uni-mannheim.de

Abstract

Many latent constructs are inherently multidimensional, but their measures do not necessarily follow a perfect subdomain structure. For example, many applications of the bi-factor model can not establish a full set of well-interpretable specific factors. The current work proposes a more flexible approach to the specification of hierarchically structured factor models. It uses a sparse set of relevant residual correlations to represent specific relationships beyond the target trait, selected using Bayesian lasso regularization. The four-step procedure, including cross-validation, combines the benefits of exploratory and confirmatory analysis, the compelling hierarchical structure of bi-factor models, and the principled Bayesian lasso selection procedure. The approach is introduced and discussed using a large open data example. In a multiverse analysis and multiple replications, consequences of several modeling choices are examined. In the example, the final model outperforms the traditional bi-factor model in both model fit and parsimony simultaneously. Furthermore, its flexibility in representing specific content matches a more realistic theoretical view of the complexity of typical questionnaire items. It is discussed how this new approach compares to the existing toolkit of factor modeling techniques.

Keywords: Bayesian lasso, confirmatory factor analysis, bi-factor models, residual correlations

Word count: 7211

Building hierarchically structured factor models with systematically selected residual correlations

Introduction

Multi-item psychological measures are routinely modelled using confirmatory factor analysis (CFA) models in the Structural Equation Modelling (SEM) framework. Often, simple structure models are used, in which the relationships between the items are fully explained by each reflecting a single latent variable from a (potentially correlated) set of latent variables. Because such simple structure models are unrealistically restrictive, the current work outlines a new approach to systematically lift two of their key assumptions to build a sparse, hierarchically structured factor model around a target trait.

Consider a covariance structure model of k items and p latent variables (Equation (1)).

$$\mathbf{Y} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{1}$$

$\mathbf{Y}$ is the $k \times 1$ response vector, $\mathbf{\Lambda}$ is the $k \times p$ factor loading matrix, $\boldsymbol{\eta}$ is the $p \times 1$ latent trait vector, and $\boldsymbol{\varepsilon}$ the $k \times 1$ vector of errors. Here, I call a factor model a simple structure model if it fulfills the following two conditions: $\mathbf{\Lambda}$ has only one nonzero entry per row (no cross-loadings) and the $k \times k$ covariance matrix of errors $\mathbf{\Psi}$ is a diagonal matrix (independently distributed residuals). The off-diagonal elements of $\mathbf{\Psi}$ are the *residual covariances* (ψ). The *residual correlations* are obtained by standardizing $\mathbf{\Psi}$ to be a correlation matrix, implying that all residual variances equal one.

Because items in psychological measures are complex, this assumption is rather restrictive. The meaning of questionnaire items is multifaceted, even if they are well-formulated and well-selected. They often reflect a large number of influences, both regarding the target construct and (unavoidable) nuisance. For example, item 16[1] of the well-established MACH IV machiavellianism questionnaire (Christie & Geis, 1970) seems simple at first glance: "It is possible to be good in all respects." Yet, it has been argued that (reversely scored) it reflects four specific content areas of machiavellianism: immorality, duplicity/dishonesty, misanthropy, and cynicism (Rauthmann, 2013). When considering this item by itself, it seems naively optimistic to assume that nothing beyond machiavellianism meaningfully influences the responses. For the example of machiavellianism, Rauthmann and Will (2011) identified no less than 46 machiavellianism content areas that potentially result in specific common variance among subsets of items. Even if most of them had a negligible specific influence on the data in practice, excluding all would be very restrictive.

Expectably, the assumptions of simple structure models are frequently violated in practice. For example, simple structure models of the Big Five usually have an unacceptable degree of model misfit (Marsh et al., 2010; Vassend & Skrondal, 1997). Most prominently, the often desired single factor model of the target construct (e.g., one factor per Big Five trait) rarely fits empirical data well by common standards of model fit index cut-offs (Hu & Bentler, 1999). A prominent example of that is the Rosenberg Self-Esteem Scale (RSES, Rosenberg, 1965). This short, straightforward, and supposedly unidimensional measure has produced a large amount of literature on its factor structure (e.g., Alessandri et al., 2015; Gnambs et al., 2018; Schmitt & Allik, 2005). The misfit of this simplest of all models leads many researchers to split their target construct into a correlated set of latent factors. For example, Corral and Calvete (2000) considered models with one to four (correlated) factors to represent the MACH-IV machiavellianism scale, which is discussed in the empirical example below. This prevents the assessment of the target construct as a whole and mixes the variance shared by all items with that of specific content domains (F. F. Chen et al.,

[1] The item enumeration may vary across publications. Here, it consistently matches the example dataset discussed below.

2012).

The current work outlines an approach to lift model restrictions just enough to account for the multitude of influences on item responses. It does so systematically and parsimoniously, and preserves the target trait as a core latent variable. Before describing the overall approach, two core building blocks are briefly reviewed: the bi-factor model and a state-of-the-art method to identify relevant ψs. After outlining the general approach, the chosen ψ selection method is described in more detail. After this introduction, an empirical example study is reported and used to inform the subsequent discussion of the proposed approach.

Building Blocks: Complexity beyond a single trait in latent variable models

The bi-factor model (Bornovalova et al., 2020; Holzinger & Swineford, 1937; Reise, 2012; see also Bader & Moshagen, 2022; and Petras & Meiser, 2024) is a currently popular approach to account for complexity beyond a singular target trait. In the bi-factor model, all items load on one general factor. A set of specific factors, which are orthogonal to the general factor and each other, represents the common variance of subsets of items. More formally, the loading matrix $\mathbf{\Lambda}$ (cf. Equation (1)) of a bi-factor model includes two non-zero entries per row: one in the column of the general factor and one in the column of the assigned specific factor. Traditionally, all items are assigned to a specific factor (Holzinger & Swineford, 1937), and recent variations assign all but some items to a specific factor by default (Eid et al., 2017). Specifying such a "full" bi-factor model is not strictly necessary, it is a modelling choice. Problematically, such bi-factor models often produce weak specific factors that are hard to interpret (Eid et al., 2017; Petras & Meiser, 2024). The schematic approach to only specify full bi-factor models ignores potentially more parsimonious models. For example, specific factors may capture a strong ψ between two of their items and show near zero loadings on all other items. In extreme cases, a specific factor has near zero loadings on all items or its inclusion in the model triggers convergence problems. The expectation that the data are best represented by a model with a symmetrical factor structure in which each item has two substantial factor loadings is restrictive and should be questioned.

