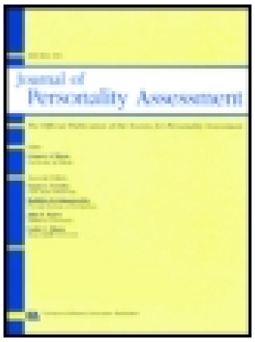

Routledge
Taylor & Francis Group

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: http://www.tandfonline.com/loi/hjpa20

A Profile-Based Framework for Factorial Similarity and the Congruence Coefficient

Anselma G. Hartley & R. Michael Furr

JOURNAL OF PERSONALITY ASSESSMENT
http://dx.doi.org/10.1080/00223891.2017.1279167

STATISTICAL DEVELOPMENTS AND APPLICATIONS

A Profile-Based Framework for Factorial Similarity and the Congruence Coefficient

Anselma G. Hartley and R. Michael Furr

Department of Psychology, Wake Forest University

ABSTRACT

We present a novel profile-based framework for understanding factorial similarity in the context of exploratory factor analysis in general, and for understanding the congruence coefficient (a commonly used index of factor similarity) specifically. First, we introduce the profile-based framework articulating factorial similarity in terms of 3 intuitive components: general saturation similarity, differential saturation similarity, and configural similarity. We then articulate the congruence coefficient in terms of these components, along with 2 additional profile-based components, and we explain how these components resolve ambiguities that can be—and are—found when using the congruence coefficient. Finally, we present secondary analyses revealing that profile-based components of factorial are indeed linked to experts' actual evaluations of factorial similarity. Overall, the profile-based approach we present offers new insights into the ways in which researchers can examine factor similarity and holds the potential to enhance researchers' ability to understand the congruence coefficient.

ARTICLE HISTORY
Received 8 June 2016
Revised 28 November 2016

Across many areas of psychology, including personality assessment, the factorial structure of sets of variables has important implications. For example, when developing and evaluating assessment instruments, researchers must pay careful attention to the factorial structure of the instruments' items. An important issue when examining an instrument's factorial structure is whether that structure is similar across groups of respondents. If, for example, an instrument's factorial structure is not similar across males and females, then the instrument is not "working" in the same way across sexes. It might reflect different psychological constructs in the two sexes, or it might reflect the same construct to differing degrees in the two sexes. Indeed, the instrument's scores might not be psychologically comparable across sexes. Thus, factorial similarity (or measurement invariance or differential item functioning) is a crucial psychometric issue that should be examined when an instrument is to be used across groups of respondents. The purpose of this article is to present a novel profile-based framework for understanding factorial similarity in the context of exploratory factor analysis (EFA).[1]

Researchers have long examined profiles of scores, and ways of characterizing profiles are well understood (Baker & Block, 1957; Cattell, 1949; Cronbach & Gleser, 1953; Furr, 2008, 2010). The profile-based framework we put forth in this article views factor similarity in terms of several components of profiles and profile similarity. In this approach, a set of factor loadings is seen as a profile of values, and it provides a way of understanding both a given set of loadings and the similarity between two sets of loadings. As summarized in Table 1, a given set of factor loadings—and thus the similarity between two sets of factor loadings—can be characterized in terms of several key profile-based components. In this article, we articulate each component, describe its meaning in the context of factor analysis and of factorial similarity, and examine the degree to which experts' actual judgments of factorial similarity are shaped by each component.[2]

We believe that the profile-based framework is important for at least two reasons. First and most broadly, it could reveal new ways of conceptualizing factor similarity. The different components of factorial similarity might have importantly different psychological and psychometric implications, each revealing unique and crucial information about the factor solutions. Ultimately, this framework might shape the questions that researchers ask in their factor analytic work.

Second, it resolves confusion that can—and does—arise with the most common index of factorial similarity in an EFA context, the congruence coefficient (φ; Burt, 1948; Tucker, 1951). Although φ's calculation is straightforward (as described later), its precise meaning might be less clear. For example, φ has been explained in terms of "the locations of the factor axes" (Jensen, 1998, p. 100), "a standardized measure of proportionality of elements in both vectors [i.e., both sets of factor loadings]" (Lorenzo-Seva & ten Berge, 2006, p. 57), an "unadjusted correlation" between two sets of factor loadings (Harman, 1976, p. 343), and "the cosine of the angle of the two factor loading vectors, taken from the origin" (Revelle, n.d., p. 174). Although such descriptions might be intuitive to some

CONTACT Anselma G. Hartley aghartley@gmail.com Amazon.com, Seattle, WA 98102.
Anselma G. Hartley is now at Amazon.com in Seattle, WA.

