Psychological Methods
2017, Vol. 22, No. 3, 541–562

1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000083

Anomalous Results in *G*-Factor Models: Explanations and Alternatives

Michael Eid
Free University of Berlin

Christian Geiser
Utah State University

Tobias Koch
Leuphana University of Lüneburg

Moritz Heene
Ludwig-Maximilian University of Munich

G-factor models such as the bifactor model and the hierarchical *G*-factor model are increasingly applied in psychology. Many applications of these models have produced anomalous and unexpected results that are often not in line with the theoretical assumptions on which these applications are based. Examples of such anomalous results are vanishing specific factors and irregular loading patterns. In this article, the authors show that from the perspective of stochastic measurement theory anomalous results have to be expected when *G*-factor models are applied to a single-level (rather than a 2-level) sampling process. The authors argue that the application of the bifactor model and related models require a 2-level sampling process that is usually not present in empirical studies. We demonstrate how alternative models with a *G*-factor and specific factors can be derived that are more well-defined for the actual single-level sampling design that underlies most empirical studies. It is shown in detail how 2 alternative models, the bifactor-($S - 1$) model and the bifactor-($S{\cdot}I - 1$) model, can be defined. The properties of these models are described and illustrated with an empirical example. Finally, further alternatives for analyzing multidimensional models are discussed.

Keywords: *G*-factor, bifactor model, nested factor model, ctc(m-1) model, stochastic measurement theory

The analysis of *G*-factor structures has a long tradition in psychology going back to Spearman's (1904) seminal article titled "*General intelligence," objectively determined and measured*. In this article, Spearman concluded that "all branches of intellectual activity have in common one fundamental function (or groups of functions), whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all the others" (p. 284). The idea that each cognitive ability represents a general ability and a specific component that is not shared with other abilities has strongly influenced intelligence research over the last 110 years. This idea has also had a strong impact on psychometric theory and modeling. In 1937, Holzinger and Swineford developed the bifactor method of factor analysis as an approach that enables researchers to separate a general factor from uncorrelated factors that are specific to a group of tests that measure the same ability domain (specific factors) through analyzing correlations of different ability scales.

Although the bifactor approach is rather old, it has only become very popular in recent years (Reise, 2012). For example, a literature research using PsycINFO revealed 249 hits for the search term "bifactor or bi-factor" for the period 2009–2014, but only 24 hits for 2003–2008, five hits for 1997–2002, and even only 44 hits for the whole period from 1937–1996. Many researchers from quite different areas are using this approach to model the multidimensional structure of their data. The applications are not limited to the analysis of cognitive abilities, but cover many different areas of psychology. Examples are the measurement of quality of life (Garin et al., 2013), well-being (Chen, Jing, Hayes, & Lee, 2013), and psychopathology (Urbán et al., 2014), just to mention a few.

The bifactor model and related *G*-factor models are typically applied in situations in which different domains of a construct are assessed by multiple observed variables. The domains can, for example, be different life domains for the assessment of life satisfaction. The different domains can also be different contexts as they are considered, for example, in the testlet approach. According to Wainer, Bradlow, and Wang (2007) a testlet is "a packet of test items that are administered together" (p. 44). For example, in research on text comprehension different text passages are presented with different items. A group of items referring to one text passage is a testlet. The text passage is the context. Researchers are often interested in the general reading comprehension ability that is not specific to a specific text passage.

The bifactor model has many advantages that make its application attractive. For example, it enables researchers to decompose an observed test score variable into three components: (a) The general factor, (b) specific factors, and (c) measurement error. It allows estimating the reliability of measurements and decomposing the true score (i.e., error-free) variance into components due to general and specific effects. These coefficients reflect, for exam-

This article was published Online First August 15, 2016.

Michael Eid, Department of Education and Psychology, Free University of Berlin; Christian Geiser, Department of Psychology, Utah State University; Tobias Koch, Center for Methods, Leuphana University of Lüneburg; Moritz Heene, Department of Psychology, Ludwig-Maximilian University of Munich.

Correspondence concerning this article should be addressed to Michael Eid, Department of Education and Psychology, Free University of Berlin, Habelschwerdter Allee 45.D-14195 Berlin, Germany. E-mail: michael.eid@fu-berlin.de

ple, the degree to which an intellectual performance is due to a general trait or to a more specific ability. It also allows testing specific hypotheses about the structure of the correlations of different observed variables. Besides these more conceptual advantages, the bifactor model also has some technical advantages, in particular, when the observed variables are categorical. Full information maximum likelihood estimation methods for item factor analysis are usually restricted to a small number of factors because the dimensionality of the integrals increase with the number of factors and becomes intractable with a higher number of factors. However, the specific structure of the bifactor model in which an observed variable is decomposed only into two factors makes it possible to use efficient maximum likelihood estimation methods for parameter estimates independently of the number of specific factors considered. This is due to the fact that these estimation procedures only require a chain of two-dimensional integrals and not integrals of higher dimensionality (see Cai,Yang, & Hansen, 2011, for a deeper discussion).

Despite these advantages and the high popularity of bifactor models, however, "many conceptual as well as technical issues in the application of bifactor models remain poorly understood in the psychometric and assessment communities" (Reise, 2012, p. 669). Indeed, many applications of the bifactor model and related *G*-factor models in recent years revealed anomalous results—anomalous results that were even found in the very first applications of the bifactor method presented by Holzinger and Swineford (1937). These anomalous results, which we describe in detail in the next section, have led us to question the appropriateness of the bifactor model and related *G*-factor models for many types of applications in psychological research.

In this article, we take a closer look at these anomalous results and application problems and discuss potential reasons why these problems might occur. We focus on models of confirmatory factor analysis because these models are now the most often applied models for analyzing *G*-factor structures; we review alternative approaches (e.g., principal component analysis, aggregation approach) in the Discussion section. We show that from the perspective of stochastic measurement theory (SMT; Eid, 1996; Steyer, 1989) the proper application of bifactor and related models requires a two-level sampling design—a design that is not present in most substantive applications of these models. By using an example from latent state-trait theory, we clarify under which conditions the application of a traditional bifactor model and related *G*-factor models is justified. We conclude that the typical applications of *G*-factor models require different models. We explain how SMT can be used to define alternative *G*-factor models. We present these alternative models in detail and illustrate them with an empirical example. Finally, we discuss further alternatives to *G*-factor models.

In our presentation, we focus on the conceptual problems and do not consider all statistical methods for analyzing bifactor structures that have been developed in recent years. The conceptual problems we discuss are related to the general structure of the model in combination with the area of application, and they are not related to a specific statistical method (such as estimation or rotation methods). For an overview of these more practical issues see, for example, Brunner, Nagy, and Wilhelm (2012), Cai, Yang, and Hansen (2011), Reise (2012), Reise, Moore, and Haviland (2010) as well as Reise, Moore, and Maydeu-Olivares (2011).

Overview

The article is organized as follows. First, we explain the basic structure of the bifactor and hierarchical *G*-factor models as the two most prominent approaches for the analysis of a *G* factor and specific factors. Then, we discuss some anomalous, but common results found in applications of these models. In the next step, we discuss some additional conceptual issues and give an introduction to the basic idea of SMT. We show that from the perspective of SMT, the factors in the bifactor model are only well-defined in the case of a two-step sampling procedure (two-level structure). This two-level structure requires the interchangeability of the different domains of a construct that are supposed to measure a common *G* factor. We explain that in the majority of applications the measurement design or types of domains do not meet this requirement. We demonstrate how alternative models with a *G* factor and domain-specific factors can be derived that are more well-defined for the actual single-level sampling design that underlies most empirical studies that currently use the bifactor or related approaches. These alternative models differ from the traditional bifactor and hierarchical *G*-factor model in some important ways. We explain what a *G* factor and specific factors mean in these alternative models and describe how these insights can guide future research. We present an application of the alternative models to the study of emotion intensity. Finally, we discuss further alternatives for measuring *G* and specific factors.

G-factor Models of Confirmatory Factor Analysis

There are two general *G*-factor models of confirmatory factor analysis (CFA) that are widely applied (Brunner, Nagy, & Wilhelm, 2012): The bifactor model (Holzinger & Swineford, 1937), which is sometimes also referred to as *nested factor model*, and the higher-order *G*-factor model, which is also called hierarchical *G*-factor model (Gustafsson & Balke, 1993). A bifactor model for three domains and three indicators per domain is depicted in Figure 1. In the bifactor model, there is a general factor (G) that is assumed to influence all observed variables. In addition, there is one specific factor (S_k; $k = 1, \ldots, K$; K: number of facets) for each domain. These specific factors were called group factors by Holzinger and Swineford (1937). Indicators that pertain to the same domain load onto the same specific factor (in addition to their loadings on G). Finally, there is a residual variable ε_{ik} for each observed variable Y_{ik} ($i = 1, \ldots, I_k$; I_k: number of indicators i belonging to domain k). The specific factors are assumed to be uncorrelated with all other specific factors as well as with G. The residual variables are assumed to be uncorrelated with (a) each other, (b) G, and (c) all latent S_k. This model represents Spearman's (1904) basic idea of a general factor and uncorrelated specific factors.

The hierarchical *G*-factor model is depicted in Figure 2. In this model, there is one first-order factor for each domain (F_k). In addition, there is a general second-order factor (G) having an influence on all first-order factors. The latent residuals S_k indicate the specific part of a domain factor F_k that is not determined by the general factor G. The residual variable ε_{ik} of an observed variable Y_{jk} indicates the unique part of an observed variable that is not shared with the other observed variables. All residual variables S_k and ε_{ik} are uncorrelated with each other as well as with G. When there are only three domains this model is equivalent to a first-

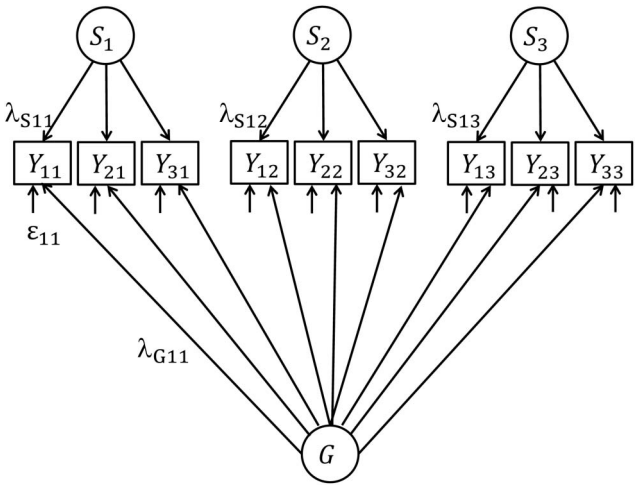

Figure 1. Traditional bifactor model with one general factor (G), three specific factors (S_k) and three observed variables Y_{ik} per domain. ε_{ik}: error variables, λ_{Gik}: G-factor loadings, λ_{Sik}: specific factor loadings $k = 1, \ldots, K$; K: number of domains; $i = 1, \ldots, I_k$; I_k: number of indicators i belonging to domain k. For simplicity, not all parameters and variables are labeled.

order correlated-factors model without a G-factor (see Figure 3). In the case of more than three domains the hierarchical G-factor model is more restrictive. The hierarchical G-factor model is formally equivalent to the testlet model (Rijmen, 2010), a model that has been developed for analyzing testlets (Wainer et al., 2007).

Conceptually, the variables S_k represent the specific part of a domain that is not due to the general factor in both types of models (see Figure 1 and 2). The two models are not generally equivalent. However, if one puts some specific restrictions on the bifactor model the two models can be defined in such a way that they are equivalent and that the G and S_k variables are exactly the same in both models (Mulaik & Quartetti, 1997; Rijmen, 2010; Yung, Thissen, & McLeod, 1999). Besides these two types of models, there are other models such as the two-tier model (Cai, 2010), in which there are different general factors for different domains (groups) of measures. In the present article, we focus on the bifactor model, because it is generally less restrictive than the hierarchical G-factor model and currently the most widely applied model.

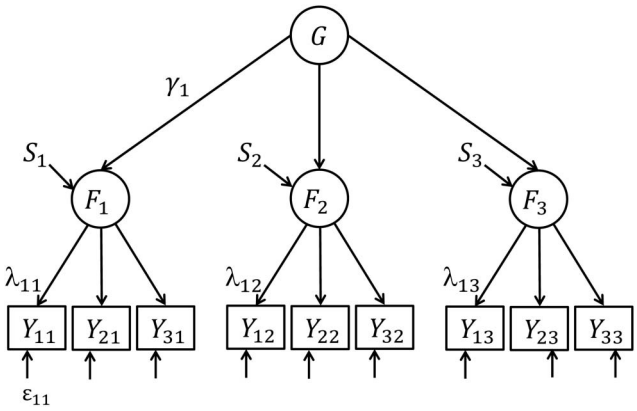

Figure 2. Hierarchical G-factor model with one general factor (G), three first order domain-specific factors (F_k) and three observed variables Y_{ik} per domain. ε_{ik}: error variables, S_k: latent domain-specific residuals, γ_k: G-factor loadings, λ_{ik}: specific domain factor loadings $k = 1, \ldots, K$; K: number of domains; $i = 1, \ldots, I_k$; I_k: number of indicators i belonging to domain k. For simplicity, not all parameters and variables are labeled.

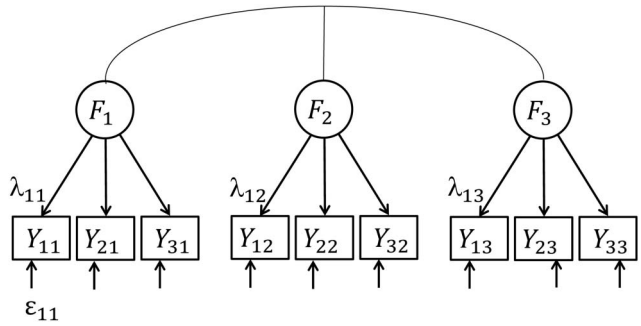

Figure 3. Multidimensional CFA model with three correlated first-order domain-specific factors (F_k) and three indicator variable Y_{ik} per domain. ε_{ik}: error variables; λ_{ik}: specific domain factor loadings $k = 1, \ldots, K$; K: number of domains; $i = 1, \ldots, I_k$; I_k: number of indicators i belonging to domain k. For simplicity, not all parameters and variables are labeled.