The key feature of the bi-factor model is its inclusion of two independent levels: that of the general factor and that of the specific factors. An alternative approach to account for complexity beyond a singular to-be-measured trait is to allow for a sparse set of ψs. This can be done in a systematic way, using Bayesian lasso regularization to circumvent the problem of non-identification due to the large number of parameters (Pan et al., 2017; Zhang, Pan, Dubé, et al., 2021). The Bayesian lasso approach was originally suggested as a principled alternative to modification indices. Modification indices can be used to add ψs until the model fit reaches a "publishable" level. This ad-hoc modification strategy is typically seen very critically, because it involves multiple questionable statistical steps and is prone to capitalize on chance, inspiring questionable post-hoc rationalization (for a more detailed argument, see Pan et al., 2017). The well-founded hesitancy to use modification indices may have led to a hesitancy to include ψs in general. The more principled approach by Pan et al. (2017) estimates all ψs at the same time, using a restrictive double-exponential prior on the off-diagonal elements of $\mathbf{\Psi}^{-1}$. Importantly, the result can not only be used as a final model itself, but also to select a sparse set of ψs to keep for a final model, based on a reasonable cut-off (Zhang, Pan, & Ip, 2021; Zhang, Pan, Dubé, et al., 2021). The resulting model also has two independent hierarchical levels: factor(s) and the residual distribution with a sparse set of ψs.

Alternative modelling strategy

The current work examines the following suggested modeling strategy, using a detailed empirical example (below):

1) Select a baseline model including the target trait(s). Test for potentially relevant specific factors. Several potential baseline models may be compared on an exploratory sample.

2) Estimate a hierarchy of relevance among all possible ψs on the exploratory sample. For this step, Bayesian lasso regularization is principled and seems optimal.
3) Use a predefined inclusion criterion (or procedure) to decide how many ψs to add to the baseline model. For this, a cut-off of on the posterior means of the Bayesian lasso model estimates has been proposed to be a good rule of thumb (Zhang, Pan, & Ip, 2021).
4) Use a second, confirmatory sample to estimate the final model. Interpret only the estimates and model fit of this model.

The final model, through a combination of the bi-factor approach and the systematic inclusion of ψs, has three orthogonal hierarchical levels: The level of the general factor represents the target construct. The (optional) level of specific factors captures more specific common variance of a subset of items. The sparse set of covariances in the error distribution captures any further relationships between pairs of items.

Ultimately, the goal is to strike an optimal balance between the model's fit to the data and its parsimony. On the one hand, a sparse, well-structured model can more closely represent, and thereby deductively test, scientific theory. To that end, the resulting extreme restrictions of simple structure models have failed to accurately describe most empirical data. This introduces a danger of misinterpretation due to missed patterns in the data. On the other hand, a well-fitting model implies that no major structures in the data have been missed. To that end, the exploratory nature of approaches such as Exploratory Factor Analysis (EFA) or Exploratory Structural Equation Modeling (ESEM, Asparouhov & Muthén, 2009) involve an excessive amount of free parameters. This makes models so flexible that a good model fit no longer indicates a close fit to the theory and a large number of nuisance parameters (such as the full set of all potential cross-loadings) has to be included in the interpretation. In this case, researchers do no longer test the restrictions of the model, but rather interpret the match between parameter estimates and a desired or anticipated pattern. The suggested modelling strategy strikes a balance between these extremes.

Residual correlation selection using bayesian lasso regularization

Regularization methods reduce the number of model parameters (e.g. the number of included predictors in a multiple regression model) to those who efficiently describe the data (e.g. predict the criterion). Such a sparse model has a low likelihood of including mere random noise in its structure. The “Least absolute shrinkage and selection operator” (Lasso) adds a punishment in the estimation process that increases with the absolute values of parameter estimates. Thereby, solutions with small parameter estimates are prefered (Tibshirani, 1996). In this way, some parameter estimates are reduced to (almost) zero and can subsequently be fixed (or excluded from the model altogether). The Bayesian Lasso achieves this using a double-exponential (Laplace) prior (Park & Casella, 2008).

In SEMs, the value off the main diagonal of $\boldsymbol{\Psi}^{-1}$ can be interpreted as the conditional relationship of the two variables after accounting for all other variables (Pan et al., 2017). The double-exponential prior to regularize ψs is best applied to the off-diagonal entries of $\boldsymbol{\Psi}^{-1}$, not $\boldsymbol{\Psi}$ (Dempster, 1972; Pan et al., 2017). In this way, the posterior density is reduced for parameter combinations with a higher sum of absolute values off the main diagonal of $\boldsymbol{\Psi}^{-1}$. The strength of punishment for a lack of parsimony depends on the rate λ of the double-exponential prior. To reduce the subjectivity of this choice, a gamma hyperprior for the common rate of this prior and the exponential prior of the residual variances is used. The current work also adopts the further choice of priors by Pan et al. (2017) (see Appendix A). This approach estimates all ψs at the same time and alongside the other parameters, by using a block Gibbs sampler for MCMC sampling from the posterior distribution. This is the crucial advantage of the Bayesian lasso compared to other estimation procedures: They would fail from a lack of degrees of freedom given the excessive number of model parameters in this full model.

From the resulting posterior mean estimates of $\boldsymbol{\Psi}$, the residual correlations can be computed and ordered by absolute value, to complete step 2) of the proposed approach. Selecting those exceeding $|\hat{r}| \geq .1$ for inclusion in the model completes step 3). This cut-off was shown to be superior to alternative values and methods in a simulation study (Zhang, Pan, & Ip, 2021). As an alternative to the suggested cut-off by Zhang, Pan, and Ip (2021), I suggest that the ψs of the fully standardized model can be used directly. Since the fully standardized model (with item and factor variances equal to one) is the most commonly interpreted one, this seems to match common analysis best. This standardization sets item variances to be equal, instead of setting residual variances equal. If there is variation in the reliability of items, this common standardization can lead to different rank orders of the values in $\mathbf{\Psi}$ compared to the residual correlation matrix. It seems intuitive that including a residual correlation of the same value should have more impact on model estimates and model fit if the factors explain a smaller proportion of the item variances. For this reason, the ranking of standardized residual covariances can be seen as a more informative inclusion criterion than the ranking of residual correlations. Notably, the same cut-off value is more conservative on ψs than residual correlations, since the residual variances underlying ψs in fully standardized models (main diagonal of $\mathbf{\Psi}$) are all smaller than or equal to one. Therefore, a different cutoff value than 0.1 may be optimal when using ψ.

After selecting which ψs to estimate freely, the resulting model can then be estimated on a new, confirmatory sample (step 4). This avoids capitalizing on chance: if uncorrected hypothesis tests were conducted on estimates from the same sample, the alpha errors would be inflated due to the preselection of the most promising candidates. I suggest that the a posteriori model derived from the exploratory data should become the a priori model to be confirmed on a new dataset, resulting in valid hypothesis tests on the new sample.

Methods

Dataset

The example data are open data made available by openpsychometrics.org, a website that collects large amounts of questionnaire data online (http://openpsychometrics.org/_rawdata/MACH_data.zip, downloaded 18 July 2023). Data of 73489 participants on the MACH IV machiavellianism scale (Christie & Geis, 1970) and two other questionnaires are included in the dataset. The data was collected online between July 2017 and March 2019. Demographic statistics and a Figure showing the distributions of item responses can be found in the online supplement.