Supplemental material for this article can be accessed on the publisher's website.

[1]Factorial similarity is also frequently examined in the context of confirmatory factor analysis (CFA), and much attention has been given to those procedures in the form of multigroup CFA (e.g., Millsap & Meredith, 2007). Given the frequent use of EFA in many domains of psychological research, our current focus is on procedures mostly likely to be used by researchers conducting EFA, rather than CFA.

[2]These components are sometimes referred to as "elevation," "scatter," and "shape" in profile-based approaches (see Furr, 2010).

Table 1. Profile-based components of factor loading sets and factor similarity.

Component	Index	Description (The degree to which …)
Components of a set of factor loadings (set X or set Y)		
General saturation	$\overline{X}$, $\overline{Y}$	A set of items is generally affected by, or "saturated" with the latent variable; indexed by the mean factor loading in the set
Differential saturation	s_X, s_Y	The items within a set are differentially affected, or saturated with the latent variable; indexed by the standard deviation or variance of loadings
Configuration		The particular patterning with which items are differentially affected by the latent variable
Components of factorial similarity		
General saturation similarity	$\lvert\overline{X}-\overline{Y}\rvert$	Two sets of items are generally affected by a latent variable to the same level; indexed as the absolute value of the difference between means of each set of factor loadings
Differential saturation similarity	$\lvert s_X - s_y\rvert$	The differential saturation in one set of items matches the differential saturation in the other set; indexed as the absolute value of the difference between standard deviations of each set of factor loadings
Configural similarity	r_{XY}	The patterning of loadings is similar across two sets of factor loadings; indexed by the correlation between the two sets of factor loadings
Additional components related to the congruence coefficient		
Joint general saturation	$\overline{XY}$	The items are strongly affected by the latent variable in general, across both groups; indexed by the geometric mean of the (squared) mean factor loading of each set (i.e., $\overline{XY}=\sqrt{\overline{X}^2\overline{Y}^2}$).
Joint differential saturation	$s_X s_y$	The two sets of items have relatively large versus small degrees of differential saturation; indexed by the geometric mean of the variances of each set of factor loadings (i.e., $s_X s_y=\sqrt{s_X^2 s_Y^2}$).

experts, they might not be intuitively meaningful to many researchers conducting personality assessment research. Moreover, φ can yield counterintuitive results. Indeed, it has been described as "ambiguous" or difficult to interpret (e.g., Pinneau & Newhouse, 1964), and researchers have noted several "difficulties" that can produce "erroneous conclusion[s]" (Teel & Verran, 1991, p. 70) or even "ridiculous results" (Gorsuch, 1983, p. 285). For example, consider a recent examination of the factor structure of borderline personality disorder (BPD) symptoms (Hawkins et al., 2014). Analyses comparing a one-factor structure across two groups (individuals with or without BPD) produced a φ value indicating extremely strong factorial similarity. A different perspective, however, seems to suggest weak similarity—several symptoms that loaded strongly in the BPD group loaded weakly in the non-BPD group. In other words, the two groups had very different patternings, or configurations of loadings—the symptoms that were most fundamental to the factor in the BPD group seemed to be meaningfully different than those that were most fundamental in the non-BPD group. Thus, the φ value is inconsistent with a seemingly straightforward and intuitive examination of patterns of loadings. Such seemingly counterintuitive findings can create confusion over the meaning and interpretation of φ. A profile-based framework provides relatively intuitive insight into the meaning of φ and can help researchers understand, resolve, and interpret results that might initially seem counterintuitive, ambiguous, erroneous, or ridiculous.

A profile-based framework for factor similarity

Characterizing a set of factor loadings

As outlined in Table 1, this article characterizes a set of factor loadings in terms of three components: general factorial saturation, differential factorial saturation, and their patterning, or configuration. As an illustrative example, imagine that a researcher in moral psychology develops a 9-item assessment of moral cognition. Knowing the importance of factorial structure, she evaluates the factorial similarity of her assessment across groups of respondents—moral exemplars, typical moral individuals, and adolescent moral individuals. Males and females from each group complete her assessment, she conducts EFA within each group, and she extracts one factor within each group. Table 2 presents five hypothetical sets of loadings we have created, with φ values reflecting similarity

Table 2. Sets of factor solutions.