Anomalous Results

Applications of bifactor models frequently result in anomalous results or even in improper solutions. We consider results as anomalous when they are not in line with the general structure of the bifactor or the hierarchical G-factor model. Many applications report that at least one specific factor S_k had a negative variance estimate, a positive but nonsignificant variance estimate, and/or a full set of nonsignificant factor loading estimates (e.g., Brown, Finney, & France, 2011; Chen, West, & Sousa, 2006). In addition, in some cases specific factors are found to be correlated. These results are in contrast to the basic idea of the models (and related psychological theories) that each domain of a construct and each valid indicator of a domain can be decomposed into a general factor having an influence on all domains and a specific factor being unique to a domain and uncorrelated with other domains. We explain in detail later why we consider such empirical results to be problematic.

In order to illustrate the frequency of problems encountered in bifactor models, we conducted a more in-depth literature search of recently published bifactor applications. A PSYCInfo search of studies using the bifactor approach published between 2013 and 2014 revealed 143 articles for the search terms "bi-factor" or "bifactor" (in all fields). Out of these articles, 82 articles presented applications of a bifactor model of confirmatory factor analysis (CFA). About 61% ($n = 50$ articles) of the articles included in our search showed anomalous results (see Tables 1–3 for a list of articles that showed anomalous results). This is a lower bound as not all articles with applications provided sufficient information (e.g., tests of significance of factor loadings or variances) required for evaluating the applications in detail.

We found that about 16% ($n = 13$) of the studies reported at least one specific factor variance estimate as not significantly

Table 1
Studies Published in 2013/2014 in Which One or More Facet-Specific Factors Did Not Have a Significant Variance Estimate

Construct	Authors
Arithmetic, learning, and reading	Norwalk, diPerna, and Lei (2014)
Client-centered care	Muntinga, Mokkink, Knol, Nijpels, and Jansen (2014)
Cognition	Gavett, Crane, and Dams-O'Connor (2013)
Depression	Blanco et al. (2014)
Intelligence	Beaujean, Parkin, and Parker (2014); Booth et al. (2013)
Internet addiction	Watters, Keefer, Kloosterman, Summerfeldt, and Parker (2013)
Maladaptive cognitive intrusions	Meyer and Brown (2013)
Motivation for educational attainment	Cham, Hughes, West, and Im (2014)
Psychopathological symptoms	Tackett et al. (2013); Urbán et al. (2014)
Psychotic experiences	Betts, Williams, Najman, Scott, and Alati (2014)
Verbal fluency and creativity	Silvia, Beaty, and Nusbaum (2013)

different from zero. In about 48% (n = 39) of the applications irregular loading patterns occurred in which loadings on specific factors and/or the general factor did not differ significantly from 0. In about 5% (n = 4) of applications, some or all specific factors were correlated.

Vanishing Specific Factors

One unexpected result when applying bifactor models is when one or more specific factors vanish empirically. This is the case when a specific factor has a variance estimate that is not significantly different from zero or that is even negative. Studies published in 2013 and 2014 with this unexpected result are presented in Table 1. Beaujean, Parkin, and Parker (2014), for example, proposed a model with one *G*-factor and five specific factors (verbal comprehension, visual spatial, fluid reasoning, working memory, processing speed) for the Wechsler Intelligence Scale for Children. They found that the specific factors for fluid reasoning and working memory were not necessary for explaining the cova-

Table 2
Studies Published in 2013/2014 in Which Some Indicators Did Not Load Significantly on Facet-Specific Factors

Construct	Authors
Anxiety	Balsamo et al. (2013); Gomez (2013)
Anxiety and depression	Burns, Höfer, Curry, Sexton, and Doyle (2014); Luciano, Barrada, Aguado, Osma, and García-Campayo (2014)
Attention-deficit/hyperactivity disorder	Burns, de Moura, Beauchaine, and McBurnett (2014); Gomez, Kyriakides, and Devlin (2014)
Behavioral risk	diStefano, Greer, and Kamphaus (2013)
Burnout	Mészáros, Adám, Szabó, Szigeti, and Urbán (2014)
Callous-unemotional traits	Byrd, Kahn, and Pardini (2013)
Career motivation	Deemer, Smith, Thoman, and Chase (2014)
Caregiver interaction behavior	Colwell, Gordon, Fujimoto, Kaestner, and Korenman (2013)
Dark triad (narcissism, psychopathy, Machiavellianism)	Jonason, Kaufman, Webster, and Geher (2013)
Depression	Brouwer, Meijer, and Zevalkink (2013); Young, Hutman, Enggasser, and Meesters (2014)
Disgust	Olatunji, Ebesutani, Haidt, and Sawchuk (2014)
Emotional disorders	deSousa, Zibetti, Trentini, Koller, Manfro, and Salum (2014)
Emotional distress	Hyland, Shevlin, Adamson, and Boduszek (2013)
Ethnic identity	Yap et al. (2014)
Fatigue	Varni, Beaujean, and Limbers (2013)
Foreign language listening	Cai (2013)
Inclusive practices of teachers	Park, Dimitrov, Das, and Gichuru (2014)
Interpersonal sexual objectivation	Davidson, Gervais, Canivez, and Cole (2013)
Irritability	Burke et al. (2014)
Loneliness	Grygiel, Humenny, Rebisz, Switaj, and Sikorska (2013)
Medically unexplained syndromes	Witthöft, Hiller, Loch, and Jasper (2013)
Motivation for educational attainment	Cham, Hughes, West, and Im (2014)
Psychological well-being	Chen, Jing, Hayes, and Lee (2013)
Psychopathological symptoms	Urbán et al. (2014)
Quality of life	Garin et al. (2013); Zheng, Chang, and Chang (2013)
Reasoning	Primi, Rocha da Silva, Rodrigues, Muniz, and Almeida (2013)
Strength and difficulties	Kóbor, Takács, and Urbán (2013)
Sun protection behavior	Tripp et al. (2013)
Susceptibility to emotional contagion	Lo Coco, Ingoglia, and Lundqvist (2014)
Teacher-child interactions	Hamra, Hatfield, Pianta, and Jamil (2014)
Youth antisocial behavior	Tackett, Daoud, de Bolle, and Burt (2013)

Table 3
Studies Published in 2013/2014 in Which Some Indicators Did Not Load Significantly on the G-Factor

Construct	Authors
Irritability	Burke et al. (2014)
Callous-unemotional traits	Byrd et al. (2013)
Strength and difficulties	Kóbor et al. (2013)
Tinnitus acceptance	Weise, Kleinstäuber, Hesser, Westin, and Andersson (2013)

riances of the observed variables. A similar result was found in the very first application of the bifactor model by Holzinger and Swineford (1937) who postulated a bifactor model with four specific factors for mental abilities. Holzinger and Swineford found only three specific factors to be substantial. In addition, two of the observed variables had loadings only on the *G*-factor, but not on the hypothesized specific factor. One might argue that this is not a problem as the general factor could be interpreted in terms of this domain representing a particularly good indicator of *G*. However, only in very few studies the presence of fewer than *K* specific factors was expected by theory and modeled accordingly (for exceptions see, e.g., Tackett et al., 2013; Watters, Keefer, Kloosterman, Summerfeldt, & Parker, 2013). In almost all studies this result was unexpected and at odds with the basic idea that each domain of a multidimensional construct represents a combination of a common *G*-factor and a specific factor.

Irregular Loading Patterns

A frequently observed result in our review was that loadings of different indicators on a given specific factor often differ strongly, with some indicators having very small loadings, non-significant loadings, or even negative loadings. Examples are listed in Table 2. In most cases, this result is at odds with a researcher's expectation that all variables should load positively and significantly on all factors. This result is often unexpected because the loading pattern for the same data may be regular when a simpler model with correlated domain-specific first-order factors (and no *G* factor) is considered. In many studies that report both a model with correlated first-order factors and a bifactor model the loading pattern on at least one of the specific factors changes strongly. For example, Davidson, Gervais, Canivez, and Cole (2013) analyzed the Interpersonal Sexual Objectification Scale with a CFA model with three correlated first-order factors and a bifactor model. In the first-order factor model the standardized factor loadings on the body evaluation factor ranged between .71 and .85, indicating a relatively high homogeneity of the indicators of this construct. In the bifactor model, however, the standardized factor loadings on the specific body evaluation factor ranged between .07 and .72, indicating a high degree of heterogeneity. This result is theoretically unexpected because one would expect similar loading patterns in both the correlated first-order factor model and the bifactor model. Another frequently seen result is that not all indicators load on the *G*-factor, which is usually unexpected as well. Examples are given in Table 3.

Correlated Specific Factors

Another anomalous result is that specific factors are not independent from each other as is postulated in the bifactor model and a priori hypothesized by many researchers applying the bi-factor model. Examples can be found in Table 4. If specific factors are correlated there are additional sources of common variance in addition to the *G*-factor (e.g., minor factors). This is in contrast to most *G*-factor theories.

In summary, our literature review showed that anomalous results are quite common in applications of the bifactor approach. We think that the frequency with which these issues occur warrants a critical evaluation of the bifactor model (and related models) for analyzing *G*-factor structures in the research areas in which they are typically applied. This does not mean that the bifactor model and related models are generally inappropriate. However, as we show later in this paper, the application of such models requires different measurement designs than the ones that are traditionally used. Besides the anomalous empirical results there are conceptual problems that have not been sufficiently addressed in the literature so far. These issues are described in the next section.

Conceptual Problems

Psychometric Meaning of *G* and Specific Factors

One important question is what the *G*-factor and the specific factors mean from a psychometric perspective. A traditional assumption is that the *G*-factor is a "common factor" that has an influence on all domains of a multidimensional construct. The specific factors are considered "residualized factors" (Reise, 2012, p. 691). This would imply that the means of the specific factors have to be zero in bifactor models, because residuals in regression theory have means of zero by definition (e.g., Steyer, 1988). However, Chen, West, and Sousa (2006) mention as an advantage of the bifactor model that the means of the *G*-factor as well as the specific factors can be compared between groups (of individuals) if measurement invariance between groups can be established. But what is the meaning of specific factors and the *G*-factor if all factors can have means differing from zero? If their means are estimated, should these factors still be interpreted in the same way? Under which conditions is it reasonable to allow the means of specific factors to differ from zero? Or is this not reasonable at all?

(Non-)Invariance of the *G* Factor Across Different Sets of Domains

If the *G*-factor represents a meaningful construct, its influence on the different domains as well as the loadings on the specific

Table 4
Studies Published in 2013/2014 With Correlated Facet-Specific Factors

Construct	Authors
Intelligence	Watkins and Beaujean (2014)
Irritability	Burke et al. (2014)
Male role norms	Levant, Hall, and Rankin (2013)
Strengths and difficulties	Kóbor et al. (2013)

factors should not change when a researcher adds or removes individual domains (Reise, 2012). That means, for example, that the *G*-factor of intelligence should stay the same (i.e., "general") when one takes out four of 10 domains of intelligence. From a conceptual perspective this assumption is very reasonable. If there is a general ability why should its influence on a specific domain change when other domains are taken out? Reise (2012), however, found that the *G* factor loadings can change when domains are removed. This causes some conceptual problems, as it means that *G* factors as measured in the bifactor and related models are not generally invariant across different sets of domains used to measure them. This can cause problems, for example, in literature reviews or meta-analyses that summarize data from different studies or in so-called conceptual replications in which different domains were used to measure a given *G* factor, because the *G* factors may not be comparable across studies. How can a *G*-factor model be defined in such a way that the influence of the *G*-factor on specific domains does not depend on the presence or absence of other domains in the model?

Correlations Between *G* and Specific Factors

Brunner et al. (2012) mentioned as a limitation of the bifactor approach that general and specific factors are assumed to be mutually uncorrelated. Allowing correlations between the *G*-factor and specific factors would change the meaning of the *G*-factor and the specific factors and can cause identification problems. Under which conditions would it make sense to estimate a correlation between the *G*-factor and specific domains or are such correlations not reasonable at all?

Correlations Between Specific Factors

Based on the empirical finding of nonzero correlations between specific factors, some authors have proposed bifactor models in which all specific factors are correlated (e.g., Levant et al., 2013). But what does a *G*-factor mean if all domains are correlated indicating that there is a common source of variance in addition to *G*? Under which conditions would it make sense to have (some) correlated specific factors or is this assumption not reasonable at all?

In the next section, we show how SMT can guide researchers in finding answers to these fundamental questions that we consider of high importance for applying and interpreting the results of the bifactor model and related models in psychology. Moreover, we demonstrate that SMT can help understand why anomalous results can occur.

Basic Ideas of Stochastic Measurement Theory

In models of confirmatory factor analysis it is assumed that the covariances of the observed variables can be explained by latent variables. In most applications these latent variables are based on theoretical considerations without making it explicit how these variables can be formally defined using probability theory. However, factors in CFA are commonly considered to be random variables (e.g., Bollen, 1989). For example, some estimation methods in confirmatory factor analysis (such as maximum likelihood estimation) require that the latent variables follow a multivariate normal distribution. In probability theory, random variables assign values to the elements of a sample space (e.g., Hays, 1994). For example, if one tosses a coin there are two possible outcomes (head, tail). Therefore, the sample space Ω consists of two elements: $\Omega = \{\text{head, tail}\}$. A random variable Y assigns values to the elements ω of the sample space, for example, the value 1 to head and the value 0 to tail: $Y(\text{head}) = 1$ and $Y(\text{tail}) = 0$.