For the purpose of this study, only the MACH IV scale and the demographic data are used. The sample is reduced to two subsamples of a more typical sample size for psychological studies ($n = 1000$), by taking the first one thousand and the second one thousand cases. The first subsample is used as the exploratory subsample and the second is used as the confirmatory subsample. Eight further chunks of $n = 1000$ cases are used for replicability checks.

To understand the practical limitations of schematic modelling approaches, it is worthwhile to take a closer look at the structure of the application example. The analysis of a measure's data structure should be complemented by a theoretical structure from which expectations about the data structure can be derived and which explains patterns in the data. The MACH IV scale (Christie & Geis, 1970) comprises 20 items measuring machiavellianism. The construct of machiavellianism is named after Niccolo Machiavelli's (1469-1527) writings on manipulative social strategies. The scale's authors write that "Traditionally, the 'Machiavellian' is someone who views and manipulates others for his own purposes." (Christie & Geis, 1970, p. 1). Although the scale was designed to measure one target construct (machiavellianism), several different factor structures with multiple factors were proposed on various translations (Corral & Calvete, 2000; Hunter et al., 1982; O'Hair &

Cody, 1987; P. Monteiro et al., 2022; Williams et al., 1975). Rauthmann (2013) provide an analysis of the content of the MACH-IV items (Rauthmann, 2013, Table 1). They did not consider alterations to their unidimensional IRT model, but instead focussed on further item selection. In their analysis of item content, Rauthmann (2013) conclude that the items mix many different aspects of machiavellianism in various combinations, and these aspects are represented by varying numbers of items. Their analysis implies that no simple structure model (or bi-factor model) should be able to describe the data properly. Such models would not allow for the content areas to be fully or partly represented. There is only one clear content-based feature of the MACH IV scale that can easily be translated to a standard statistical model: it contains 50% reversely scored items.

Statistical models

Two possible baseline models are considered: a) a single factor model with one factor across all items and b) a model with one general factor across all items and one specific factor across all negatively keyed items[2]. In model b) the factor covariances were set to zero for the general and specific factor to be orthogonal. In all analyses, the observed and latent variables are standardized ($\mu = 0$, $\sigma = 1$). Using Maximum Likelihood (ML) estimation, it is decided which of these models is preferable, using a Likelihood Ratio Test (LRT) for model comparison to test if the inclusion of the method factor is worthwhile. Next, relevant ψs are selected using Bayesian lasso regularization (Pan et al., 2017) on the exploratory subsample. The analysis is repeated with two different cut-offs: $|\hat{\psi}| \geq .1$[3] and $|\hat{r}| \geq .1$. For each cut-off,

[2] The model with one specific factor for only the positively keyed items fits worse and is therefore discarded, see supplementary code. A replication of the four factor model by Corral and Calvete (2000) showed several flaws, despite a good fit to the data: items 17 and 19 show factor loadings with signs opposite to the reported direction and two of the factors correlate almost perfectly ($r = .99$). For this reason, and because their data are based on a translation, this model – and similar correlated factor models – are not explored further.

[3] To obtain a similarly strict cut-off to the proposed $|\hat{r}| \geq .1$ by Zhang, Pan, and Ip (2021), this value would need to be lowered. Without any clear strategy to do so, the current work just uses the same value. As this turned out to work just fine and arguably even better than the more liberal cut-off proposed by Zhang, Pan,

one final model with the selected ψs is specified. For brevity, the final model based on $|\hat{r}| \geq .1$ is not discussed in detail. These final models are estimated on the confirmatory subsample. The replicability of the selection of residual correlations is explored on a total of ten subsamples of $n = 1000$ participants. A full bi-factor model is estimated for comparison, in which both negatively keyed and positively keyed items are each related to a specific factor. To examine the trade-off between model fit and parsimony, a multiverse analysis across models selecting between zero and 75 ψs is presented.

Bayesian lasso regularization was used to 1) create a hierarchy of relevance among all potential ψs and 2) check against the cut-off of $|\hat{r}| \geq .1$ (Zhang, Pan, & Ip, 2021) and $|\hat{\psi}| \geq .1$. 10000 MCMC samples were drawn using Gibbs sampling, of which 3000 were discarded as burn-in. The replicability of the Bayesian lasso based selection is examined by repeating it on nine further subsamples of $n = 1000$ cases. The replicability is judged by a) the number of times a ψ is selected by the cut-off rule, b) the average of the posterior means in the nine subsamples beyond the exploratory subsample, and c) the retest reliability of the relevance measure (correlation between the posterior means (of ψs) across replications).

What if a larger or smaller number of ψs was selected? The descriptive multiverse analysis compares models with a range of 0-75 included ψs, adding ψs in order of their absolute posterior mean in the Bayesian lasso regularization results. For comparison, both the selection based on residual correlations and the analysis based on the alternative baseline model are included. The unreasonably high number of 75 ψs is used to map the overall development of the model statistics. It is not suggested to be a reasonable candidate for model selection. Although the cut-off proposed by Zhang, Pan, and Ip (2021) ($|\hat{r}| \geq .1$) was shown to be a good one-size-fits-all compromise across studies, a different number of selected ψs may be optimal in individual studies. To explore where an optimal balance on the trade-off between model fit and parsimony could be, several fit indices (AIC, BIC, χ^2, CFI,

and Ip (2021), strategies to arrive at an optimal cut-off might need to be reconsidered anyways.

RMSEA, SRMR) are computed for all models. Adding any ψ to the model can only improve the model fit – at least when ignoring parsimony. Therefore, a bootstrapping sample of 100 random orders was drawn to compare the performance of the Bayesian lasso selection to random selection.

To consider the effect on the estimated factors, the sum of the squared factor loadings per factor was computed. It could be expected that adding more and more ψs gradually chips away at the factors, lowering their factor loadings and affecting their usefulness and interpretability. To compare the effect of including the method factor with the effect of selecting ψs, the same analysis was repeated with both potential baseline models.

All analyses were performed in R version 4.3.1 (R Core Team, 2020) running under Windows 11. Maximum likelihood estimation was performed using the `lavaan` package (version 0.6-16, Rosseel, 2012). For the Bayesian lasso analysis, the code supplement by Pan et al. (2017) was used and extended.[4] The code underlying the analysis and the current manuscript can be found in the OSF supplement (https://osf.io/8s3ez/). The manuscript is rendered using the `papaja` R package for reproducible manuscripts (version 0.1.2, Aust & Barth, 2020).

Results

Step 1: Baseline model

The baseline model including the item wording method factor fits the confirmatory data subset well ($CFI = .919$, $RMSEA = .055$, $SRMR = .043$, $\chi^2(160) = 644.43$) and much better ($\chi^2_{diff}(10) = 434.98$, $p < .001$) than the simpler, single-factor model ($CFI = .848$, $RMSEA = .073$, $SRMR = .055$, $\chi^2(170) = 1,079.41$). In the baseline model, the absolute values of the factor loadings on the general Machiavellianism factor range from 0.21 to 0.73 ($M = 0.50$) and their sign was consistently in the expected direction. The factor

[4] The R package `blcfa` performs similar tasks (Zhang, Pan, Dubé, et al., 2021) to produce MPlus code of modified models. The current analysis is done entirely in the free and open source software R.

loadings on the method factor of this model are all positive and range from 0.17 to 0.55 ($M = 0.34$). A look at the distribution of factor loadings reveals that this factor is more than a glorified residual correlation between two of its items.