	Factor solutions for moral samples				
Item	Female moral exemplars	Male moral exemplars	Typical females	Typical males	Adolescent typical females
1	.90	.30	.60	.54	.24
2	.80	.60	.50	.53	.23
3	.70	.40	.40	.52	.22
4	.60	.50	.30	.51	.21
5	.50	.70	.20	.50	.20
6	.40	.80	.10	.49	.19
7	.30	.90	.00	.48	.18
8	.20	.10	−.10	.47	.17
9	.10	.20	−.20	.46	.16
General saturation ($\overline{X}$, $\overline{Y}$)	.50	.50	.20	.50	.20
Differential saturation (s_X, s_Y)	.27	.27	.27	.03	.03
With female exemplars					With typical females:
Congruence coefficient (φ)		.79	.91	.90	.69
General saturation similarity $\lvert\overline{X}-\overline{Y}\rvert$		.00	.30	.00	.00
Differential saturation similarity $\lvert s_X - s_y\rvert$		.00	.00	.24	.24
Configural similarity (r_{XY})		.10	1.00	1.00	1.00
Joint general saturation ($\overline{XY}$)		.50	.32	.50	.20
Joint differential saturation ($s_X s_y$)		.27	.27	.09	.09

between the female moral exemplars' loadings and loadings from the other groups.

General saturation

General saturation is the degree to which a set of items is generally affected by, or "saturated" with the latent variable, indexed by the mean factor loading in the set ($\overline{X}$, $\overline{Y}$). Consider the female moral exemplars in Table 2—their mean loading is .50, suggesting a moderate level of saturation. Larger values suggest that, in general, the items are strongly connected to the relevant latent variable. Smaller values (e.g., among the typical females in Table 2) suggest that, in general, the items are weakly connected to the latent variable.

Differential saturation

Differential saturation is the degree to which the items within a set are differentially affected, or saturated with the latent variable, indexed by the standard deviation or variance of the loadings (s_X, s_Y). Again consider Table 2's female exemplars—their loadings range widely from .10 to .90, with a standard deviation of .27. Thus, some items load strongly, indicating a robust connection to the latent variable. Other items load weakly, indicating little connection to the latent variable. Greater differential saturation suggests that items are widely dispersed in terms of their factorial saturation, whereas less differential saturation suggests that they are more uniformly saturated with the latent variable (e.g., the typical males in Table 2).

Configuration

Configuration is the particular patterning with which items are differentially affected by the latent variable. Within Table 2's female moral exemplars, Items 1 and 2 are strongly linked to the latent variable, but Items 8 and 9 are weakly linked. This configuration of loadings might reveal something meaningful about the latent variable driving the female moral exemplars' responses to the items. Unlike general and differential saturation, no index reflects configuration in a general sense. However, as we shall see, we can index the degree to which a set of loadings reflects a particular configuration of interest.

Characterizing factor similarity

Just as general saturation, differential saturation, and configuration can characterize a given set of loadings, they can characterize the similarity between sets of loadings (again, see Table 1).

General saturation similarity

General saturation similarity is the degree to which two sets of items are generally affected by a latent variable to the same level. Indexed as the absolute value of the difference between general saturation levels, $|\overline{X} - \overline{Y}|$ (i.e., between means of each set of factor loadings), low values represent similarity and high values represent dissimilarity—see Table 2's general saturation similarity row.

For example, female and male moral exemplars are identical in general saturation. The difference is zero ($|\overline{X} - \overline{Y}| = |.50 - .50| = 0$), indicating perfect similarity. Thus, these groups' responses to the scale are affected by the relevant latent variables to the same general degree. In contrast, the female exemplars' loadings differ from those of the typical females, whose general saturation is only .20, with a dissimilarity of .30 ($|\overline{X} - \overline{Y}| = |.50 - .20| = .30$). Thus, as compared to responses by female exemplars, responses by typical females are less connected to the latent variable and are more affected by factors other than that latent variable.

Differential saturation similarity

As defined in Table 1, differential saturation similarity is the degree to which the differential saturation in one set of items matches the differential saturation in the other set. Indexed as the absolute value of the difference between differential saturation levels (i.e., between standard deviations of each set of factor loadings, $|s_X - s_y|$), low values represent similarity and high values represent dissimilarity; see the differential saturation similarity row in Table 2.[3]

Female and male moral exemplars are again identical, with a difference of zero ($|s_X - s_y| = |.27 - .27| = 0$), indicating perfect similarity. Thus, the dispersion of factor loadings is the same within each group. In contrast, the female exemplars' factor loadings differ from the typical males' loadings, where differential saturation is only .03, for a dissimilarity of .24 ($|.27 - .03| = .24$). Thus, as compared to the responses by female moral exemplars, responses by typical males are more uniformly affected by the latent variable.