In order to understand the meaning of the values of a random variable one has to know the sample space and the assignment rule. The sample space characterizes the random experiment, that means, the procedure that leads to the possible outcomes. The sample space $\Omega = \{\text{head, tail}\}$ characterizes the random experiment "tossing a coin." If factors are supposed to be interpreted as random variables, the sample space on which they are defined has to be explicated. This is necessary in order to understand what the values of a factor mean. This argument is important for the present considerations, because—as we show later on—not all factors in CFA can be defined as random variables. That is, there can be CFA models in which the factors are not random variables and hence not well-defined in the sense of SMT. In cases in which this is not possible, it remains an open question what the factors mean and whether the model considered is a reasonable model.

Although it is relatively uncommon for researchers to think about these issues when specifying and testing factor models, the history of psychometrics has shown that taking the approach of explicitly defining latent variables as random variables can help clarify important misconceptions about psychometric theories and about measurement models. For example, Novick (1966) has shown how the random experiment underlying classical test theory (CTT) can be explicated and in which way the variables of CTT can be explicitly defined. The aim of his article was to "show that classical test theory may be placed on a firm theoretical foundation, that its necessary assumptions are very weak and hence generally satisfied" (p. 1). His article was motivated by the fact that CTT suffered "from some imprecision of statement so that, from time to time, controversies arise that appear to raise embarrassing questions concerning its foundations" (p. 1).

In our view, the field is currently in a similar situation regarding the bifactor model and other related *G*-factor models. Many researchers apply these models, because they find them intuitively plausible for modeling multidimensional constructs. These models seem to fit the basic theoretical assumption of a general factor underlying different domains of a multidimensional construct in addition to specific factors that are specific to a domain. However, from the perspective of SMT, the *G*-factor in bifactor and related models is a formally well-defined random variable only if a very specific type of random experiment is considered. Similar to Novick (1966), who clarified the measurement theoretical foundations and assumptions of CTT, we would like to place the bifactor model "on a firm theoretical foundation" (p. 1). We show that many conceptual problems can be resolved (and anomalous empirical results understood) by taking this approach.

SMT is a theory that shows how latent variables can be defined as random variables on a well-explicated sample space. Defining the models of CTT as stochastic measurement models, for example, made it possible to clarify what assumptions have to be made to define the variables and what the consequences of these definitions are. In particular, key properties of the true score and error variables could be derived. For example, it could be shown that the

uncorrelatedness of the true score and the error variables are not assumptions, but logical consequences of defining the true score as an expected value (Zimmerman, 1975, 1976). Similarly, it could be shown that the expected value of an error variable in CTT is always zero by definition. Moreover, important measurement theoretical questions concerning the uniqueness and meaningfulness of latent variables could be clarified (Steyer, 1989). Using the concepts of conditional expectations (Steyer, 1988; Zimmerman, 1975, 1976) it could be shown that models of CTT are formally equivalent to models of item response theory (IRT) and that they have testable consequences (Steyer, 1989; Steyer & Eid, 2001). This made it possible to integrate CTT and IRT into a common psychometric framework such as the generalized linear item response theory (Eid & Schmidt, 2014; Mellenbergh, 1994).

In the next section we show how the factors of a multidimensional factor model with domain-specific first-order factors (without a *G*-factor) can be defined as random variables on a set of possible outcomes. We then show how *G*-factor models can be defined. This has strong implications for the application and interpretation of the bifactor model and related *G*-factor models. We then show how alternative models for *G*-factor structures can be defined. This approach leads to the definition of bifactor models that are in line with many of the anomalous results found in empirical applications. From the point of view of SMT these results have to be expected for theoretical reasons. For didactical reasons, we only refer to the set of possible outcomes and not to the whole probability space (for the definition of probability spaces see, e.g., Steyer, 1988, 1989; Zimmerman, 1975). We start with a single-level sampling process and continue with a two-level sampling process.

SMT has some similarities with generalizability theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In generalizability theory different facets (here termed GT-facets) that have an influence on observed variables are considered. In GT, a researcher has to decide explicitly whether a level of a GT-facet is randomly sampled from a universe of units (random effects) or whether a level of a GT-facet is not randomly selected from a universe but selected by the researcher according to some criteria (fixed effects). For example, consider two researchers who are interested in the assessment of mathematical abilities. Researcher A randomly selects four items from a universe consisting of all multiplication tasks that are possible for two two-digit numbers. Researcher B formulates one addition task, one subtraction task, one multiplication task and one division task. In the case of Researcher A the items are interchangeable. Consequently, differences between items (item effects) are random. In the case of Researcher B the items are not interchangeable, but structurally different. Item effects are fixed effects. The distinction between random and fixed effects has important consequences for the definition of variance components (e.g., generalizability coefficients). In GT different designs can be considered in which GT-facets can be crossed and/or nested. Similar to GT, SMT also requires to explicitly formulate whether a unit of a domain is randomly selected or fixed. It also allows decomposing an observed variable into different components and defining variance components. However, SMT also differs from GT in some important ways. In particular, SMT deals with some limitations of GT that have been formulated by Hunter (1968) and Werts, Linn, and Jöreskog (1974). According to Hunter (1968) the assumptions on which models of GT are based are often not explicitly described, mathematical proofs are missing, and the sampling theory is not well explicated. According to Werts et al. (1974) CFA allows to consider less restrictive measurement models than GT in the case of multiple items. SMT can be considered as an extension of GT by integrating Hunter's (1968) and Werts et al.'s (1974) approaches. Because of the strong similarities between GT and SMT and because many researchers might be more familiar with GT than with SMT we will also refer to concepts of GT when we introduce the basic concepts of single- and two-level sampling processes.

Single-Level Sampling Process

In the single-level sampling process, observational units u (e.g., individuals) are randomly selected from a set Ω_U of possible observational units u. Then, scores of these observational units on different items or scales are registered. For example, a researcher interested in the intensity of negative emotions might ask individuals to rate with which intensity they experience emotions of fear, anger, shame, and sadness. Because measurement error cannot be avoided in the measurement of most psychological attributes, the recorded outcomes are considered to stem from a set of possible outcomes Ω_{M_i} for a measurement (item or scale) i (Novick, 1966; Zimmerman, 1975). If there are p different items or scales, the set of possible outcomes (Ω) of this random experiment can be defined in the following way (Eid & Koch, 2014; Steyer, 1989):

$$\Omega = \Omega_U \times \Omega_{M_1} \times \ldots \times \Omega_{M_i} \times \ldots \times \Omega_{M_p}. \quad (1)$$

In order to define latent variables, two mappings are required (Steyer, 1989). The mapping $p_U : \Omega \rightarrow \Omega_U$ maps the possible outcomes into the set of observational units. The mapping $Y_i : \Omega \rightarrow \mathbb{R}$ maps the possible outcomes of item or scale i into the set of real numbers ($\mathbb{R}$). These general concepts apply to both continuous and categorical observed variables. For simplicity, we explain the concepts for continuous observed variables. Given that bifactor models for dichotomous and ordinal observed variables have gained increasing interest in IRT and many applications (e.g., Cai, 2010, 2015; Cai et al., 2011; Cho, Cohen, & Kim, 2014; DeMars, 2006, 2013; Gibbons et al., 2007; Gibbons & Hedeker, 1992; Han & Paek, 2014; Jeon, Rijmen, & Rabe-Hesketh, 2013, 2014; Liu & Thissen, 2014; Yang et al., 2013), we show in the Appendix how the latent variables of *G*-factor models can be properly defined for ordinal response variables.

Definition of Latent Variables in Classical Test Theory

The latent factors underlying continuous observed variables in a CFA model can be defined on the basis of true score variables as defined in CTT (Eid & Koch, 2014; Steyer, 1989). A true score variable τ_i of an item or scale i is defined as the conditional expectation of Y_i given the person variable p_U: $\tau_i = E(Y_i \mid p_U)$. A value $E(Y_i \mid p_U = u)$ of the true score variable is the outcome that is expected for a specific individual u and a specific item or scale i. The measurement error variable ε_i is defined as the difference between the observed variable Y_i and the true score variable τ_i: $\varepsilon_i = Y_i - E(Y_i \mid p_U)$.

The true score variables can be used to define factors in CFA models as random variables on the same probability space. For

example, if one assumes that all true score variables belonging to different items or scales i and j are linear functions of each other ($\tau_i = \alpha_{ij} + \lambda_{ij}\tau_j$), a common factor η can be defined as a linear function of an arbitrary true score variable. This has the advantage that the common factor is defined as a random variable on a well-explicated sample space. For example, if one defines $\eta = \tau_1$ the common factor η equals the true score variable of the first item or scale, with that true score variable itself being well-defined as $\tau_1 = E(Y_1 \mid p_U)$. This definition of the common factor η implies the following equation for the unidimensional (congeneric) measurement model: $\tau_i = \alpha_i + \lambda_i\eta$ with $\alpha_1 = 0$ and $\lambda_1 = 1$. This unidimensional model is related to the Person × Item design of GT. The GT-facet *persons* is considered random, the GT-facet *items* is considered fixed.

Definition of Latent Variables: Multidimensional Models

A multidimensional CFA model with multiple correlated first-order factors (see Figure 3) can be defined by making the assumption that only true score variables that belong to the same domain of a multidimensional construct are linear functions of each other (Eid & Koch, 2014). If the index k represents the specific domain, the multidimensional factor model is defined by the assumption $\tau_{ik} = \alpha_{ijk} + \lambda_{ijk}\tau_{jk}$, which is equivalent to the assumption $\tau_{ik} = \alpha_{ik} + \lambda_{ik}\eta_k$. That means that there is a common factor for each domain with a congeneric measurement model within domains. For example, if one sets $\alpha_{1k} = 0$ and $\lambda_{1k} = 1$, the common domain-specific factor η_k is the true score variable of the first indicator pertaining to this domain. In the one-factor model as well as the multidimensional factor model, it is assumed that all error variables ε_i are uncorrelated. This multidimensional model is related to the Person × Constructs design of GT with items nested within constructs. The GT-facet *persons* is random, the GT-facets *constructs* and *items* are fixed.

Based on SMT it is easy to define common first-order factors that are related to different groups of observed variables as functions of the true score variables. This ensures that the common factors are random variables on a well-explicated random experiment. These factors are clearly defined as true score variables or functions of true score variables. In order to define a common higher-order factor (as in the hierarchical G-factor model; see Figure 2) or a G-factor (as in the bifactor model; see Figure 1) as a random variable based on the single-level random experiment explained so far, it would have to be shown that these factors can be expressed as functions of the true score variables. This, however, does not seem to be possible (Eid & Koch, 2014). At least, we did not succeed in defining the latent variables of the models presented in Figures 1 and 2 as random variables on a well explicated set of possible outcomes.

In order to define G-factors in addition to first-order domain factors a two-level sampling process is required. It will become clearer why the G-factor and the specific factors in Figures 1 and 2 cannot be defined as random variables based on a single-level sampling process after we present the two-level sampling process.

Two-Level Sampling Process

In order to define G factors in the way it is done in the bifactor approach, a more complex type of random experiment has to be considered. With respect to the different domains of a construct, the domains have to be randomly chosen from a set of possible domains. Whereas the domains are considered levels of a fixed GT-facet in the terminology of GT in the single-level sampling process, they are considered levels of a random GT-facet in the two-level sampling process. We will refer to a design where domains are nested within individuals first, and we will then briefly discuss a design in which individuals and domains are crossed. In the nested design domains are randomly chosen from a set of domains for an individual that has been selected from a set of individuals. That means that the randomly selected domains differ between individuals. For testlet research this would mean that different individuals have to work on different testlets (e.g., text passages). Another example for a nested design is when examiners grade different essays (e.g., in writing ability research). In clinical psychology, researchers might be interested in phobic reactions to spiders and randomly select spiders from a set of spiders whereas the spiders differ for different individuals. The nested design is also used in longitudinal data analysis such as in latent state-trait (LST) theory (Eid, 1996; Steyer, Mayer, Geiser, & Cole, 2015; Steyer, Schmitt, & Eid, 1999; Steyer, Ferring, & Schmitt, 1992). Consider, for example, a researcher who is interested in the stability of mood across time. He randomly selects a sample of individuals and measures each individual's mood with multiple indicators on two occasions of measurement. He assumes that the individuals are in different (inner) situations on different occasions of measurement. The (inner) situations differ between individuals. He wants to separate a stable component (the mood trait) from time-specific influences by specifying a bifactor model. The G factor would represent the stable component that could be called mood trait or habitual mood level. The two time-specific factors would represent the occasion-specific influences (with k indicating the occasion of measurement or situation [instead of domain] and $K = 2$ in this example). The LST model for this application is equivalent to the bifactor model presented in Figure 1 (but with only two specific factors). According to an LST theory of mood (e.g., Eid & Diener, 2004; Eid, Schneider, & Schwenkmezger, 1999) the occasion-specific influences are due to situational influences and/or the interaction between the individual and the situation. This nested design is also used in multirater studies when a randomly selected target is assessed by randomly selected peers, where the peers differ between targets (e.g., Nussbeck, Eid, Geiser, Courvoisier, & Lischetzke, 2009). Peers are considered as domains. We will illustrate the random experiment of the nested design with respect to this multiple rater example.

Based on a nested design, the G-factor and the specific factors can be defined as random variables on the following random experiment (Eid, 1996; Steyer et al., 1992, 1999; for simplicity, we restrict ourselves to two domains and two items/scales per domain):

$$\Omega = \Omega_U \times \Omega_{D_1} \times \Omega_{M_{11}} \times \Omega_{M_{21}} \times \Omega_{D_2} \times \Omega_{M_{12}} \times \Omega_{M_{22}} \qquad (2)$$

According to this random experiment an individual u is drawn from a set of individuals (Ω_U). This is the same as in the single-level random experiment discussed earlier. Then, a domain (rater) is randomly drawn for this individual from a set of domains (raters; Ω_{D_1}), and the outcomes on two different items or scales are registered (elements of the sets $\Omega_{M_{11}}$ and $\Omega_{M_{21}}$). After that, a second domain (rater) is randomly drawn for each individual and

the scores on two items or scales are recorded. In addition to the mappings considered in the single-level random experiment, the mappings $p_{D_1} : \Omega \rightarrow \Omega_{D_1}$ and $p_{D_2} : \Omega \rightarrow \Omega_{D_2}$ can be defined. They map the possible outcomes into the sets of domains (raters). The mapping $Y_{ik} : \Omega \rightarrow \mathbb{R}$ maps the possible outcomes of item or scale i on occasion k into the set of real numbers. It is important to note that according to this type of random experiment the domains are considered interchangeable. It is related to the design of GT in which the random GT-facet *conditions* (here: domains) is nested with the random GT-facets *persons* and the fixed GT-facet *items* either nested within conditions (if the items differ between the conditions) or crossed with the conditions (if the items do not differ between conditions).