Step 2 + 3: Residual correlation hierarchy and selection

There are 5 ψs whose estimates exceed the cut-off value $|\hat{\psi}| \geq .1$ in the Bayesian lasso regularization model (Table 1). Repeating this analysis on nine further subsamples shows a massive variation in the number of samples in which a given ψ exceeds the cut-off. This should not be too surprising: for a true value exactly at the cut-off, one can expect a replicability of the selection decision of exactly 50%. Here, all posterior means of the selected ψs are close to the cut-off value. At least two of the ψs ($\psi_{14,4}$ and $\psi_{7,6}$) were selected from the set of 190 potential ψs in 80% or more of the repetitions. The retest reliability (pairwise correlations between replications) of the relevance measure ranges from 0.67 to 0.82 ($M = 0.74$).

A total of 15 estimates exceed the cut-uff of $|\hat{r}| \geq .1$ that was proposed by Zhang, Pan, and Ip (2021). Table 2 shows that many of these are selected in fewer than 50% of replications, with one notable exception ($\hat{r}_{15,2}$) that is selected in all replications. The average posterior mean in the replications of these additional selections is consistently lower than $|\hat{r}| \leq .1$ for all but $\hat{r}_{15,2}$. Taking a closer look at the estimates in the final model reveals that some of them are inconsistent with the posterior means (especially $\hat{r}_{6,3}$). All in all, it seems that this cut-off is too liberal in the current analysis, although it further includes one very consistent residual correlation compared to $|\hat{\psi}| \geq .1$.

Step 4: Final model

The addition of the ψs to the final model reduces the model misfit substantially (Table 3). A key finding is that the lasso-based model ("lasso-informed (cov)") shows a superior model fit compared to the symmetrical bi-factor model that combines all remaining (positively keyed) items in a second specific factor. Its absolute fit is superior, even though it adds only five parameters to the baseline model, compared to the bi-factor model's ten. The ψs are more relevant to describe the data than the factor completing the bi-factor model, even though their ML-estimates are all in the range of $-.2 < \hat{\psi} < .2$ (Table 1, $-.3 < \hat{r} < .3$ Table 2). The option to model several distinct relationships of one item with other items pays off in this application: All five ψs estimated at $|\hat{\psi}| \geq .1$ involve at least one negatively keyed item that is already assigned to the method factor. A full bi-factor model without ψs could not have represented these relationships properly. Nevertheless, the inclusion of specific factors can be of great importance, as the difference in model fit of the baseline models, both on the exploratory sample (see above) and on the confirmatory sample (Table 3) indicates. The alternative lasso-based model using $|\hat{r}| \geq .1$ ("lasso-based (cor)") shows a further improvement in model-fit, although the BIC indicates that this might not outweigh the loss of parsimony. For an overview of the estimates of the factor loadings in the lasso-informed model on the confirmatory sample and a visual display of these factor loading estimates, see the Appendix B.

Multiverse analysis

Figure 1 shows the results of the multiverse analysis. All models shown are estimated on the confirmatory subsample but developed on the exploratory subsample. The x-axis always shows the degrees of freedom with more parsimonious models to the right and models with up to 75 added ψs to the left. The upper six panels show fit index values on the y-axes. Higher values of the CFI indicate a better model fit. For all other fit indexes, lower values indicate a better fit of the model.

The main analysis based on the baseline model including the method factor and the cut-off of $|\hat{\psi}| \geq .1$ for ψ selection is shown in red. The red dot represents the final lasso-informed model. It lies on a red trace indicating all the possible models when selecting 0-75 ψs in order of their absolute posterior mean in the Bayesian lasso model. The black dot at the right end of the red trace represents the baseline model without ψs. All fit indices but the BIC indicate that adding more ψs improves the model, even though the RMSEA and the AIC penalize increasing the number of model parameters. The BIC indicates an optimal number of ψs near the chosen value of the final model.

The alternative selection strategy based on $|\hat{r}| \geq .1$ is shown in blue. It barely differs from the main analysis in the trace, because the hierarchy barely changes during the standardization of covariances to correlations. A major difference lies in the cut-off: the more liberal cut-off includes many more ψs. Judging by the only fit index that seems to strike a meaningful balance between parsimony and model fit, the BIC, the two alternatives surround the range of flat lines in which adding a ψ is barely worth its "cost" in parsimony loss. Given the results of the replicability analysis and the added difficulty of interpreting 15 instead of 5 ψs, this cut-off might still be seen as too lenient.

The cloud of transparent and overlapping gray background dots indicate models with a random selection of ψs added to the baseline model. This bootstrap analysis is based on 100 repetitions of random rank orders of ψs, from which the top 1-75 are selected, resulting in $100 \times 75 = 7500$ total models. The lasso-based selection is meaningful: On all fit indices and all numbers of added ψs, all (in rare exceptions: almost all) of the random bootstrap repetitions show a worse model fit than the lasso-informed models. This is an important finding, because the Bayesian lasso selection was performed on a different (exploratory) dataset and does not use model fit as a selection criterion.

To compare the relevance of added ψs based on the Bayesian lasso and substantial specific factors, the green dots represent lasso-informed models without the method factor for negatively keyed items. Using specific factors can be highly efficient: Up until about 50 added ψs, the addition of the method factor or a mix of the method factor and ψs (red and blue traces) is much more efficient than the addition of the same number of parameters all as ψs (green trace).

The lower two panels of Figure 1 show the sum of squared loadings of the general factor and the method factor for negatively keyed items. The “loss” of factor variance when adding ψs is insubstantial in size, as indicated by the rather flat traces. The only major loss of general factor variance is caused by the addition of the method factor (green trace versus red and blue traces). It reduces the sum of squared loadings of the general MACH factor by approximately 0.5 (comparing the baseline models), and itself has a sum of squared loadings of approximately 1.2 in the final model.

Discussion

The current study introduces a new and systematic approach to the specification of confirmatory factor analysis (CFA) models. In four steps, 1) a baseline model is selected, 2) a hierarchy of relevance among residual covariances (ψs) is established, 3) a number of ψs is selected, and 4) the resulting model is estimated on new data. This approach accounts for both the theoretical understanding of questionnaire items as complex and the practical difficulties in achieving well-fitting models – while still yielding sparse models. The use of two distinct (sub-)samples combines the advantages of data exploration (not missing patterns in the data) and cross-validation (confirmatory hypothesis testing, not capitalizing on chance). The reanalysis of open data in a multiverse analysis further demonstrates the importance of both specific factors and ψs in CFA models, which the suggested approach selects systematically. In these ways, the current approach strikes a balance between extremely flexible exploratory models (EFA, ESEM, Asparouhov & Muthén, 2009) and the often overly restrictive, simple structure confirmatory models. Importantly, it does so without engaging in post-hoc modification, compared to the similar approach of using modification indices after obtaining a bad model fit. Inspired by the logic of bi-factor models, the resulting models have a hierarchical factor structure with orthogonal levels. One level comprises the general target trait and one (or two, in case of specific factors) further levels comprise separate, specific relationships between items.