Configural similarity

Configural similarity is the degree to which the patterning of loadings is similar across two sets of factor loadings (see Table 1). Indexed as the Pearson correlation between two sets of factor loadings, r_{xy}, positive values reflect similarity (with larger values representing greater similarity), values close to 0 reflect dissimilarity, and negative values suggest that the pattern of loadings is opposite in the two groups—see the configural similarity row in Table 2.

For example, the female moral exemplars' factor loadings are configurally similar to the typical females' loadings. The items loading closest to 1.0 (or –1.0) in one group also load closest to 1.0 (or –1.0) in the other. In fact, the correlation is $r_{XY} = 1.0$, indicating perfect configural similarity. In contrast, the female exemplars' configuration of loadings differs from the male exemplars' configuration, with similarity of only $r_{XY} = .10$. Thus, the patterning with which the latent variable affects the items among female exemplars is quite unlike the patterning with which it affects items among male exemplars.

Researchers have used the Pearson correlation to index factorial similarity (Hopwood & Donnellan, 2010; Louks, Hayne, & Smith, 1989; Pinneau & Newhouse, 1964; Teel & Verran, 1991). Although we know of no research on this issue, informal discussions with colleagues and our review of the literature suggest that researchers might assume that configural similarity is precisely what is meant by factorial similarity and they also might assume that the φ coefficient is essentially a Pearson correlation between two sets of loadings. As we shall see, such

[3]Differential (and general) saturation similarity can also be indexed in ways other than taking the difference in standard deviations (or means), including taking the ratio of the variances or the standard deviations. We use the difference method here for its relative simplicity.

assumptions are mistaken and could lead to confusion when examining factorial similarity.

The congruence coefficient

Several indexes are available for evaluating factorial similarity across groups, including the R_V coefficient (Abdi, 2007), the KHB index (Kaiser, Hunka, & Bianchini, 1971), and the salient variable similarity index (SVS; Cattell & Baggaley, 1960). However, the *congruence coefficient* appears to be the most widely used (e.g., Bellmann, 2016; Chan, Ho, Leung, Chan, & Yung, 1999; Chmielewski & Watson, 2008; Cooke & Michie, 2001; DeYoung, Weisberg, Quilty, & Peterson, 2013; Fossati et al., 2007; Harpur, Hakstian, & Hare, 1988; Hawkins et al., 2014; Hopwood & Donnellan, 2010; Jensen, 1983; Livesley, Jackson, & Schroeder, 1992; Rammstedt & Farmer, 2013; Soto & John, 2014; Weiss et al., 2015; Yamagata et al., 2006).

A φ reflects the similarity between two set of factor loadings, most commonly when comparing factors across groups. Based on the cross-products of factor loadings across groups, it is:

$$\varphi = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \tag{1}$$

Here, x and y represent two sets of factor loadings, with both sets based on the same or highly corresponding measured variables—for example, the same set of items.

Table 2 also presents φ values reflecting the similarity between the female moral exemplars' loadings and the loadings from the other groups. For example, φ between the female exemplars and typical females is relatively large at .91:

$$\varphi = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$
$$\varphi = \frac{(.90)(.60) + (.80)(.50) + (.70)(.40) + \cdots + (.10)(-.20)}{\sqrt{[(.90^2)(.80^2)\ \ldots\ (.10^2)][(.60^2)(.50^2)\ \ldots\ (-.20^2)]}}$$
$$\varphi = \frac{1.50}{\sqrt{(2.85)(.96)}}$$
$$\varphi = .91$$

Although large compared to the Table 2's other φ values, this reflects "fair" similarity by conventional guidelines. Ranging from −1 to +1, larger φ values indicate greater factorial similarity: Values $\geq .95$ indicate that factors are roughly equivalent, values of .85 to .94 indicate "fair" similarity, and values $\leq .85$ indicate dissimilarity (Lorenzo-Seva & ten Berge, 2006).[4]