Definition of Latent Variables

In this context, the true score variables are defined as $\tau_{ik} = E(Y_{ik} \mid p_U, p_{D_k})$, and the error variables as $\varepsilon_{ik} = Y_{ik} - E(Y_{ik} \mid p_U, p_{D_k})$. A value of the true score variable $\tau_{ik} = E(Y_{ik} \mid p_U = u, p_{D_k} = d_k)$ is the expected outcome of item or scale i for a specific individual u measured with respect to a specific domain d_k (rated by a specific rater). The conditional expectation $\xi_{ik} = E(\tau_{ik} \mid p_U) = E(E(Y_{ik} \mid p_U, p_{D_k}) \mid p_U) = E(Y_{ik} \mid p_U)$ is the latent general variable or G factor. A value $E(Y_{ik} \mid p_U = u)$ of this variable is the expected outcome of an item or scale Y_{ik} for an individual u. It is the expected value for an individual across the different domains (raters). A value of the residual $\zeta_{ik} = \tau_{ik} - \xi_{ik}$ represents the specific influences (e.g., rater-specific influences). In the context of a two-level sampling plan, an observed variable can be decomposed into a true score variable and an error variable:

$$Y_{ik} = \tau_{ik} + \varepsilon_{ik} \tag{3}$$

In addition, a true-score variable can be decomposed into a general and a specific variable:

$$\tau_{ik} = \xi_{ik} + \zeta_{ik} \tag{4}$$

As a consequence, an observed variable can be decomposed into a general variable, a specific variable and an error variable:

$$Y_{ik} = \xi_{ik} + \zeta_{ik} + \varepsilon_{ik} \tag{5}$$

Based on these decompositions, the factors of a bifactor model (see Figure 1) can be constructively defined as random variables. If one assumes that all general variables ξ_{ik} are linear functions of each other ($\xi_{ik} = \alpha_{Gijk} + \lambda_{Gijk}\xi_{jk}$) a general factor ξ can be defined as a linear function of a general variable ξ_{ik} that is arbitrarily chosen. That means that it does not matter which variable ξ_{ik} is taken. For example, if one defines $\xi = \xi_{11}$ one obtains the equation $\xi_{ik} = \alpha_{Gik} + \lambda_{Gik}\xi$ with $\alpha_{G11} = 0$ and $\lambda_{G11} = 1$.

It is essential to recognize that defining the general factor in this way gives the general factor a very clear meaning. In our example, it is the general variable of the first item/scale belonging to the first domain (rater). In a similar way a specific factor can be defined by assuming that $\zeta_{ik} = \lambda_{Sijk}\zeta_{jk}$ which is equivalent to the assumption that $\zeta_{ik} = \lambda_{Sik}\zeta_k$. If one chooses, for example, $\lambda_{S11} = 1$, the specific factor ζ_1 is the specific variable of the first item/scale Y_{11}. Note that there are no additive constants (intercepts) for the specific variables because they are defined as residual variables so that their expectations are zero by definition. As a consequence of the two assumptions according to which (a) all general variables are linear functions of each other, and (b) the specific variables that are measured on the same occasion of measurement are linear functions of each other, each observed variable Y_{ik} can be decomposed into a linear combination of a general (ξ) and a specific factor (ζ_k) and an error variable:

$$Y_{ik} = \alpha_{Gik} + \lambda_{Gik}\xi + \lambda_{Sik}\zeta_k + \varepsilon_{ik} \tag{6}$$

This model is identical to the bifactor model. In the case of interchangeable domains (e.g., raters) it is additionally assumed that the intercepts and loadings do not differ between domains (Nussbeck et al., 2009). What are the advantages of defining general and specific factors in this way? First of all, this definition gives the factors a clear meaning. They are functions of conditional expectations that have a very clear meaning. It also shows in which way factors can be defined as random variables. This helps to avoid including factors in a CFA model that do not have a clear psychometric meaning. Moreover, the definition of the factors has some important consequences. Because the specific factors are linear functions of residual variables their expected values (means) have to be zero. Moreover, they have to be uncorrelated with the *G*-factor. If one defines the factors in this way it would not make sense to estimate means of specific factors or correlations between general and specific factors.

Whereas the nested design has often been realized in longitudinal data analysis and multiple (interchangeable) rater studies, it is rather unusual in other areas of psychology. In many studies individuals do not differ in the randomly selected domains but they were assigned the same domains and items. Domains and persons are fully crossed (in the terminology of GT). A possible example is a design in which at first different domains are randomly selected from a universe of domains and then all randomly selected individuals are exposed to the same (randomly selected) domains. For example, in research on reading ability the text passages can be randomly selected from a universe of text passages and all individuals have to work on the same text passages and their reading abilities are assessed. Because the domains (text passages) are randomly selected, a *G*-factor and specific factors can be defined in the same way as in the nested model and a bifactor model can be applied. However, because individuals and domains are fully crossed and there are multiple observed scores for a person-domain combination also the interaction between the persons and the domains can be identified and separated from random domain effects. Consequently, in addition to the bifactor structure also a random domains factor can be defined. In the terminology of multilevel analysis this model would be a latent cross-classified model which is more complex than the bifactor model, and will therefore not be further considered in the present paper (for a definition and application of such a model see Koch et al., 2016).

In summary, the above discussions show that based on a single-level sampling process, it is not possible to define the *G*-factor and specific factors of a bifactor model. We were, however, able to define general and specific factors when the domains are randomly selected. This shows that the general and specific factors would not have a clear meaning when a single-level sampling process is considered. From the perspective of SMT, it is questionable whether the bifactor model is a reasonable model when a single-level sampling design is considered. However, the above discussions make clear that a bifactor model is a very reasonable model

if a two-level sampling process is considered. In this case both the general and specific factors have a very clear meaning.

Areas of Application of the Bifactor and Related Models

What are the areas of application in which the bifactor model is reasonable? From the perspective of SMTM, reasonable applications would be ones that involve a two-step sampling process as described above. In other words, the different domains would have to be considered as interchangeable (i.e., randomly chosen from a set of equivalent domains). For example, bifactor models or related models can be applied in multiple-rater studies when the raters are interchangeable (i.e., Nussbeck, et al., 2009). In applications in which domains or raters are selected truly at random from a set of interchangeable domains or raters, we could expect the *G* factor to be invariant across different sets of randomly selected domains.

In applied research, domains are typically not formally selected from a universe of domains. However, the application of the bifactor model would still make sense if it is reasonable to assume that the domains are interchangeable. For example, in research on reading comprehension using testlets the application of the bifactor models is reasonable if the text passages (the contexts) can be considered interchangeable. However, the application of the bifactor model might not be reasonable if the different domains cannot be considered interchangeable, but are structurally different. For example, in Holzinger and Swineford's (1937) first application of the model the different domains of intelligence that were considered (spatial, mental speed, motor speed, verbal) were not interchangeable but structurally different. We believe that in many other applications of the bifactor model (e.g., in the applications listed in Tables 1 to 4) the domains also cannot be considered as interchangeable but are structurally different. In the case of structurally different domains, the bifactor model might not be appropriate. For these applications, which are based on a single-level sampling process, other models are required.

How can a *G* Factor be Defined in a Single-Level Sampling Process?

If it is not possible to define the bifactor model based on a single-level sampling process how can a *G* factor and specific factors be constructively defined as random variables within a single-level sampling process? There are several possibilities. We now present two different ways that are in line with the typically found anomalous results presented in Tables 1 and 2. The first way is based on the idea of taking one domain as comparison standard or reference domain to which the other domains are compared. The second way is to take one item or scale as reference indicator to which the other items/scales are compared.

Model With a Reference Domain: The Bifactor-($S - 1$) Model

One way to define a *G* factor in a single-level random experiment is to take one domain as a reference domain. Without loss of generality, we may choose the first domain ($k = 1$) as reference domain and take the first indicator of this domain ($i = 1$) as a reference indicator. This choice of the reference domain and indicator depends on a researcher's theory and goals. Our starting point is the true score variable τ_{11}. A *G*-factor model can be defined by the following steps (see Table 5):

1. We assume that all true score variables τ_{i1} belonging to the reference domain are linear functions of the reference true score variable τ_{11}:

$$\tau_{i1} = \alpha_{i1} + \lambda_{Gi1}\tau_{11} \tag{7}$$

2. We assume that the regressions (conditional expectations) of all true score variables τ_{ik} belonging to a nonreference domain ($k \neq 1$) on the reference true score variable τ_{11} are linear:

$$E(\tau_{ik}|\tau_{11}) = \alpha_{ik} + \lambda_{Gik}\tau_{11} \tag{8}$$

3. For each domain we select the first indicator as reference indicator and assume that all regression residuals $\zeta_{ik} = \tau_{ik} - E(\tau_{ik} \mid \tau_{11})$ belonging to the same domain k are linear functions of the regression residual of the first indicator: $\zeta_{ik} = \lambda_{Si1}\zeta_{1k}$. A regression residual ζ_{ik} represents that part of the true score variable that is specific to the nonreference domain and cannot be predicted by the reference domain.

4. We assume that all error variables $\varepsilon_{ik} = Y_{ik} - \tau_{ik}$ are uncorrelated.

For three domains and three indicators per domain these assumptions imply a model that is depicted in Figure 4. In this model there is a *G* factor that has an influence on all items considered. The *G* factor equals the true score variable τ_{11}, that is, the true score variable of the reference indicator pertaining to the reference domain. 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 a domain that is not shared with the reference domain. The residual factor equals the residual variable of the reference indicator of the domain considered. The specific factors can be correlated. These correlations indicate partial relationships between domains after accounting for variance that all domains share with the reference domain. We call this model *bifactor-(S – 1) model* because there is one specific factor less than domains considered (with $S = K$, K: number of domains considered). This model is based on the idea of the CT-C($M - 1$) model that has been developed in the context of multitrait-multimethod analysis (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Eid et al., 2008; Geiser, Eid, & Nussbeck, 2008; Nussbeck, Eid, & Lischetzke, 2006). Mulaik and Quartetti (1997) have described such a model for analyzing general and specific abilities, but they assumed uncorrelated specific factors. Using SMT as the underlying theoretical framework shows that this model is a reasonable model for analyzing a *G* factor and specific factors, because both *G* and the specific factors can be defined as direct functions of true score variables. However, there are only very few applications in which such a structure has been hypothesized a priori, that is, *before* the data analysis (e.g., Brunner et al., 2010; Tackett et al., 2013; Watters et al., 2013). In most cases, such a structure was found as a result of an empirical analysis (see Table 1).

Table 5
Definition and Important Properties of the Bifactor-(S − 1) and Bifactor-(S·I − 1) Models

	Bifactor-($S-1$) model	Bifactor-($S\cdot I-1$) model
Model definition	(1) $\tau_{i1} = \alpha_{i1} + \lambda_{Gi1}\tau_{11}$ (2) $E(\tau_{ik} \mid \tau_{11}) = \alpha_{ik} + \lambda_{Gik}\tau_{11}$ (3) $\zeta_{ik} = \lambda_{Si1}\zeta_{1k}$, with $\zeta_{ik} = \tau_{ik} - E(\tau_{ik} \mid \tau_{11})$ and $k \neq 1$ (4) $Cov(\varepsilon_{ik}, \varepsilon_{jl}) = 0$, for $(i, k) \neq (j, l)$	(1) $E(\tau_{ik} \mid \tau_{11}) = \alpha_{ik} + \lambda_{Gik}\tau_{11}$ (2) $\zeta_{ik} = \lambda_{Si1}\zeta_{1k}$, with $\zeta_{ik} = \tau_{ik} - E(\tau_{ik} \mid \tau_{11})$ and $k \neq 1$ (3) $\zeta_{i1} = \lambda_{S21}\zeta_{21}$ and $k = 1$ (4) $Cov(\varepsilon_{ik}, \varepsilon_{jl}) = 0$, for $(i, k) \neq (j, l)$
General model equation	$Y_{ik} = \alpha_{ik} + \lambda_{Gik}G + \lambda_{Sik}S_k + \varepsilon_{ik}$, with $G = \tau_{11}$ $S_k = \zeta_{1k}$ $S_1 = 0$ $\alpha_{11} = 0$ $\lambda_{G11} = 1$ $\lambda_{S1k} = 1$ $\lambda_{Si1} = 0$	$Y_{ik} = \alpha_{ik} + \lambda_{Gik}G + \lambda_{Sik}S_k + \varepsilon_{ik}$, with $G = \tau_{11}$ $S_k = \zeta_{1k}$, for $k \neq 1$ $S_1 = \zeta_{21}$ $\alpha_{11} = 0$ $\lambda_{G11} = 1$ $\lambda_{S1k} = 1$, for $k \neq 1$ $\lambda_{S21} = 1$ $\lambda_{S11} = 0$
Consistency	$Con(\tau_{ik}) = \frac{\lambda^2_{Gik}Var(\tau_{11})}{Var(\tau_{ik})}$	
Specificity	$Spe(\tau_{ik}) = \frac{\lambda^2_{Sik}Var(\zeta_{1k})}{Var(\tau_{ik})}$	
Reliability	$Rel(Y_{ik}) = 1 - \frac{Var(\varepsilon_{ik})}{Var(Y_{ik})}$	
Properties	(1) General factor is uncorrelated with all facet-specific factors (2) Specific factors can be correlated (3) Mean values of facet-specific factors have to be 0 (4) Meaning of the *G* factor will not change when adding or removing facets (5) The meaning of the *G* factor changes when the reference facet changes	

The bifactor-($S-1$) model allows estimating the proportion of variance in a nonreference domain true-score variable that is determined by the *G*-factor (consistency coefficient)

$$Con(\tau_{ik}) = \frac{\lambda^2_{Gik}Var(\tau_{11})}{Var(\tau_{ik})}. \quad (9)$$

The counterpart of the consistency coefficient is the specificity coefficient:

$$Spe(\tau_{ik}) = \frac{\lambda^2_{Sik}Var(\zeta_{1k})}{Var(\tau_{ik})}. \quad (10)$$

It represents the proportion of specific variance in a nonreference domain true-score variable that is not shared with the reference true-score variable. The reliability coefficient is defined as in other latent variable models:

$$Rel(Y_{ik}) = 1 - \frac{Var(\varepsilon_{ik})}{Var(Y_{ik})}. \quad (11)$$

This model has some important properties and gives answers to conceptual questions that are related to bifactor models:

1. The *G* factor and the specific factors cannot be correlated because the specific factors are defined as residual factors with respect to the *G* factor.