Several findings in the empirical example validate the outlined four-step approach and demonstrate its relevance. First, the lasso-informed model obtained in this way fits the data better than a full bi-factor model, on top of being more parsimonious. This is partly because of its flexibility to involve items in multiple specific relationships with other items. Furthermore, it only includes relevant specific factors instead of defining a schematic bi-factor structure. The resulting model includes a minimal set of parameters in an efficient way: the ψs are added where the exploratory analysis on another (sub-)sample indicates that there is a relevant pattern in the data.

Second, the clear importance of the method factor (Table 3, Figure 1) shows the importance of searching a good baseline model in step 1. Specific factors are able to represent the common variance of a set of items (in this case, 10) much more parsimoniously than the pairwise ψs. In combination with the first finding, it is clear that both the addition of ψs and the addition of specific factors have the potential to best improve the model. In this case, the added factor could have several different interpretations, although I termed it "method factor" for its relationship to the item keying. It could be a methodological artifact from the measure design, a result of response processes, or a relevant content domain. It is important to seriously consider the substantive meaning of this factor, since the items are not mere negations: They differ in the content they cover. For ψs it is equally important to consider substantive and methodological explanations. Both substantive and method-induced variance can be relevant to model as a factor or as a residual correlation.

Third, the replicability of the selection of ψ is mixed in the empirical example (Table 1) and proved questionable if a cut-off of $|\hat{r}| \geq .1$ (Zhang, Pan, & Ip, 2021) is used (Table 2). On the one hand, there are well-replicated selections (Table 1) and the lasso-informed selection is clearly superior over random selection (Figure 1). The retest-reliability (correlation between individual repetitions) of the relevance measure for the residual correlations (posterior mean of Bayesian lasso estimation) is consistently close to its mean of

$M = 0.74$ and the average posterior mean of the replications is mostly consistent with the selection decision (Table 1). This shows that the suggested procedure generally selects the most relevant ψs with some reliability. On the other hand, the observed uncertainty shows the importance of using a confirmatory (sub-)sample to estimate the final model. There is a real danger to capitalize on chance when selecting ψs post-hoc. One of the selected ψs would not have been selected in any of the nine replication attempts (Table 1). In the two-sample approach, the interpreted final estimate on the confirmatory model is independent of the (presumed) random variation that lifted the ψ over the cut-off in the selection procedure. In this way, the hypothesis tests on the final estimates are valid and unbiased. The current application therefore provides a promising proof of concept of the suggested approach.

The substantial uncertainty in the ψ selection can be explained easily: selection problems like this are only partly solvable. Given an essentially continuous distribution of the true relevance of ψs on a metric scale, any inclusion rule will produce severe uncertainty regarding the ψs close to the selection criterion (i.e. cut-off). For this reason, a guiding principle for selection would be crucial, but is currently lacking. The proposed cut-off at $|\hat{r}| = .1$ does not represent a universal principle but rather a practical convention informed by simulation studies and the literature on other kinds of parameters in various models (Zhang, Pan, & Ip, 2021). One important innovation by Zhang, Pan, and Ip (2021) is to use an absolute cut-off instead of significance testing or posterior density intervals ensures that sample size does not systematically influence the number of selected ψs. If the selection is made based on $\hat{r}$ or $\hat{\psi}$ might be of little relevance in practice, but could be discussed for principled reasons. The choice of a cut-off (or a procedure to find a data-informed cut-off) seems most important. The current multiverse analysis hints to the possibility that many potential choices might be almost equally optimal.

Comparison to existing approaches

The suggested approach aims to build a model of a target trait whilst simultaneously acknowledging the complexity of the data and the (potential) multidimensionality of the target construct. How does this compare to alternative approaches?

One alternative is to consider a model that directly reflects an abstract theory, such as a straightforward CFA model without any hierarchical factor structure, cross-loadings, or ψs. The suggested approach allows a more realistic representation of nuisance influences, without substantially reducing factor variance, even if many ψs are added (Figure 1). It also carries over all the advantages of bi-factor models. If the target construct strongly violates the assumption of unidimensionality, this can be accounted for and the suggested approach still provides a latent variable representing the overall target trait. With domain-specific variance represented explicitly in the model, researchers can better decide how meaningful the obtained target trait is, and if its measurement needs to be improved. Finally, the suggested approach provides more security to not miss an important pattern in the data than the use of fit indices alone: it systematically uncovers all the ψs for which fixing them to zero causes substantial misfit. On the flipside, it has the potential to include chance findings in the model, for which there need to be precautions (such as the two-sample approach).

The suggested approach seems more principled than the standard bi-factor model in the modeling of specific variance based on content-domains or item-content based nuisance. It is much more flexible, yet at the same time optimizes parsimony (see Table 3), because it allows researchers to systematically replace weak specific factors of a bi-factor model (Eid et al., 2017; Petras & Meiser, 2024) by a sparse set of ψs. It can sparsely represent a theory that implies specific relationships between pairs or sets of items more flexibly than a bi-factor model. Although, it has to be acknowledged that in the empirical example, the match between the data analysis and the content analysis by Rauthmann (2013) is minimal. The proposed approach allows each item to be involved in multiple specific relationships with other items. Lastly, its exploratory first steps make it more robust against a potential misrepresentation of the data: instead of proposing a schematic bi-factor structure, Step 1 of the current approach encourages testing multiple candidate models with or without certain specific factors. In cases where specific variance reflects the influence of different raters, testlets, or other rather systematically structured method effects, the suggested approach might not be sensible and the bi-factor model may be much preferable.

Compared to more flexible models, such as exploratory factor analysis (EFA) and exploratory structural equation modeling (ESEM, Asparouhov & Muthén, 2009), the suggested approach produces much more sparse models. This simplifies the interpretation and makes model fit indices meaningful. The suggested approach explicitly avoids using the Bayesian lasso model as the final model. It only selects the few relevant ψs for inclusion in the final model. This greatly simplifies the model by removing the information that is indistinguishable from random noise. A similar selection procedure has been proposed for cross-loadings (J. Chen et al., 2021; Zhang, Pan, Dubé, et al., 2021).