A profile-based perspective on the congruence coefficient

As summarized in Table 1, the profile-based components provide a framework for understanding the φ, and they reveal conditions under which φ will be large or small. In terms of these components:

$$\varphi = \frac{(r_{XY} S_X S_Y) + \overline{XY}}{\sqrt{\left(S_X^2 + \overline{X}^2\right)\left(S_Y^2 + \overline{Y}^2\right)}} \tag{2}$$

(see proof in the Appendix). Here, X and Y are two sets of factor loadings, r_{XY} is the correlation between the sets (i.e., configural similarity), s_X, s_Y, s_X^2, s_X^2 are standard deviations and variances within each set (i.e., differential saturations), and $\overline{X}$ and $\overline{Y}$ are means of each set (i.e., general saturations). For example, φ between female exemplars and typical females (from Table 2) is:

$$\varphi = \frac{(1.00)(.27)(.27) + (.50)(.20)}{\sqrt{[(.27^2 + .50^2)(.27^2 + .20^2)]}}$$
$$\varphi = \frac{.073 + .100}{\sqrt{(.323)(.113)}}$$
$$\varphi = \frac{.173}{.191}$$
$$\varphi = .91$$

To understand the widely used congruence coefficient in general, it is useful to understand each component's specific effects on the coefficient.

Effect of configural similarity

Figure 1a reveals the positive linear impact of configural similarity on φ. To focus on these effects, configural similarity (i.e., the degree to which two sets of loadings have similar pattern of high-vs.-low loadings across items) varies from $r_{XY} = -1$ to $+1$, and general saturation is held constant across the two sets of factor loadings at .50 (i.e., $\overline{X} = \overline{Y} = .50$), as is differential saturation (i.e., $S_X = S_Y$). All else being equal, greater configural similarity produces a higher φ.

Figure 1a also reveals that differential saturation moderates configural similarity's effect. At low levels of differential saturation (e.g., $S_X = S_Y = .1$, $S_X S_Y = .01$), configural similarity has minimal effect on φ; however, at greater levels of differential saturation ($S_X = S_Y = .8$; $S_X S_Y = .64$), configural similarity has a dramatic effect on φ.

Figure 1a also reveals examples of potentially "ridiculous" or "ambiguous" results to which φ is susceptible. On one hand, φ can be quite robust ($> .90$) when configural similarity is low, zero, or even negative. For example, with little differential saturation (e.g., $S_X = S_Y = .01$), φ can be robust (e.g., $\varphi = .96$), even though configural similarity suggests severe dissimilarity, at $r_{XY} = .05$. On the other hand, φ can be *small* when configural similarity is quite strong. This is apparent in Table 2, comparing typical females and adolescent females. In this case, φ of .69 indicates dissimilar factors, despite perfect configural similarity ($r_{XY} = 1.0$).

As we mentioned earlier, our impression is that researchers assume that φ is essentially configural similarity (i.e., that is very close to r_{xy}) and that configural similarity is at the heart of factorial similarity. Under such assumptions, obtaining a large φ when configural similarity is weak (or obtaining a small φ when configural similarity is robust) would indeed seem

[4] However, it should be noted that there is some disagreement about the guidelines for interpreting the congruence coefficient, and there is relatively little research on what these guidelines are based on (see Lorenzo-Seva & ten Berge, 2006).