2. The mean values of the specific factors have to be zero, because residual factors always have means of zero by definition.

3. The specific factors can be correlated. These correlations are partial correlations—corrected for common influences of the *G* factor.

4. The meaning of the *G* factor does not change when domains are added or removed, because the *G* factor is defined as the common factor of the reference domain. As long as the indicators of the reference domain do not change the *G* factor also does not change.

5. The meaning of the *G* factor changes when the reference domain changes. That means that the *G* factor is always the common factor of the reference domain.

6. The fit of the model can change when the reference domain changes (see applications below). These differences in the fit coefficients are due to the fact that choosing a different reference domain represents the potential item-heterogeneity in a different way. There are two ways to handle this difference in model fit. One possibility is to restrict the model in such a way that the fit does not change and equals exactly the fit of the multidimensional first-order model. Geiser, Eid, and Nussbeck (2008) have shown how this can be done in the context of multitrait-multimethod modeling. The second possibility is to fit a less restrictive model to the data that represents item heterogeneity more generally. We now present this second approach.

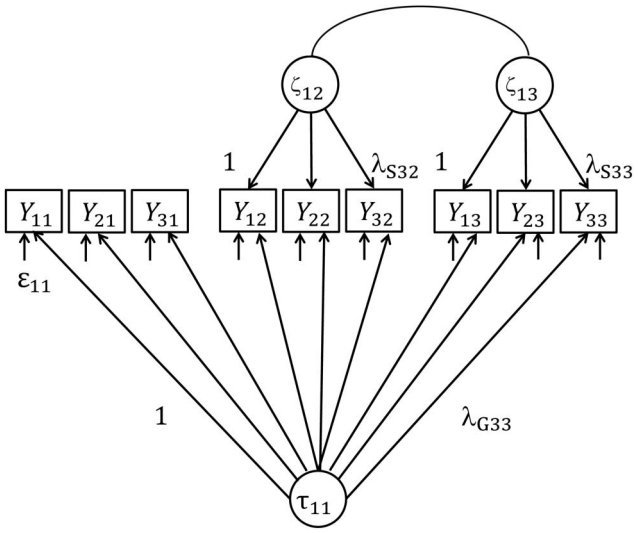

Figure 4. Bifactor-(S – 1) model with one general factor, two specific factors and three observed variables Y_{ik} per domain. τ_{ik}: true-score variables; ζ_{ik}: residual variables; ε_{ik}: error variables; λ_{Gik}: G-factor loadings, λ_{Sik}: specific factor loadings $k = 1, \ldots, K$; K: number of domains; $i = 1, \ldots, I_k$; I_k: number of indicators i belonging to domain k. For simplicity, not all parameters and variables are labeled.

Model With a Reference Indicator: The Bifactor-($S{\cdot}I$ – 1) Model

A less restrictive G-factor model that allows a higher degree of item heterogeneity can be defined by the following steps: We assume that the first domain ($k = 1$) serves as reference domain and that the first indicator of the reference domain (Y_{11}) serves as reference indicator. All nonreference true score variables τ_{ik} $(i, k) \neq (1, 1)$ are linearly regressed on the true score variable τ_{11} that pertains to the first indicator of the reference domain:

$$E(\tau_{ik} \mid \tau_{11}) = \alpha_{ik} + \lambda_{Gik}\tau_{11} \tag{12}$$

1. For each nonreference domain we select the first indicator (Y_{1k}, $k \neq 1$) as reference indicator and assume that all regression residuals $\zeta_{ik} = \tau_{ik} - E(\tau_{ik} \mid \tau_{11})$ belonging to the same domain k are linear functions of the regression residual of the first indicator:

$$\zeta_{ik} = \lambda_{Si1}\zeta_{1k} \tag{13}$$

2. For the reference domain we take the second indicator as a further reference indicator and assume that all regression residuals $\zeta_{i1} = \tau_{i1} - E(\tau_{i1} \mid \tau_{11})$ belonging to the reference domain ($k = 1$) are linear functions of the regression residual of the second indicator:

$$\zeta_{i1} = \lambda_{S21}\zeta_{21} \quad (i \neq 1) \tag{14}$$

3. We assume that all error variables $\varepsilon_{ik} = Y_{ik} - \tau_{ik}$ are uncorrelated.

We refer to this model as the bifactor-($S{\cdot}I$ – 1) model (with $S = K$, K: number of domains considered). An example of the bifactor-($S{\cdot}I$ – 1) model for three domains and three indicators per domain is shown in Figure 5. This model differs from the bifactor-(S – 1) model in one important way: There is now also a specific factor for the reference domain. However, the first indicator of the reference domain (the reference indicator for the G factor) is not allowed to have a loading on a specific factor. Hence, this model includes one specific factor loading less than the total number of indicators considered. If there is an equal number of items per domain there are in total $S{\cdot}I$ observed variables (indicators). Therefore, we call this model bifactor-($S{\cdot}I$ – 1) model.

The model has the same properties as the bifactor-(S – 1) model, and the factors have the same meaning. The only difference is that there is one specific factor more in the bifactor-($S{\cdot}I$ – 1) model. This model explains why in many applications of the conventional bifactor approach one loading is missing on at least one specific factor. According to the bifactor-($S{\cdot}I$ – 1) model, one measured variable serves as a marker or "gold standard" measure for G and therefore does not load onto a specific factor. The difference between the bifactor-($S{\cdot}I$ – 1) model and empirical applications with close-to-zero loadings is that the marker indicator is chosen a priori based on theory in the model, whereas it is data-driven in applications in which one or more indicators do not load significantly onto a specific factor. In the bifactor-($S{\cdot}I$ – 1) model consistency, specificity, and reliability coefficients can be calculated in the same way as in the bifactor-(S – 1) model.

Applications

We now illustrate the two new approaches with an analysis of nine items measuring the intensity of negative emotions with respect to three emotion groups: (a) anger (items: anger, fury, rage); (b) depression (items: depression, sorrow, unhappiness); and (c) guilt (items: guilt, shame, embarrassment). Participants were

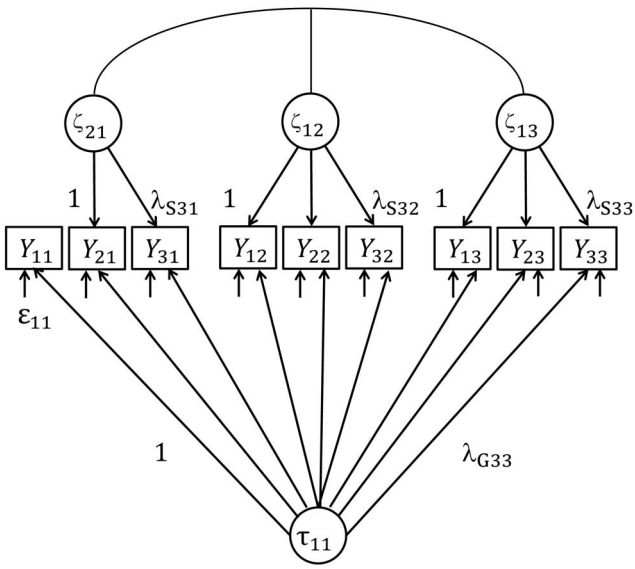

Figure 5. Bifactor-($S{\cdot}I$ – 1) model with one general factor, three specific factors and three observed variables Y_{ik} per domain. τ_{ik}: true-score variables; ζ_{ik}: residual variables; ε_{ik}: error variables; λ_{Gik}: G-factor loadings, λ_{Sik}: specific factor loadings $k = 1, \ldots, K$; K: number of domains; $i = 1, \ldots, I_k$; I_k: number of indicators i belonging to domain k. For simplicity, not all parameters and variables are labeled.

asked to assess the intensity with which they usually experience this emotion on a scale with four response categories (very weakly, rather weakly, rather strongly, very strongly). The items stem from a longer list of 28 emotions. For didactical reasons we selected only these nine emotions. The sample consisted of $N = 482$ individuals (Trierweiler, Eid, & Lischetzke, 2002).

Many emotion psychologists might be tempted to fit a bifactor model to the data. It could be hypothesized that there is a general disposition of emotional intensity that has an influence on all emotions in addition to specific anger, depression, and guilt factors. However, from a psychometric point of view a traditional bifactor model does not make sense in this application, because the three emotions are not randomly selected from a universe of interchangeable emotions. Instead, the emotions are clearly distinct and should be viewed as "fixed factors." As a consequence, the bifactor-$(S - 1)$ and -$(S \cdot I - 1)$ models should be more appropriate for these data.

Because the item responses were on a 4-point rating scale, we applied CFA estimation methods suitable for ordered categorical (ordinal) variables. We show in the appendix how the specific models presented in this article can be properly defined for ordinal response variables on the basis of SMT. For the present application, it is sufficient to know that in CFA models for ordinal response variables it is assumed that there is a continuous response variable Y^*_{ik} underlying each observed ordinal response variable Y_{ik}. The response variable Y^*_{ik} is decomposed into a linear combination of the general factor and specific factors in the same way as in the models for continuous observed variables. However, for identification reasons all intercepts have to be fixed to 0 (see Appendix for details). The WLSMV estimator for categorical observed variables was applied (Muthén & Muthén, 1998–2012). We used the THETA parameterization because this parameterization is in line with the formulation of the CFA model as a model of item response theory (see Appendix). All analyses were done using the computer program Mplus Version 7.3 (Muthén & Muthén, 1998–2012).

Multiple Correlated First-Order Factor Model

A multiple correlated first-order factor model with three domain factors as depicted in Figure 3 fit the data well $\chi^2(24, N = 482) = 44.03$, $p = .01$; RMSEA = 0.04; CFI = 0.99. The loading parameters as well as the variances and covariances of the factors are given in Table 6. All items had substantial loadings all of which were significantly different from zero ($p < .01$). The correlations between the latent factors were as follows: $\hat{\rho}$ (anger, depression) = .33, $\hat{\rho}$ (anger, guilt) = .30, $\hat{\rho}$ (guilt, depression) = .56. All latent correlations were significantly different from zero ($p < .01$). The correlations show that all factors were significantly correlated, but that the correlation between guilt and depression was stronger than the correlations between anger and the other emotions.

Traditional Bifactor Model

An application of the traditional bifactor model did not converge for these data. An in-depth analysis of the convergence problems showed that Mplus did not succeed in estimating the loading parameters of one indicator (shame) appropriately (the loading parameter estimate was getting increasingly large and the estimated variance of the corresponding specific factor was getting very low). This is in line with our new bifactor models derived from SMT, in which one indicator (or the entire set of indicators pertaining to a specific domain) have loadings only on *G*, but not on a specific factor.

Table 6
First-Order Factor Model: Estimated Loading Parameters as Well as Variances, Covariances, and Correlations (Standard Errors Are Given in Brackets)

Item	Factor loadings	Standardized factor loadings
Factor I		
Anger	1.000	0.595 (0.040)
Fury	2.631 (0.606)	0.890 (0.035)
Rage	1.820 (0.261)	0.803 (0.033)
Factor II		
Depression	1.000	0.710 (0.039)
Sorrow	0.929 (0.145)	0.684 (0.039)
Unhappiness	1.006 (0.159)	0.712 (0.035)
Factor III		
Guilt	1.000	0.549 (0.056)
Embarrassment	1.134 (0.245)	0.598 (0.056)
Shame	1.133 (0.234)	0.597 (0.051)

	Anger	Depression	Guilt
Anger	0.549 (0.114)	0.326 (0.056)	0.304 (0.064)
Depression	0.243 (0.055)	1.017 (0.225)	0.555 (0.056)
Guilt	0.148 (0.042)	0.368 (0.077)	0.432 (0.127)

Note. Variances are diagonal, covariances are subdiagonal, and correlations are superdiagonal. Standard errors are given in parentheses.

Bifactor-$(S - 1)$ Model

We now present an application of the bifactor-$(S - 1)$ model to our data example in which we used anger as a reference domain. Anger was chosen as a reference domain, because anger represents a more externalizing emotion, whereas depression and guilt are more internalizing emotions. The Mplus syntax for this model is presented in Appendix A.5.1. The model fit the data well, $\chi^2(20, N = 482) = 27.57$, $p = .12$; RMSEA = 0.03; CFI = 1.00. The better fit of this model compared to the multiple correlated first-order factor model is due to the fact that the bifactor-$(S - 1)$ model is better able to take item heterogeneity into account.

In the model, we fixed the loadings of the first indicators of each domain to 1. Hence, the reference indicator is the item anger. The *G* factor in this model represents anger intensity. The specific depression factor measures the deviations of the error-free depression scores from the values expected on the basis of the anger intensity variable. A positive score on the specific depression factor would indicate that a person experiences feelings of depression more intensively than one would expect given her or his anger intensity score. A negative value would represent a lower depression intensity than one would expect given the person's anger intensity score. The meaning of the specific guilt factor can be interpreted in the same way. The estimated factor loadings, variances, covariances, and the coefficients of consistency and specificity are presented in Table 7 and Table 8.