Limitations and future directions

Bayesian-lasso regularization is not the only approach to create a rank order of importance of ψs. In many applications, its results may not differ much from less sophisticated approaches, such as subtracting the model-implied $\hat{\mathbf{\Sigma}}$ from the observed $\mathbf{\Sigma}$ and selecting by the order of absolute difference values. Yet, the current approach seems optimal for two reasons: First, it is a systematic approach that can be easily planned a priori. It produces a fully confirmatory final model without any post-hoc modifications. Second, it estimates all ψs and all other model parameters simultaneously. This avoids any problems with ripple effects in stepwise approaches. For example when using modification indices, the model has to be estimated again after adding a ψ, because the evaluation of all other ψs may change when one is added. Using the differences between the observed $\mathbf{\Sigma}$ and model-implied $\hat{\mathbf{\Sigma}}$ (of the baseline model) runs into the same problem, because the inclusion of one ψ in the model might change the whole pattern of the differences between the matrices. The Bayesian lasso instead considers the value of ψ when all other ψs are included in the model. In other words, it allows all estimates of the model to adapt to one another, before ψs are ranked by relevance. Nevertheless, a crude selection based on a crude criterion might frequently result in the exact same selection as the Bayesian lasso and be more practical regarding computation time.

There are alternatives to the Bayesian lasso in regularization. For the case of selecting regression coefficients, Park and Casella (2008) showed that the Bayesian Lasso shrank small estimates towards zero quicker than the alternative Ridge regression. To my best knowledge, this advantage has not yet been confirmed for the use on ψs. Therefore, ridge regression priors (or spike and slab priors), as used for cross-loadings in Bayesian Structural Equation Modeling (Muthén & Asparouhov, 2012), are potential alternatives that future research could explore.

Systematically adding ψs should not lower the standard for the acceptability of measures. In the example, there are multiple items with rather small factor loadings (Table B1), especially item 19 ("People suffering from incurable diseases should have the choice of being put painlessly to death.") which has a questionable relation to the target construct Machiavellianism. This item also happens to be involved in two[5] negative ψs in the final model (Table 1). Both of these are are small, not well replicated, and hard to interpret. The overlap of assigned content areas of these item pairs is minimal (Rauthmann, 2013). The ability to account for such imperfections with flexible statistical models – and achieving a good model fit – should not be used as an excuse to forego the improvement of measures. To the contrary, it should be used to inform the improvement of measures.

[5] Item 6: "Honesty is the best policy in all cases.", item 7: "There is no excuse for lying to someone else."

Conclusion

The current work proposes and demonstrates a four-step approach to the specification of sparse, yet flexible factor models with a hierarchical structure. It combines solutions to several existing problems in measurement models: It replaces schematic modelling approaches by a systematic procedure to obtain sparse models. It represents the overall target trait while also accounting for the inherent multidimensionality of psychological constructs and the inherent complexity of items in psychological measurement. It strikes a balance between arbitrarily flexible exploratory approaches (EFA, ESEM) and overly strict assumptions in basic confirmatory models. It encorporates the systematic selection of ψs by the Bayesian lasso (Pan et al., 2017; Zhang, Pan, & Ip, 2021; Zhang, Pan, Dubé, et al., 2021), which elegantly solves most problems underlying modification indices. Its systematic data exploration finally results in a fully confirmatory model on a new (sub-)sample.

References

Alessandri, G., Vecchione, M., Eisenberg, N., & Łaguna, M. (2015). On the factor structure of the rosenberg (1965) general self-esteem scale. *Psychological Assessment*, *27*(2), 621. https://doi.org/10.1037/pas0000073

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* https://github.com/crsh/papaja

Bader, M., & Moshagen, M. (2022). *No probifactor model fit index bias, but a propensity toward selecting the best model.* https://doi.org/10.1037/abn0000685

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*, *88*(1), 18–27. https://doi.org/10.1016/j.biopsych.2020.01.013

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

Chen, J., Guo, Z., Zhang, L., & Pan, J. (2021). A partially confirmatory approach to scale development with the bayesian lasso. *Psychological Methods*, *26*(2), 210. https://doi.org/10.1037/met0000293

Christie, R., & Geis, F. L. (1970). *Studies in machiavellianism.* Academic Press.

Corral, S., & Calvete, E. (2000). Machiavellianism: Dimensionality of the mach IV and its relation to self-monitoring in a spanish sample. *The Spanish Journal of Psychology*, *3*, 3–13. https://doi.org/10.1017/S1138741600005497

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 157–175.

https://doi.org/10.2307/2528966

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541. https://doi.org/10.1037/met0000083

Gnambs, T., Scharl, A., & Schroeders, U. (2018). The structure of the rosenberg self-esteem scale. *Zeitschrift für Psychologie*. https://doi.org/10.1027/2151-2604/a000317

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hunter, J. E., Gerbing, D. W., & Boster, F. J. (1982). Machiavellian beliefs and personality: Construct invalidity of the machiavellianism dimension. *Journal of Personality and Social Psychology*, *43*(6), 1293. https://doi.org/10.1037/0022-3514.43.6.1293

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471. https://doi.org/10.1037/a0019227

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313. https://doi.org/10.1037/a0026802

O'Hair, D., & Cody, M. J. (1987). Machiavellian beliefs and social influence. *Western Journal of Communication (Includes Communication Reports)*, *51*(3), 279–303. https://doi.org/10.1080/10570318709374272

P. Monteiro, R., Lins de Holanda Coelho, G., Medeiros Cavalcanti, T., Moura Grangeiro, A. S. de, & V. Gouveia, V. (2022). The ends justify the means? Psychometric parameters of the MACH-IV, the two-dimensional MACH-IV and the trimmed MACH in brazil.

Current Psychology, 1–10. https://doi.org/10.1007/s12144-020-00892-0

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The bayesian lasso. *Psychological Methods*, *22*(4), 687. https://doi.org/10.1037/met0000112

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Petras, N., & Meiser, T. (2024). Problems of domain factors with small factor loadings in bi-factor models. *Multivariate Behavioral Research*, *59*(1), 123–147. https://doi.org/10.1080/00273171.2023.2228757

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rauthmann, J. F. (2013). Investigating the MACH–IV with item response theory and proposing the trimmed MACH. *Journal of Personality Assessment*, *95*(4), 388–397. https://doi.org/10.1080/00223891.2012.742905

Rauthmann, J. F., & Will, T. (2011). Proposing a multidimensional machiavellianism conceptualization. *Social Behavior and Personality: An International Journal*, *39*(3), 391–403. https://doi.org/10.2224/sbp.2011.39.3.391

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Rosenberg, M. (1965). Rosenberg self-esteem scale. *Journal of Religion and Health*. https://doi.org/10.1037/t01038-000

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the rosenberg self-esteem scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, *89*(4), 623. https://doi.org/10.1037/0022-3514.89.4.623

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vassend, O., & Skrondal, A. (1997). Validation of the NEO personality inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*(2), 147–166. https://doi.org/10.1002/(SICI)1099-0984(199706)11:2%3C147::AID-PER278%3E3.0.CO;2-E

Williams, M. L., Hazleton, V., & Renshaw, S. (1975). The measurement of machiavellianism: A factor analytic and correlational study of mach IV and mach v. *Communications Monographs*, *42*(2), 151–159. https://doi.org/10.1080/03637757509375889