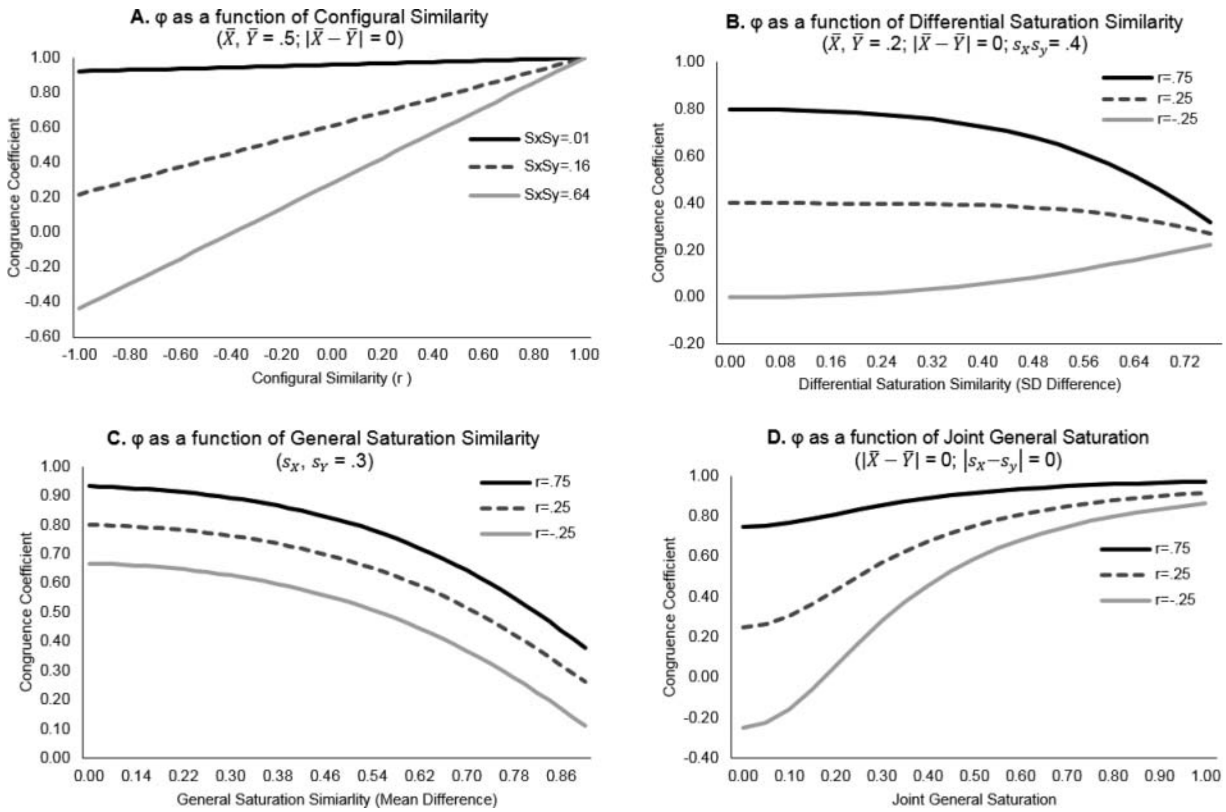

Figure 1. The impact of (a) configural similarity, (b) differential saturation similarity, (c) general saturation similarity, and (d) joint general saturation on the congruence coefficient. $|\bar{X} - \bar{Y}|$ = general saturation similarity; $|s_X - s_y|$ = differential saturation similarity; $s_X s_y$ = joint differential saturation; φ = congruence coefficient; *SD* = standard deviation; *r* = Pearson's *r*. In Figure 1c, the arithmetic mean of the factor solution means was held constant at .5.

problematic. That is, researchers who expect φ to reflect r_{xy} or who view r_{xy} as truly indicating factorial similarity are likely to be confused by robust discrepancies between φ and r_{xy}. This could lead to doubts about the utility of φ and confusion about appropriate conclusions regarding factorial similarity.

Effect of differential saturation similarity

Figure 1b shows the somewhat more complex impact of differential saturation similarity. Here, differential saturation similarity (i.e., the degree to which two sets of loadings have similar variability in their loadings, $|s_X - s_y|$) varies from 0 (perfect similarity) to .80 (relatively extreme dissimilarity).[5] General saturation is constant across the two sets of loadings at .20 (i.e., $\bar{X} = \bar{Y} = .20$), with effects shown at several levels of configural similarity.

In general, as there is more dissimilarity between the groups' differential saturation, φ moves toward the average saturation (here, .20). In some cases, dissimilarity in differential saturation decreases φ, but in others, it increases φ. For example, when configurations are dissimilar ($r = -.25$), greater differential saturation dissimilarity moves φ upward toward .20. This effect might seem counterintuitive—lower similarity (in terms of differential saturation) can produce higher values of apparent congruence.

Effect of general saturation similarity

Figure 1c illustrates the straightforward effect of general saturation similarity. General saturation (dis)similarity (i.e., the degree to which two sets of loadings have similar mean factor loadings ($|\bar{X} - \bar{Y}|$) varies from 0 to .90, whereas differential saturation is held constant across the two sets of loadings at $s_X = s_y = .30$ ($|s_X - s_y| = 0$), with effects shown at configural similarity levels of $r_{XY} = -.25$, .25, and .75.

As Figure 1c shows, φ decreases as dissimilarity in general saturation increases. That is, φ is larger when two sets of factor loadings have similar levels of general saturation (e.g., $|\bar{X} - \bar{Y}| = 0$) than when they have different levels of general saturation (e.g., $|\bar{X} - \bar{Y}| = .86$).