Table 7
Bifactor-(S – 1) Model With the Anger Domain as Reference Domain: Estimated Loading Parameters, Consistency and Specificity Coefficients as Well as Variances, Covariances, and Correlations of the Factors (Standard Errors are Given in Parentheses)

Item	*G*-factor loadings	Standardized *G*-factor loadings	*S*-factor loadings	Standardized *S*-factor loadings	Consistency	Specificity
Anger	1.000	.593 (.040)			1	0
Fury	2.671 (.616)	.891 (.034)			1	0
Rage	1.821 (.260)	.802 (.033)			1	0
Depression	.512 (.114)	.274 (.049)	1.000	.630 (.044)	.159	.841
Sorrow	.444 (.103)	.242 (.049)	.974 (.157)	.626 (.042)	.130	.870
Unhappiness	.360 (.119)	.173 (.053)	1.302 (.236)	.737 (.040)	.052	.948
Guilt	.370 (.095)	.235 (.053)	1.000	.449 (.057)	.215	.785
Embarrassment	.327 (.095)	.195 (.051)	1.302 (.292)	.551 (.061)	.111	.889
Shame	.187 (.099)	.100 (.052)	1.780 (.453)	.677 (.063)	.021	.979

These coefficients show that the anger factor could only explain a rather small amount of variance in the two other emotion domains (see Table 7). The highest consistency values were found for guilt and depression, the lowest for shame and unhappiness. The correlation of the two specific factors ($\hat{\rho}$ = .49; $p < .01$) shows that the guilt and depression domains had more in common than could be explained by the anger factor. This means that there was a rather strong tendency for people who experience guilt more (or less) intensely than would be expected based on their anger intensity scores to also experience depression more (or less) intensely than would be expected based on their anger intensity.

We also estimated the model with depression as the reference domain. The fit of this model version was also good, $\chi^2(20, N = 482) = 33.01$, $p = .03$; RMSEA = 0.04; CFI = 0.99. For guilt as reference facet the fit was: $\chi^2(20, N = 482) = 40.60$, $p < .01$; RMSEA = 0.05; CFI = 0.99. These differences in the fit coefficients are due to the fact that the multidimensional first-order model showed some misfit. This misfit was a result of the indicators of one domain not being perfectly homogeneous. Choosing a different reference domain represents the item heterogeneity in a different way.

Bifactor-(*S·I* – 1) Model

In applying the bifactor-(*S·I* – 1) to our data one has to take into consideration that each domain had three indicators in this application. As a consequence, the specific factor of the reference domain had only two indicators in this application. In this case, the model is only identified if the specific factor pertaining to the reference domain is correlated with at least one other specific factor (or external variable included in the model). When the correlations of this specific factor with other variables in the model are close to 0, this can cause empirical underidentification and estimation problems. Because this was the case in some versions of the present model, we fixed both loadings on the specific reference factor to 1 to ensure the identification of the model.

Table 8
Bifactor-(S-1) Model With the Anger Domain as Reference Domain: Variances (Diagonal), Covariances (Subdiagonal), and Correlations (Superdiagonal) Between the General Factor (G-Anger) and the Specific Factors (S-Depression, S-Guilt) (Standard Errors are Given in Parentheses)

	G-Anger	*S*-Depression	*S*-Guilt
G-Anger	.542 (.112)		
S-Depression		.751 (.177)	.490 (.060)
S-Guilt		.221 (.054)	.272 (.087)

The bifactor-(*S·I* – 1) model also fit our empirical example well. With the anger item used as reference indicator, the fit was: $\chi^2(17, N = 482) = 29.60$, $p = .03$; RMSEA = 0.04; CFI = 0.99. With the depression item used to define the specific reference domain factor, the fit was $\chi^2(16, N = 482) = 19.79$, $p = .23$; RMSEA = 0.02; CFI = 1.00. However, the variance of the specific factor of depression was estimated to be negative indicating that this represented an overfactorization and that the bifactor-(*S* – 1) model should be chosen for the anger reference domain, given that all versions of the bifactor-(*S* – 1) model resulted in proper solutions and showed an appropriate fit. Only when guilt was chosen as reference domain the bifactor-(*S·I* – 1) model showed a superior fit compared with the bifactor-(*S* – 1) model: $\chi^2(17, N = 482) = 26.28$, $p = .07$; RMSEA = 0.03; CFI = 0.99. The Mplus syntax for this model is presented in Appendix A.5.2. The superiority of the bifactor-(*S·I* – 1) model in this case also makes sense from a substantive point of view. The additional specific factor is a shame factor as the items shame and embarrassment have loadings on this factor and these two emotions differ from guilt. These applications show that the very general bifactor-(*S·I* – 1) model should only be applied in cases in which the additional specific factor makes sense from a substantive point of view.

The estimated coefficients are presented in Table 9 and Table 10. The consistency coefficients were relatively large for shame, embarrassment, sorrow and depression. This shows that the intensity of guilt is strongly related to these emotions and more weakly to anger. However, the specificity coefficients of shame and embarrassment were also substantial indicating that a specific factor for the reference domain was necessary. The correlations of the specific factors were small and not significantly different from 0 ($\hat{\rho}_{\text{depression-anger}} = .06$; $\hat{\rho}_{\text{depression-shame}} = -.06$; $\hat{\rho}_{\text{anger-shame}} = -.16$). Therefore, a model without correlations of specific factors also fitted the data very well, $\chi^2(20, N = 482) = 27.49$, $p = .12$; RMSEA = 0.03; CFI = 1.00. In this model, the standard errors of the *G*-factor loadings were smaller (between 0.14 and 0.44) showing that omitting nonsignificant correlations between the specific factors can improve the precision of estimating *G*-factor loadings.

Table 9
Bifactor-(S·I – 1) Model With the Item Guilt as Reference Indicator: Estimated Loading Parameters as Well as Variances, Covariances, and Correlations (Standard Errors are Given in Brackets)

Item	*G*-factor loadings	Standardized *G*-factor loadings	*S*-factor loadings	Standardized *S*-factor loadings	Consistency	Specificity
Guilt	1.000	.537 (.122)			1	0
Embarrassment	1.065 (.586)	.508 (.130)	1.000	.426 (.148)	.587	.413
Shame	1.008 (.562)	.488 (.130)	1.000	.432 (.148)	.561	.439
Anger	.502 (.289)	.258 (.086)	1.000	.530 (.054)	.192	.808
Fury	1.457 (.927)	.362 (.119)	3.315 (1.110)	.847 (.065)	.154	.846
Rage	.943 (.492)	.376 (.098)	1.665 (.221)	.683 (.062)	.233	.767
Depression	1.016 (.555)	.457 (.119)	1.000	.542 (.111)	.416	.584
Sorrow	1.040 (.490)	.498 (.094)	.749 (.166)	.432 (.110)	.571	.429
Unhappiness	1.015 (.619)	.427 (.139)	1.221 (.351)	.618 (.112)	.323	.677

Discussion

Bifactor and similarly structured CFA models are frequently used to study multidimensional data, and the use of such models has recently been advocated (Reise, 2012). At the same time, empirical applications of such models frequently produce anomalous results that can challenge the intended a priori interpretation of the factors and cause interpretation problems. 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. This sheds new light on applications of these models in psychology. 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. Based on SMT, it is clear that for single-level sampling designs, models with one specific factor less than domains considered—or models with a reference indicator that does not load onto a specific factor—are more appropriate than traditional bifactor approaches. Only these reduced model versions allow researchers to specify *G* and specific factors as functions of well-defined true score variables. Therefore, we recommend applying the two alternative models instead of the more traditional models that cannot be defined on the basis of SMT when the research design is characterized by a single-level sampling process. Selecting one reference domain or one reference indicator might lead to a model that is even less restricted, as correlations between the specific factors are allowed and have a clear meaning as partial correlations in these models.

Table 10
Bifactor-(S·I-1) Model With the Item Guilt as Reference Indicator: Variances (Diagonal), Covariances (Subdiagonal), and Correlations (Superdiagonal) Between the General Factor (G-Guilt) and the Specific Factors (S-Shame, S-Anger, S-Depression) (Standard Errors are Given in Brackets)

	G-Guilt	*S*-Shame	*S*-Anger	*S*-Depression
G-Anger	.406 (.258)			
S-Shame		.323 (.228)	−.155 (.322)	−.055 (.484)
S-Anger		−.058 (.100)	.430 (.109)	.056 (.184)
S-Depression		−.024 (.200)	.028 (.099)	.590 (.277)

There is an important difference between the traditional bifactor model and the bifactor models we have presented. In our models, different domains are contrasted against a reference domain. To understand this difference, consider an example from medicine. Let's assume that a physician has measured both height and weight with two measures each. The height and weight indicators are strongly correlated. A proponent of the classical bifactor approach might model a traditional *G* factor and two specific factors, one for weight and one for height. He might call the *G* factor "constitution." But what would the *G* factor "constitution" measure, and what would a *G*-factor score represent? In which way would the knowledge that a person has a large or a low constitution score help the physician to make a decision about a treatment? A proponent of the bifactor-(*S* – 1) model might choose the height domain as reference to define a *G* factor and model a specific factor for weight. The *G* factor has now a very clear meaning—it represents individuals' height corrected for measurement error. The specific weight factor also has a clear meaning. It indicates whether a person has a higher or lower weight than one would expect given the person's height. This knowledge would clearly help the physician to make a decision about a treatment. Indices used in medicine—such as the body mass index—are based on such ideas. They are not built on ideas of general "constitution" factors.

The application of the bifactor-(*S* – 1) model and the bifactor-(*S·I* – 1) model is especially reasonable when one of the domains considered is particularly outstanding and thus a clear candidate for a reference domain. In intelligence research, this could be a domain that is important for most or all other domains, for example, fluid intelligence. In satisfaction with life studies this could be, for example, the domain "satisfaction with the self." We have shown that the bifactor-(*S·I* – 1) model can be used if the indicators of the reference domain are not unidimensional.

The newly proposed bifactor models are one possibility to deal with modeling multidimensionality in the case of a single-level sampling process. There are, however, several alternatives:

1. The researcher could simply use a model with (psychometrically well-defined) correlated domain-specific first-order factors without integrating a *G* factor in the model (see Figure 3). Based on the first-order factors, the researcher can look at the individual profiles of scores. Profile analysis might be much more interesting for visualizing multidimensional data than just looking at a

single *G* factor (e.g., von Steinbüchel, Lischetzke, Gurny, & Eid, 2006).

2. The researcher can measure the *G* factor more directly. For example, in life satisfaction research one could use items that directly target *general* life satisfaction (i.e., by using a general life satisfaction scale). The common factor pertaining to the general life satisfaction items could be integrated as an independent variable in a model with the specific factors as dependent variables. The regression residuals of the dependent factors would then be interpreted as specific factors in a *G* factor model. Brunner et al. (2010) have shown how this idea can be applied to the assessment of academic self-concept. The self-assessed general academic self-concept factor is used as an independent variable in predicting the academic self-concept in different domains and domain-specific factors are defined as residual factors that are allowed to be correlated.

3. A *G* factor can be defined in a formative measurement model as a linear combination of the first-order domain-specific factors (Willoughby, Holochwost, Blanton, & Blair, 2014). Because the first-order domain-specific factors (see Figure 3) are well-defined even in the case of a single-sampling process also any linear combination of the first-order factors is well-defined. Principal component analysis is one way to obtain an optimal linear combination of first-order domains as a definition of a general factor. This idea of defining the *G*-factor as an emerging and not a latent property has gained increasing interest, for example, in modern theories of intelligence research. Conway and Kovacs (2015) give an up-to-date overview of new theoretical approaches of intelligence research that do not define a *G*-factor by a reflective measurement model—as is done in the traditional bifactor model—but by a formative measurement model. From the scope of stochastic measurement theory this seems to be a more appropriate definition if the research design is characterized by a single-level sampling process which is usually the case when different domains of intelligence are selected by theoretical assumptions.

4. Another possibility would be to take individual means across the first-order factor scores as a measure of *G* (aggregation approach). In this case a *G* factor also would have a clear meaning from the scope of view of stochastic measurement theory. This is usually done when test scores are defined by summing up item responses. If one takes the means, however, one has to make sure that this is reasonable. It requires, for example, that the factors are measured in the same metric. Taking the mean would also change the meaning of the *G* factor when adding or removing domains. Haberman (2007) as well as Reise, Bonifay, and Haviland (2013) discuss in detail under which conditions it is reasonable to consider general and specific scale scores when multidimensionality is present and give helpful advice for deciding whether specific subscale scores are necessary in addition to general total scores. An alternative to just summing up observed scores is to define the mean of the different first-order domain-specific factors as a latent variable in the model (Koch, Lochner, & Eid, in press; Pohl & Steyer, 2010).

We think that traditional *G*-factor models, although very popular, might not be reasonable for all kinds of multidimensional constructs. From a pure data analytical point of view a general higher-order factor can always be modeled when the observed variables are positively correlated (Krijnen, 2004), and there might be a strong temptation to represent such correlations by a general factor. From the scope of measurement theory, however, one has to give an answer to the question what such a factor means. What does an individual factor score tell us about the psychological state of an individual? If one cannot give a clear answer to this question, the assumption of a general factor might not be reasonable. In cases in which it is reasonable, however, a *G*-factor model can offer many interesting insights into the data structure. The alternative models proposed in this article might help researchers to find the most adequate model for their data.

Limitations

Similar to all other testable psychometric models the models presented in this article are defined based on homogeneity assumptions that can be wrong in specific applications. For example, the bifactor-($S-1$) model assumes that the indicators of the reference domain as well as the indicators of the specific-factors are unidimensional. This assumption can be violated in practical applications. In constructing items that should be analyzed with this model it is necessary to pay careful attention to developing unidimensional items. We have shown that the bifactor-($S{\cdot}I-1$) model can be used if the indicators of the reference domain are not unidimensional. However, this model is more prone to estimation problems if the indicators of the reference domain are strictly unidimensional. Therefore, it is necessary to think carefully about the covariance structures of the indicators one uses. Moreover, the models are not symmetrical and one has to select a reference domain and reference indicators. Also the fit of a model can change when different reference domains and indicators are selected. Geiser et al. (2008) shows how the CTC(M-1) model of multitrait-multimethod research can be restricted in such a way that the fit does not depend of the reference method. This approach can be used in an analogous way for the bifactor-($S-1$) model.