Zhang, L., Pan, J., Dubé, L., & Ip, E. H. (2021). Blcfa: An r package for bayesian model modification in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 649–658. https://doi.org/10.1080/10705511.2020.1867862

Zhang, L., Pan, J., & Ip, E. H. (2021). Criteria for parameter identification in bayesian lasso methods for covariance analysis: Comparing rules for thresholding, p-value, and credible interval. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 941–950. https://doi.org/10.1080/10705511.2021.1945456

Table 1

Residual covariances selected via Bayesian lasso regularization and a cut-off of 0.1 on the posterior mean in the exploratory subsample

Items	post mean	selections	avg rep	estimate
14, 4	.14	8	.12	.185 (.123, .247)
7, 6	.15	10	.14	.169 (.09, .248)
19, 6	-.11	1	-.05	-.087 (-.154, -.019)
19, 7	-.14	6	-.10	-.186 (-.251, -.12)
14, 11	.11	4	.09	.124 (.061, .186)

Note. all values are standardized covariances; post mean = posterior mean in the exploratory subsample, selections = number of selections out of ten total repetitions; avg rep = average posterior mean across the nine replications, estimate = ML estimate of residual covariance in final, standardized model on confirmatory subsample with 95% confidence interval in brackets

Table 2

Residual correlations selected via Bayesian lasso regularization and a cut-off of 0.1 on the posterior mean in the exploratory subsample

Items	post mean	selections	avg rep	estimate
19, 1	-0.103	3	-0.072	-.136 (-.24, -.032)
7, 2	-0.12	1	-0.058	-.123 (-.232, -.015)
15, 2	0.157	10	0.147	.204 (.11, .298)
6, 3	0.123	1	0.056	-.283 (-.499, -.066)
11, 4	0.103	5	0.084	.188 (.102, .275)
14, 4	**0.202**	**10**	**0.173**	**.288 (.201, .375)**
7, 5	0.114	4	0.074	.092 (-.015, .198)
7, 6	**0.265**	**10**	**0.244**	**.23 (.045, .415)**
9, 6	0.14	3	0.077	-.03 (-.244, .185)
19, 6	**-0.164**	**3**	**-0.078**	**-.148 (-.255, -.041)**
10, 7	0.109	3	0.076	.022 (-.109, .152)
19, 7	**-0.177**	**9**	**-0.134**	**-.237 (-.324, -.151)**
15, 10	-0.142	3	-0.067	-.077 (-.178, .025)
14, 11	**0.143**	**7**	**0.123**	**.2 (.12, .281)**
18, 11	-0.122	2	-0.054	-.111 (-.193, -.03)

Note. all values are correlations; post mean = posterior mean in the exploratory subsample, selections = number of selections out of ten total repetitions; avg rep = average posterior mean across the nine replications, estimate = ML estimate of residual correlation in final model on confirmatory subsample (95% confidence interval); bold entries are also selected using the covariance cut-off

Table 3

Model fit on confirmatory subsample, n = 1000

	chisq	df	cfi	bic	aic	rmsea	srmr
simple baseline	888.3	170	0.871	64,311	64,115	0.065	0.051
baseline	541.6	160	0.932	64,033	63,788	0.049	0.040
lasso-informed (cov)	447.3	155	0.948	63,973	63,703	0.043	0.036
lasso-informed (cor)	378.6	145	0.958	63,974	63,655	0.040	0.033
bifactor	458.7	150	0.945	64,019	63,725	0.045	0.035

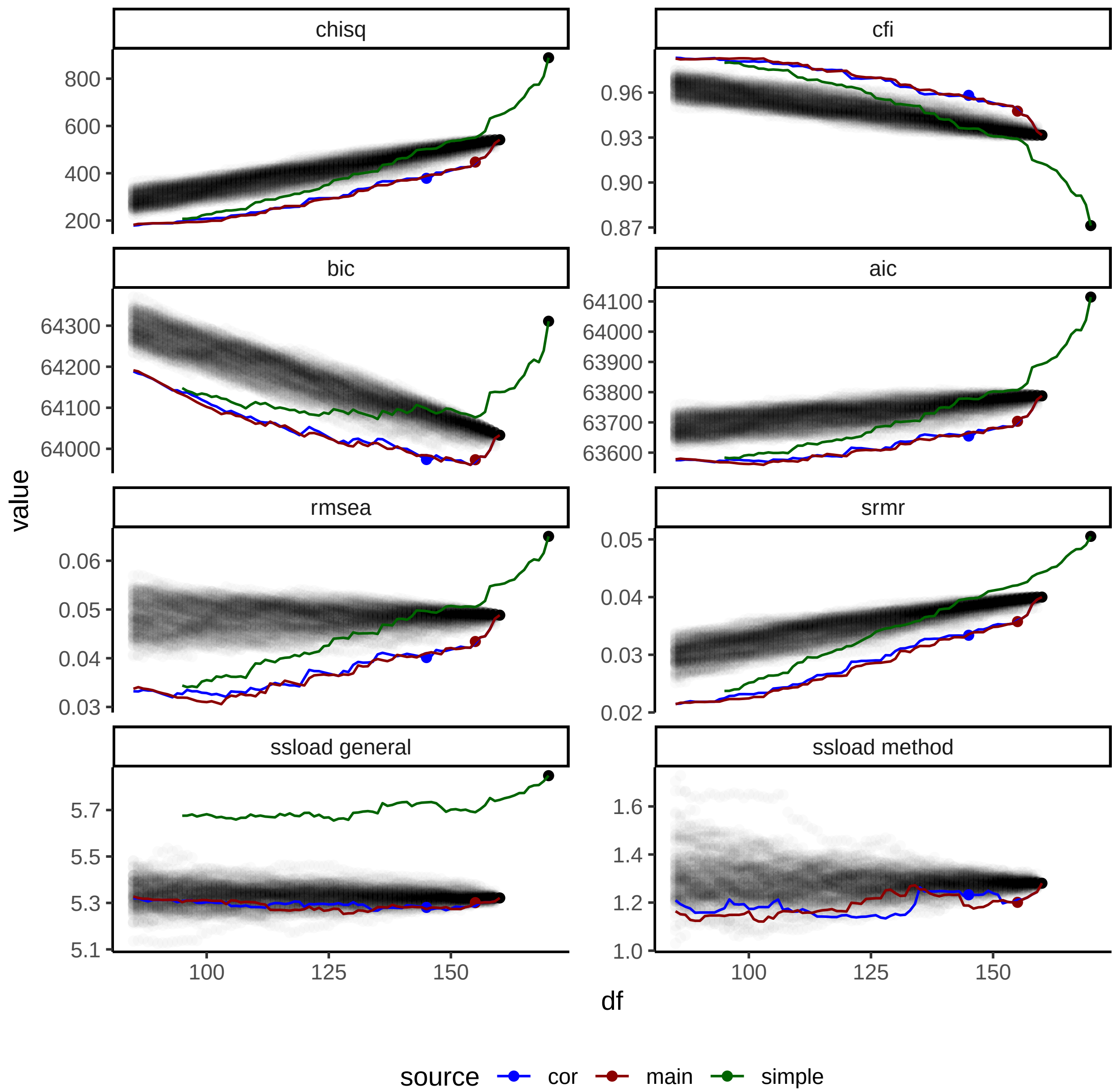


Figure 1

Multiverse analysis of model fit and sum of squared loadings of models including between 0 and 75 residual correlations. Black dots represent baseline models. The red trace represents all potential cut-offs for the inclusion of 0-75 residual covariances added to the selected baseline model. The red dot shows the model selected based on a lasso posterior mean cut-off of 0.1. The blue trace shows the same analysis with a selection based on residual correlations, not covariances, and the blue dot represents a cut-off of 0.1. The dotcloud in the background shows a bootstrap sample of 100 random traces for comparison. The green trace uses the rejected single-factor baseline model.