Effect of joint general saturation

The congruence coefficient is also directly affected by a component that can be articulated in terms of profiles—the "joint" general saturation across both sets of factor loadings. Joint general saturation is the degree to which items are affected by the latent variable in general, across both groups (e.g., when $\bar{X}$ and $\bar{Y}$ are closer to .80 than to .20). In Equation 2, this is represented as $\overline{XY}$, the geometric mean of the (squared) general saturation levels from the two sets of

[5] In addition, the arithmetic mean of the two differential saturations is held constant at .40. Similarly, Figure 1c (the effects of general saturation similarity) holds the arithmetic mean of the two general saturations constant at .50. We return to this later, in the context of joint differential saturation, joint general saturation, and the geometric mean.

loadings (i.e., $\overline{XY} = \sqrt{\overline{X}^2\overline{Y}^2}$). Note that this is also highly related to the familiar arithmetic mean of absolute values of the two general saturation values.

Figure 1d varies joint general saturation from 0 to 1.00. To highlight the effect on φ, it assumes perfect general saturation similarity ($|\overline{X} - \overline{Y}| = 0$) and perfect differential saturation similarity at $s_X = s_y = .60$ ($|s_X - s_y| = 0$), and it varies configural similarity at three levels.

Figure 1d shows the robust and straightforward positive impact of joint general saturation on φ. That is, φ is larger when two sets of factor loadings have high joint general saturation (e.g., $\overline{XY} = .80$) than when they have lower joint general saturation. Notably, this effect is more robust when configural similarity is low (e.g., $r_{XY} = -.25$).

This effect can also be seen in Table 2. Consider two different comparisons: (a) comparing typical females and adolescent typical bemales, and (b) comparing female moral exemplars and typical males. In both comparisons, the two sets of loadings are identical in their general saturation similarity, differential saturation similarity, and configural similarity (i.e., for both $|\overline{X} - \overline{Y}| = .00$, $|s_X - s_y| = .25$, and $r_{XY} = 1.00$). However, the bottom row of Table 2 reveals an important difference in these comparisons: The female moral exemplars and typical males have a large joint general saturation of .50, whereas the typical females and adolescent typical females have a lower joint general saturation of .20. This difference has a large impact on φ, which is .90 between the female exemplars and typical males, but only .69 between the typical females and the adolescent typical females. Simply stated, the higher the loadings on the factor solutions, the higher φ will be, all else being equal.

Effect of joint differential saturation

Finally, φ is affected by the joint differential saturation across both sets of factor loadings. Joint differential saturation is the degree to which two sets of items have large versus small degrees of differential saturation (e.g., when s_X and s_y are closer to .70 than to .00). In Equation 2, this is represented as $s_X s_Y$, the geometric mean of the variances within each set of factor loadings (i.e., $s_X s_y = \sqrt{s_X^2 s_Y^2}$).

Joint differential saturation can be seen as indirectly affecting φ by moderating the effect of configural similarity. When joint differential saturation is small, the effect of configural similarity is weak (e.g., the flat $s_X s_y = .01$ line in Figure 1a). In contrast, when it is larger, the effect of configural similarity is magnified (i.e., see the strongly sloped $s_X s_y =$ .64 line in Figure 1a). Put another way, if there is generally little differentiation among factor loadings, then the particular configuration of those loadings does not matter much; however, if there is greater differentiation, then the configuration of loadings has a robust impact on φ.

Integration of effects

In sum, a profile-based perspective reveals intuitive components affecting φ. Perhaps the simplest way of interpreting this is that two primary components work independently to influence the size and direction of φ: (a) to the degree that the items generally have strong factor loadings—high joint general saturation—φ will be high, all else being equal; and (b) to the degree that the items in one group's factor have the same pattern or configuration of loadings as they do in the other group's factor—strong configural similarity—φ will be high, all else being equal. The other components moderate the relative impact of these two components.

For example, a high level of joint differential saturation (the degree to which sets of items have large vs. small degrees of differential saturation) magnifies configural similarity's effect, whereas a low level minimizes that effect. The final two components—general saturation similarity and differential saturation similarity—have less direct effects on φ. Their effects can be seen, in part, as flowing through the effects of joint general saturation and joint differential saturation, respectively. That is, as the general saturation (or differential saturation) of two sets of loadings becomes less similar, the joint general saturation (or joint differential saturation) across the two sets is reduced. For a discussion of these effects related to the fact that the geometric mean represents joint general saturation and joint differential saturation in Equation 2, please see the online supplemental materials.