In the applications of the models a limited information estimation approach based on the polychoric correlations has been used. Because the specific factors in the new models can be correlated estimation methods for these models do not profit from dimension reduction (see Cai et al., 2011, for a deeper discussion). Therefore, further research is needed to figure out under which conditions other estimation methods such as full information maximum likelihood can be applied to these models.

References

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. http://dx.doi.org/10.1007/s10862-013-9354-5

Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattell-Horn-Carroll factor models: Differences between bifactor and higher order factor models in predicting language achievement. *Psychological Assessment, 26,* 789–805. http://dx.doi.org/10.1037/a0036745

Betts, K. S., Williams, G. M., Najman, J. M., Scott, J., & Alati, R. (2014). Exposure to stressful life events during pregnancy predicts psychotic experiences via behaviour problems in childhood. *Journal of Psychiatric Research, 59,* 132–139. http://dx.doi.org/10.1016/j.jpsychires.2014.08.001

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,* 214–222. http://dx.doi.org/10.1002/da.22139

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley. http://dx.doi.org/10.1002/9781118619179

Booth, T., Bastin, M. E., Penke, L., Maniega, S. M., Murray, C., Royle, N. A., . . . Deary, I. J. (2013). Brain white matter tract integrity and cognitive abilities in community-dwelling older people: The Lothian Birth Cohort, 1936. *Neuropsychology, 27,* 595–607. http://dx.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,* 136–145. http://dx.doi.org/10.1037/a0029228

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the bifactor model to assess the dimensionality of the Hong Psychological Reactance Scale. *Educational and Psychological Measurement, 71,* 170–185. http://dx.doi.org/10.1177/0013164410387378

Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A., & Martin, R. (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. *Journal of Educational Psychology, 102,* 964–981. http://dx.doi.org/10.1037/a0019644

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80,* 796–846. http://dx.doi.org/10.1111/j.1467-6494.2011.00749.x

Burke, J. D., Boylan, K., Rowe, R., Duku, E., Stepp, S. D., Hipwell, A. E., & Waldman, I. D. (2014). Identifying the irritability dimension of ODD: Application of a modified bifactor model across five large community samples of children. *Journal of Abnormal Psychology, 123,* 841–851. http://dx.doi.org/10.1037/a0037898

Burns, A., Höfer, S., Curry, P., Sexton, E., & Doyle, F. (2014). Revisiting the dimensionality of the Hospital Anxiety and Depression Scale in an international sample of patients with ischaemic heart disease. *Journal of Psychosomatic Research, 77,* 116–121. http://dx.doi.org/10.1016/j.jpsychores.2014.05.005

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, and Allied Disciplines, 55,* 393–401. http://dx.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,* 20–34. http://dx.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,* 177–199. http://dx.doi.org/10.1177/0265532212456833

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581–612. http://dx.doi.org/10.1007/s11336-010-9178-0

Cai, L. (2015). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika, 80,* 535–559. http://dx.doi.org/10.1007/s11336-014-9411-3

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16,* 221–248. http://dx.doi.org/10.1037/a0023350

Cham, H., Hughes, J. N., West, S. G., & Im, M. H. (2014). Assessment of adolescents' motivation for educational attainment. *Psychological Assessment, 26,* 642–659. http://dx.doi.org/10.1037/a0036213

Chen, F. F., Jing, Y., Hayes, A., & Lee, J. M. (2013). Two concepts or two approaches? A bifactor analysis of psychological and subjective well-being. *Journal of Happiness Studies, 14,* 1033–1068. http://dx.doi.org/10.1007/s10902-012-9367-x

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,* 189–225. http://dx.doi.org/10.1207/s15327906mbr4102_5

Cho, S.-J., Cohen, A., & Kim, S. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling, 21,* 375–395. http://dx.doi.org/10.1080/10705511.2014.915371

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver interaction scale: Results from the early childhood longitudinal study-birth cohort. *Early Childhood Research Quarterly, 28,* 218–233. http://dx.doi.org/10.1016/j.ecresq.2012.12.004

Conway, A. R. A, & Kovacs, K. (2015). New and emerging models of human intelligence. *Wiley Interdisciplinary Reviews: Cognitive Science, 6,* 419–426.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Davidson, M. M., Gervais, S. J., Canivez, G. L., & Cole, B. P. (2013). A psychometric examination of the Interpersonal Sexual Objectification Scale among college men. *Journal of Counseling Psychology, 60,* 239–250. http://dx.doi.org/10.1037/a0032075

Deemer, D. E., Smith, J. L., Thoman, D. B., & Chase, J. P. (2014). Precision in career motivation assessment: Testing the subjective science attitude change measures. *Journal of Career Assessment, 22,* 489–504. http://dx.doi.org/10.1177/1069072713498683

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43,* 145–168. http://dx.doi.org/10.1111/j.1745-3984.2006.00010.x

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

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,* 966–970. http://dx.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,* 467–476. http://dx.doi.org/10.1037/a0031393

Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research—Online, 1,* 65–85.

Eid, M., & Diener, E. (2004). Global judgments of subjective well-being: Situational variability and long-term stability. *Social Indicators Research, 65,* 245–277. http://dx.doi.org/10.1023/B:SOCI.0000003801.89195.bc

Eid, M., & Koch, T. (2014). The meaning of higher-order factors in reflective measurement models. *Measurement, 12,* 96–101.

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,* 38–60. http://dx.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,* 230–253. http://dx.doi.org/10.1037/a0013219

Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion* [Test theory and test construction]. Göttingen, Germany: Hogrefe.

Eid, M., Schneider, C., & Schwenkmezger, P. (1999). Do you feel better or worse? On the validity of perceived deviations of mood states from mood traits. *European Journal of Personality, 13,* 283–306. http://dx.doi.org/10.1002/(SICI)1099-0984(199907/08)13:4<283::AID-PER341>3.0.CO;2-0

Garin, O., Ferrer, M., Pont, À., Wiklund, I., Van Ganse, E., Vilagut, G., . . . Alonso, J. (2013). Evidence on the global measurement model of the Minnesota Living with Heart Failure Questionnaire. *Quality of Life Research, 22,* 2675–2684. http://dx.doi.org/10.1007/s11136-013-0383-z

Gavett, B. E., Crane, P. K., & Dams-O'Connor, K. (2013). Bi-factor analyses of the brief test of adult cognition by telephone. *NeuroRehabilitation, 32,* 253–265.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(M-1) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods, 13,* 49–57. http://dx.doi.org/10.1037/1082-989X.13.1.49

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement, 31,* 4–19. http://dx.doi.org/10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika, 57,* 423–436. http://dx.doi.org/10.1007/BF02295430

Gomez, R. (2013). Depression anxiety stress scales: Factor structure and differential item functioning across women and men. *Personality and Individual Differences, 54,* 687–691. http://dx.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 Rothbarth's temperament dimensions. *Personality and Individual Differences, 60,* 73–78. http://dx.doi.org/10.1016/j.paid.2013.12.023

Grygiel, P., Humenny, G., Rebisz, S., Switaj, P., & Sikorska, J. (2013). Validating the Polish adaptation on the 11-item De Jong Gierveld loneliness scale. *European Journal of Psychological Assessment, 29,* 129–139. http://dx.doi.org/10.1027/1015-5759/a000130

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28,* 407–434. http://dx.doi.org/10.1207/s15327906mbr2804_2

Haberman, S. J. (2007). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33,* 204–229. http://dx.doi.org/10.3102/1076998607302636

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,* 1257–1274. http://dx.doi.org/10.1111/cdev.12184

Han, K. T., & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement, 38,* 486–498. http://dx.doi.org/10.1177/0146621614536770

Hays, W. L. (1994). *Statistics* (5th ed.). New York, NY: Harcourt College Publishers.

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

Hunter, J. E. (1968). Probabilistic foundations for coefficients of generalizability. *Psychometrika, 33,* 1–18. http://dx.doi.org/10.1007/BF02289672

Hyland, P., Shevlin, M., Adamson, G., & Boduszek, D. (2013). The factor structure and composite reliability of the profile of emotional distress. *Cognitive Behaviour Therapy, 6,* 1–2.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38,* 32–60. http://dx.doi.org/10.3102/1076998611432173

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. *Applied Psychological Measurement, 38,* 404–405. http://dx.doi.org/10.1177/0146621614524982

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,* 81–90.

Kóbor, A., Takács, A., & Urbán, R. (2013). The bifactor model of the strengths and difficulties questionnaire. *European Journal of Psychological Assessment, 29,* 299–307. http://dx.doi.org/10.1027/1015-5759/a000160

Koch, T., Lochner, K., & Eid, M. (in press). Multitrait-multimethod-analysis: The psychometric foundation of CFA-MTMM models. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley handbook of psychometric testing*. London, UK: Wiley.

Koch, T., Schultze, M., Jeon, M., Nussbeck, F. W., Praetorius, A.-K., & Eid, M. (2016). A cross-classified CFA-MTMM model for structurally different and non-independent interchangeable methods. *Multivariate Behavioral Research, 51,* 67–85. http://dx.doi.org/10.1080/00273171.2015.1101367

Krijnen, W. P. (2004). Positive loadings and factor correlations from positive covariance matrices. *Psychometrika, 69,* 655–660. http://dx.doi.org/10.1007/BF02289861

Levant, R. F., Hall, R. J., & Rankin, T. J. (2013). Male Role Norms Inventory-Short Form (MRNI-SF): Development, confirmatory factor analytic investigation of structure, and measurement invariance across gender. *Journal of Consulting Psychology, 60,* 228–238. http://dx.doi.org/10.1037/a0031545

Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *The British Journal of Mathematical and Statistical Psychology, 67,* 496–513. http://dx.doi.org/10.1111/bmsp.12030

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,* 67–87. http://dx.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,* 395–406. http://dx.doi.org/10.1037/a0035284

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307. http://dx.doi.org/10.1037/0033-2909.115.2.300

Mészáros, V., Adám, S., 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 and Health, 30,* 82–88. http://dx.doi.org/10.1002/smi.2481

Meyer, J. F., & Brown, T. A. (2013). Psychometric evaluation of the thought-action fusion scale in a large clinical sample. *Assessment, 20,* 764–775. http://dx.doi.org/10.1177/1073191112436670

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered- categorical measures. *Multivariate Behavioral Research, 39,* 479–515. http://dx.doi.org/10.1207/S15327906MBR3903_4

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling, 4,* 193–211. http://dx.doi.org/10.1080/10705519709540071

Muntinga, M. E., Mokkink, L. B., Knol, D. L., Nijpels, G., & Jansen, A. P. D. (2014). Measurement properties of the Client-centered Care Questionnaire (CCCQ): Factor structure, reliability and validity of a questionnaire to assess self-reported client-centeredness of home care services in a population of frail, older people. *Quality of Life Research, 23,* 2063–2072. http://dx.doi.org/10.1007/s11136-014-0650-7

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115–132. http://dx.doi.org/10.1007/BF02294210

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Norwalk, K. E., DiPerna, J. C., & Lei, P. W. (2014). Confirmatory factor analysis of the early arithmetic, reading, and learning indicators. *Journal of School Psychology, 52,* 83–96. http://dx.doi.org/10.1016/j.jsp.2013.11.006

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

Nussbeck, F. W., Eid, M., Geiser, C., Courvoisier, D. S., & Lischetzke, T. (2009). A CTC(M–1) model for different types of raters. *Methodology, 5,* 88–98. http://dx.doi.org/10.1027/1614-2241.5.3.88

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analyzing MTMM data with SEM for ordinal variables applying the WLSMV-estimator: What is the sample size needed for valid results? *The British Journal of Mathematical and Statistical Psychology, 59,* 195–213. http://dx.doi.org/10.1348/000711005X67490

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,* 469–481. http://dx.doi.org/10.1016/j.beth.2014.02.006

Park, M.-H., Dimitrov, D. M., Das, A., & Gichuru, M. (2014). The teacher efficacy for inclusive practices (TEIP) scale: Dimensionality and factor structure. *Journal of Research in Special Educational Needs*. Advance online publication. http://dx.doi.org/10.1111/1471-3802.12047

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research, 45,* 45–72. http://dx.doi.org/10.1080/00273170903504729

Primi, R., Rocha da Silva, M. C., Rodrigues, P., Muniz, M., & Almeida, L. S. (2013). The use of the bi-factor model to test the unidimensionality of a battery of reasoning tests. *Psicothema, 25,* 115–122.

Reise, S. P. (2012). The rediscovery of the bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. http://dx.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,* 129–140. http://dx.doi.org/10.1080/00223891.2012.725437

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92,* 544–559. http://dx.doi.org/10.1080/00223891.2010.496477

Reise, S. P., Moore, T. M., & Maydeu-Olivares, A. (2011). Targeted bifactor rotations and assessing the impact of model violations on the parameters of unidimensional and bifactor models. *Educational and Psychological Measurement, 71,* 684–711. http://dx.doi.org/10.1177/0013164410378690

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,* 361–372.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Richmond, VA: Psychometric Society.

Silvia, P. J., Beaty, R., & Nusbaum, E. (2013). Verbal fluency and creativity: General and specific contributions of broad retrieval ability (Gr) factors to divergent thinking. *Intelligence, 41,* 328–340. http://dx.doi.org/10.1016/j.intell.2013.05.004

Spearman, C. (1904). "General intelligence" objectively determined and measured. *The American Journal of Psychology, 15,* 201–293. http://dx.doi.org/10.2307/1412107

Steyer, R. (1988). Conditional expectations: An introduction to the concept and its applications in empirical sciences. *Methodika, 2,* 53–78.