Appendix A

Priors

For the residual covariance matrix $\mathbf{\Psi}$, priors are assigned to its inverse $\mathbf{\Psi}^{-1}$. The main diagonal (variance) and off-diagonal (covariance) priors are exponential and double-exponential priors with a common rate of the form

$$\sigma_{xx} \sim \frac{\lambda}{2} e^{-\frac{\lambda}{2}\sigma_{xx}} \tag{A1}$$

$$\sigma_{xy} \sim \frac{\lambda}{2} e^{-\lambda|\sigma_{xy}|}; x \neq y \tag{A2}$$

with a Gamma-hyperprior of $\lambda \sim \Gamma(1, 0.01)$. The factor covariance matrix $\mathbf{\Phi}$ is fixed to be a diagonal matrix with ones on the main diagonal, to ensure orthogonal, standardized latent variables. The factor loading matrix $\mathbf{\Lambda}$ has a predefined factor structure. Therefore, its elements are either zero or have a normal prior of the form

$$\mathbf{\Lambda}_k \sim \begin{cases} N(\mathbf{\Lambda}_{0k}, \mathbf{H}_{0k}) \\ 0 \end{cases} \tag{A3}$$

with uninformative mean and variance parameters.

Appendix B

Lasso-informed model factor loadings

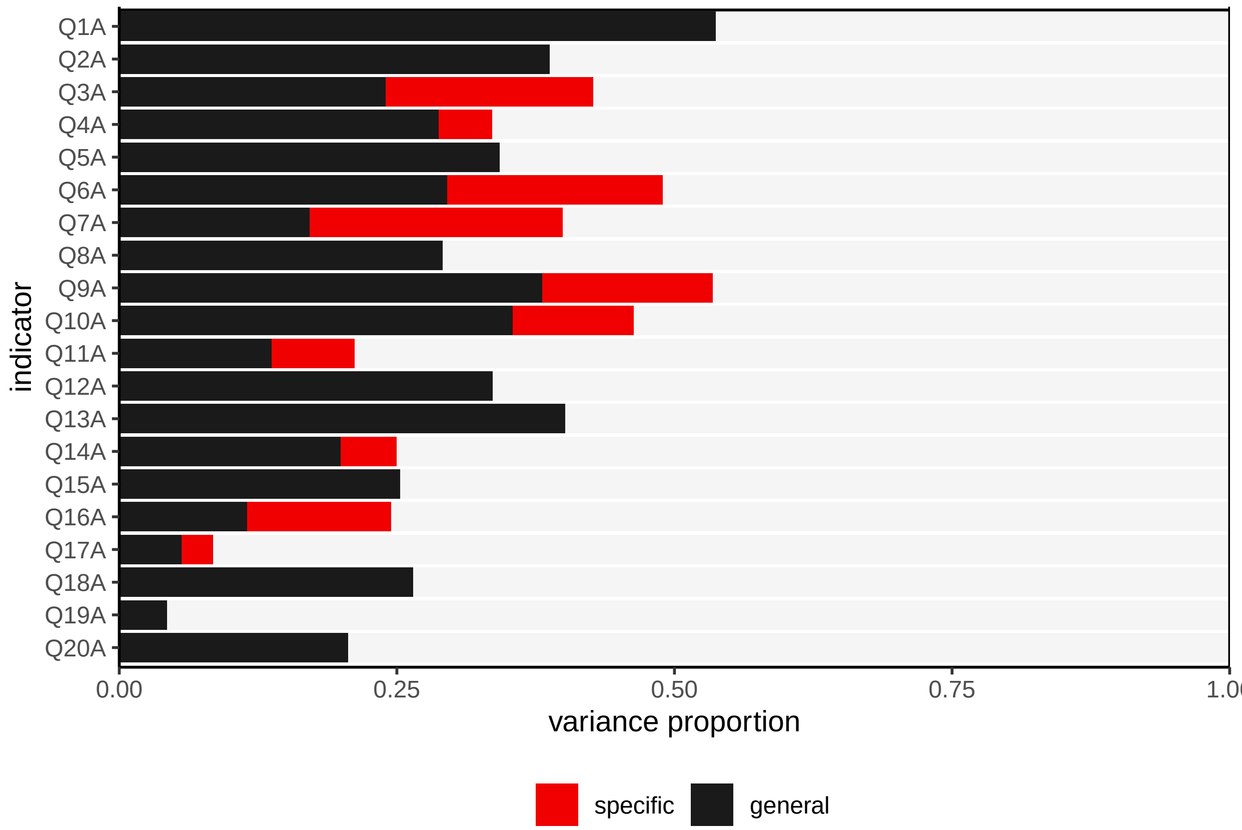


Figure B1

Variance decomposition of MACH IV items in the final model; specific = variance specific to the method factor; general = variance of the general MACH factor; the length of the stacked bars is determined by the respective squared factor loadings

[tbp]

Table B1

Standardized factor loadings of the final model

	MACH	method
Q1A	.733 (.698, .768)	0
Q2A	.623 (.579, .666)	0
Q3A	-.49 (-.543, -.437)	.432 (.365, .5)
Q4A	-.537 (-.587, -.487)	.219 (.148, .29)
Q5A	.585 (.539, .631)	0
Q6A	-.543 (-.593, -.494)	.44 (.372, .509)
Q7A	-.414 (-.471, -.357)	.477 (.406, .549)
Q8A	.54 (.491, .589)	0
Q9A	-.617 (-.662, -.573)	.392 (.329, .455)
Q10A	-.595 (-.641, -.549)	.33 (.265, .396)
Q11A	-.371 (-.43, -.311)	.273 (.198, .348)
Q12A	.58 (.533, .626)	0
Q13A	.634 (.591, .676)	0
Q14A	-.447 (-.502, -.391)	.224 (.149, .299)
Q15A	.503 (.451, .554)	0
Q16A	-.34 (-.4, -.279)	.36 (.286, .433)
Q17A	-.237 (-.301, -.173)	.168 (.087, .248)
Q18A	.514 (.464, .565)	0
Q19A	.207 (.143, .271)	0
Q20A	.454 (.4, .508)	0

Note. Values in brackets indicate the boundaries of the 95% confidence interval.