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

Steyer, R., & Eid, M. (2001). *Messen und Testen* [Measurement and testing]. Berlin, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-56924-1

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8,* 79–98.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits—Revised. *Annual Review of Clinical Psychology, 11,* 71–98. http://dx.doi.org/10.1146/annurev-clinpsy-032813-153719

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13,* 389–408. http://dx.doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Tackett, J. L., Daoud, S. L. S. B., De Bolle, M., & Burt, S. A. (2013). Is relational aggression part of the externalizing spectrum? A bifactor model of youth antisocial behavior. *Aggressive Behavior, 39,* 149–159. http://dx.doi.org/10.1002/ab.21466

Tackett, J. L., Lahey, B. B., van Hulle, C., Waldman, I., Krueger, R. F., & Rathouz, P. J. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *Journal of Abnormal Psychology, 122,* 1142–1153. http://dx.doi.org/10.1037/a0034151

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408. http://dx.doi.org/10.1007/BF02294363

Trierweiler, L. I., Eid, M., & Lischetzke, T. (2002). The structure of emotional expressivity: Each emotion counts. *Journal of Personality and Social Psychology, 82,* 1023–1040. http://dx.doi.org/10.1037/0022-3514.82.6.1023

Tripp, M. K., Diamond, P. M., Vernon, S. W., Swank, P. R., Dolan Mullen, P., & 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,* 828–842. http://dx.doi.org/10.1093/her/cys114

Urbán, R., Kun, B., Farkas, J., Paksi, B., Kökönyei, G., Unoka, Z., . . . Demetrovics, Z. (2014). Bifactor structural model of symptom checklists: SCL-90-R and Brief Symptom Inventory (BSI) in a non-clinical community sample. *Psychiatry Research, 216,* 146–154. http://dx.doi.org/10.1016/j.psychres.2014.01.027

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: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 22,* 2581–2594. http://dx.doi.org/10.1007/s11136-013-0370-4

von Steinbüchel, N., Lischetzke, T., Gurny, M., & Eid, M. (2006). Assessing quality of life in older people: Psychometric properties of the WHOQOL-BREF. *European Journal of Ageing, 3,* 116–122. http://dx.doi.org/10.1007/s10433-006-0024-2

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511618765

Watkins, M. W., & Beaujean, A. A. (2014). Bifactor structure of the Wechsler preschool and primary scale of intelligence—4th ed. *School Psychology Quarterly, 29,* 52–63. http://dx.doi.org/10.1037/spq0000038

Watters, C. A., Keefer, K. V., Kloosterman, P. H., Summerfeldt, L. J., & Parker, J. D. A. (2013). Examining the structure of the internet addiction test in adolescents: A bifactor approach. *Computers in Human Behavior, 29,* 2294–2302. http://dx.doi.org/10.1016/j.chb.2013.05.020

Weise, C., Kleinstäuber, M., Hesser, H., Westin, V. Z., & Andersson, G. (2013). Acceptance of tinnitus: Validation of the tinnitus acceptance

questionnaire. *Cognitive Behaviour Therapy, 42,* 100–115. http://dx.doi.org/10.1080/16506073.2013.781670

Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement, 34,* 25–33. http://dx.doi.org/10.1177/001316447403400104

Willoughby, M., Holochwost, S. J., Blanton, Z. E., & Blair, C. B. (2014). Executive functions: Formative versus reflective measurement. *Measurement: Interdisciplinary Research and Perspectives, 12,* 69–95. http://dx.doi.org/10.1080/15366367.2014.929453

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,* 172–183. http://dx.doi.org/10.1007/s12529-012-9237-2

Yang, Y., Sun, Y., Zhang, Y., Jiang, Y., Tang, J., Zhu, X., & Miao, D. (2013). Bifactor item response theory model of acute stress response. *PLOS ONE, 8,* 1–10. http://dx.doi.org/10.1371/journal.pone.0065291

Yap, S. C. Y., Donnellan, M. B., Schwartz, S. J., Kim, S. Y., Castillo, L. G., Zamboanga, B. L., . . . 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,* 437–446. http://dx.doi.org/10.1037/a0036253

Young, M. A., Hutman, P., Enggasser, J. L., & Meesters, Y. (2014). Assessing usual seasonal depression symptoms: The seasonality assessment form. *Journal of Psychopathological and Behavioral Assessment, 37,* 112–121.

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,* 113–128. http://dx.doi.org/10.1007/BF02294531

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

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40,* 395–412.

Zimmerman, D. W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement, 36,* 85–96. http://dx.doi.org/10.1177/001316447603600107

Appendix

Bifactor Models for Ordinal Observed Variables

Definition of Latent Variables Based on Ordinal Observed Variables

In this appendix we show how the latent variables in the factor models that we presented in the text for continuous observed variables can be defined for ordinal response variables. We refer to the graded response model (Samejima, 1969), because it is equivalent to the model of confirmatory factor analysis of categorical response variables with ordered response categories (Takane & deLeeuw, 1987). The basic idea of defining latent variables can easily be transferred to other IRT models (such as logistic test models).

A.1 Single-Level Sampling Process: Unidimensional and Multidimensional First-Order Factor Models

The starting point in the graded response model are the cumulative response probability variables $P(Y_{ik} \geq c \mid p_U)$. A value $P(Y_{ik} \geq c \mid p_U = u)$ of such a cumulative response probability variable is the probability that an individual u gives an answer at least in category c of the ordinal observed response variable Y_{ik}, $i = 1, \ldots, I_k$; I_k: number of indicators i belonging to domain k; $k = 1, \ldots, K$; K: number of domains. For simplicity, we assume that all observed variables have the same number of response categories c with $c = 0, \ldots, C$. The response probability variables are then defined for all $c > 0$. That means that there are $C - 1$ response probability variables. In order to make a linear decomposition—as is assumed in a factor analytic model—the cumulative response probability variables $P(Y_{ik} \geq c \mid p_U)$ are transformed. The graded response model uses the probit transformation. The probit variable π_{ick} is defined as $\pi_{ick} = \Phi^{-1}[P(Y_{ik} \geq c \mid p_U)]$ where Φ^{-1} is the inverse of the cumulative distribution function of the standard normal distribution. It is then assumed that all probit variables belonging to the same observed variable (item) Y_{ik} are translations of each other: $\pi_{ick} = \kappa_{icc'k} + \pi_{ic'k}$. A common item-specific latent variable π_{ik} can be defined as a translation of an arbitrary probit variable. If one defines, for example, $\pi_{ik} = \pi_{i1k}$, one obtains the measurement model $\pi_{ick} = \kappa_{ick} + \pi_{ik}$ where κ_{ick} is a threshold parameter and $\kappa_{i1k} = 0$. The item-specific latent variable π_{ik} is the probit variable belonging to the second response category ($c = 1$) and the response probability variable $P(Y_{ik} \geq 1 \mid p_U)$. In computer programs for structural equation modeling the probit variable is not defined by fixing one threshold parameter per item but by fixing the expected value of a probit variable to 0 (centered probit variables). In order to define an item response model with group-specific first-order factors it is assumed that the probit variables belonging to the same domain are linear functions of each other: $\pi_{ik} = \alpha_{ijk} + \lambda_{ijk}\pi_{jk}$. In the case of defining the item-specific probit variables by fixing their expected values to 0, this equation reduces to $\pi_{ik} = \lambda_{ijk}\pi_{jk}$. Now, a common factor can be defined as a linear function of an arbitrary probit variable. If one defines $\pi_k = \pi_{1k}$ one obtains the measurement equation $\pi_{ik} = \lambda_{ik}\pi_k$ (in the case of centered probit variables) with $\lambda_{ik} = 1$. The common domain-specific factor is then the probit variable of the first item belonging to this group of items and has a clear meaning.

(*Appendix continues*)

The probit model (graded response model) is equivalent to a model of CFA for ordinal response models (CFA-OR) under the following conditions:

1. In the CFA-OR model it is assumed that there is a latent response variable Y_{ik}^* for each observed variable Y_{ik}. Both variables Y_{ik} and Y_{ik}^* are linked by the following threshold relationship (Eid, 1996; Millsap & Yun-Tein, 2004; Muthén, 1984):

$$Y_{ik} = 0, \text{ if } Y_{ik}^* \leq \kappa_{i1k},$$

$$Y_{ik} = c, \text{ if } \kappa_{ick} < Y_{ik}^* \leq \kappa_{i(c+1)k}, \text{ for } 0 < c < C, \text{ and}$$

$$Y_{ik} = C, \text{ if } \kappa_{i(C)k} < Y_{ik}^*.$$

The threshold parameters κ_{ick} split the continuous variable Y_{ik}^* into C categories.

2. Each Y_{ik}^* variable can be decomposed in the probit variable π_{ik} and an error variable ε_{ik}:

$$Y_{ik}^* = \pi_{ik} + \varepsilon_{ik}$$

with $Var(\varepsilon_{ik}) = 1$.

A.2 Two-Level Sampling Process: Bifactor Model

Eid (1996) has shown that graded response latent state-trait models for ordinal observed variables can be defined by the following decomposition of the item-specific probit variables π_{ik}:

$$\pi_{ik} = \lambda_{Gik}\xi + \lambda_{Sik}\zeta_k$$

The item-specific probit variables π_{ik} are functions of the probit variables $\pi_{ick} = \Phi^{-1}[P(Y_{ik} \geq c \mid p_U, p_{Dk})]$.

A.3 Model With a Reference Domain: The Bifactor-($S - 1$) Model

Without loss of generality, we choose the domain $k = 1$ as reference domain and take the first indicator of this domain ($i = 1$) as reference indicator. The starting point is the probit variable π_{11}. If we consider continuous observed variables, the starting point is the true score variable τ_{11}. A *G*-factor model can be defined by the following steps:

1. We define all item-specific probit variables by fixing their expected value to 0, i.e., $E(\pi_{ik}) = 0$.

2. We assume that all probit variables π_{i1} belonging to the reference domain are linear functions of the reference probit variable π_{11}: $\pi_{i1} = \lambda_{Gi1}\pi_{11}$.

3. We assume that the regressions (conditional expectations) of all probit variables π_{ik} belonging to a nonreference domain ($k \neq 1$) on the reference probit variable π_{11} are linear: $E(\pi_{ik} \mid \pi_{11}) = \lambda_{Gik}\pi_{11}$.

4. For each domain we select the first indicator as reference indicator and assume that the all regression residuals $\zeta_{ik} = \pi_{ik} - E(\pi_{ik} \mid \pi_{11})$ belonging to the same domain k are linear functions of the regression residual of the first indicator: $\zeta_{ik} = \lambda_{Si1}\zeta_{1k}$.

5. We assume that all observed responses are stochastically independent from each other given the latent variables of the model.

A.4 Model With a Reference Indicator: The Bifactor-($S{\cdot}I - 1$) Model

A bifactor-($S{\cdot}I - 1$) model can be defined by the following steps:

1. We define all item-specific probit variables by fixing their expected value to 0, i.e., $E(\pi_{ik}) = 0$.

2. We assume that the regressions (conditional expectations) of all probit variables π_{ik} $(i, k) \neq (1, 1)$ being not equal with the reference probit variable on the reference probit variable π_{11} are linear: $E(\pi_{ik} \mid \pi_{11}) = \lambda_{Gik}\pi_{11}$.

3. For each domain that is not equal to the reference domain we select the first indicator as reference indicator and assume that all regression residuals $\zeta_{ik} = \pi_{ik} - E(\pi_{ik} \mid \pi_{11})$ belonging to the same domain k are linear functions of the regression residual of the fist indicator: $\zeta_{ik} = \lambda_{Si1}\zeta_{1k}$.

4. For the reference domain we take the second indicator as a further reference indicator and assume that the all regression residuals $\zeta_{i1} = \pi_{i1} - E(\pi_{i1} \mid \pi_{11})$ belonging to the reference domain ($k = 1$) are linear functions of the regression residual of the second indicator: $\zeta_{i1} = \lambda_{S21}\zeta_{11}$.

5. We assume that all observed responses are stochastically independent from each other given the latent variables of the model.

(*Appendix continues*)

A.5 Mplus of the Models Applied

A.5.1. Bifactor-(S – 1) Model.

```
TITLE: Bi-factor (S-1) model with anger as reference domain
DATA: FILE IS "bifactor.dat";
VARIABLE:
NAMES ARE
anger fury rage
sorrow depress unhapp
guilt embarras shame;
USEVARIABLES =
anger fury rage
sorrow depress unhapp
guilt embarras shame;
categorical=
anger fury rage
sorrow depress unhapp
guilt embarras shame;
MISSING ARE ALL (9);
ANALYSIS:
PARAMETERIZATION = THETA;
MODEL:
! Definition of the G factor
G by
anger fury rage depress
sorrow unhapp guilt
embarras shame;
! Definition of specific factors
S_Dep by depress sorrow unhapp;
S_Shame by guilt embarras shame;
! Uncorrelatedness of G factor with specific factors
G with S_Dep@0 S_Shame@0;
OUTPUT: Standardized;
```

A.5.2. Bifactor-(S·I – 1) Model.

```
TITLE: Bi-factor (S-1) model with anger as reference domain
DATA: FILE IS "bifactor.dat";
VARIABLE:
NAMES ARE
anger fury rage
sorrow depress unhapp
guilt embarras shame;
anger fury rage
sorrow depress unhapp
guilt embarras shame;
categorical=
anger fury rage
sorrow depress unhapp
guilt embarras shame;
MISSING ARE ALL (9);
ANALYSIS:
PARAMETERIZATION = THETA;
MODEL:
! Definition of the G factor
G by
guilt embarras shame
anger fury rage
depress sorrow unhapp;
! Definition of specific factors
S_Ang by anger fury rage;
S_Dep by depress sorrow unhapp;
S_Shame by embarras shame@1;
! Uncorrelatedness of G factor with specific factors
G with S_Ang@0 S_Dep@0 S_Shame@0;
OUTPUT: Standardized;
```

Received March 26, 2015
Revision received February 2, 2016
Accepted February 4, 2016 ■