Using the profile-based framework to interpret ambiguous results

A profile-based framework can help researchers interpret the results of factor analyses and how two solutions compare. Recall the examination (Hawkins et al., 2014), discussed earlier, of factorial structure of BPD symptoms, as compared across non-BPD and BPD participants. Table 3 presents the factor loadings from each analysis and each of the profile-based components.

First, configural similarity ($r_{XY} = .16$) is quite low. This suggests that, in terms of the particular symptoms that are most strongly saturated with the latent variable, the two groups differ to some degree. Indeed, the patterning of factor loadings is quite different across the two groups. For example, the highest loading symptom in the non-BPD group was impulsivity ($\lambda = .80$), which was the

Table 3. Comparisons of factor solutions from the Hawkins et al. (2014) data.

	Factor loadings		
Item	Non-BPD		BPD
Anger	.64		.74
Emotional instability	.73		.73
Relationship instability	.79		.81
Abandonment	.50		.74
Unstable identity	.70		.87
Suicide/self-harm	.50		.59
Paranoid ideation	.76		.90
Emptiness	.70		.56
Impulsivity	.80		.51
General saturation ($\overline{X}$, $\overline{Y}$)	.68		.72
Differential saturation (s_X, s_y)	.11		.13
Congruence coefficient (φ)		.98	
Configural similarity (r_{XY})		.16	
General saturation similarity $|\overline{X} - \overline{Y}|$		.04	
Differential saturation similarity $|s_X - s_y|$		.02	
Joint general saturation ($\overline{XY}$)		.49	
Joint differential saturation ($s_X s_y$)		.01	

Note. BPD = borderline personality disorder.

lowest loading symptom in the BPD group ($\lambda = .51$). Similarly, the second-highest loading symptom in the BPD group was unstable identity ($\lambda = .87$), which was only the sixth-highest loading symptom in the non-BPD group. The latent variable seems to be more strongly connected to impulsivity and relational instability in one group, but it seems to be more reflected in paranoia and identity confusion in the other group. This might be an important insight that reflects a meaningful psychological difference between people with and without BPD. It might have implications for both the measurement of that latent variable and, more deeply, for the understanding of that latent variable in general.

However, the high level of joint general saturation suggests that there is more to the story than what seems to be suggested by the apparent configural difference between groups. That is, the symptom responses across both groups are generally robustly saturated with the latent variable. This is reflected in the high level of joint general saturation (the geometric mean of the two sets of loadings was large at .70 (i.e., $\sqrt{\overline{X}\overline{Y}} = \sqrt{(.68)(.72)} = .70$, squared geometric mean = .49). There was also low joint differential saturation, revealing little variation among the factor loadings within the two sets of factor loadings ($s_X s_y = .11 \times .13 = .01$). This means that, within each group, the items or symptoms load fairly uniformly: It is not the case that some items are loading dramatically strongly and others load weakly. Together, these two pieces of information—high joint general saturation and low joint differential saturation—suggest that the items' or symptoms' factor loadings tend to be uniformly robust. With factor loadings that are uniformly of a large magnitude, we would conclude that, across groups, the items or symptoms are generally strongly reflective of the latent variable.

Strong similarity in terms of the other two similarity components underscores this latter conclusion. There is low general saturation dissimilarity ($M_{\text{nonBPD}} = .68$, $M_{\text{BPD}} = .72$, $|.68 - .72| = .04$), making it clear that the strong effect of the latent variable is indeed observed in both groups to a very similar degree. Thus, the latent variable is not more strongly connected to the items or symptoms in one group than in another (at least not to a level that is psychologically meaningful). In addition, there is low differential saturation dissimilarity ($SD_{\text{nonBPD}} = .11$, $SD_{\text{BPD}} = .13$, $|.11 - .13| = .02$), which suggests that the uniformity across items' loadings is similar across groups.

In sum, although there might be some interesting and meaningful difference between groups in terms of which specific symptoms are most strongly connected to the latent variable, the bigger picture is that the two groups are quite similar in the overall connection between the symptoms and the latent variables. That is, across both groups, the latent variable is strongly connected to all the symptoms. These findings support the idea that the two groups have similar factorial structure, but they reveal a way in which the two groups might have a subtle, but potentially interesting dissimilarity.

Understanding these components and their implications provides deeper insight into the nature of the congruence coefficient, and can clarify and resolve results that might, on the surface, seem counterintuitive or ambiguous. In this example, the high φ of .98 (again, see Table 3) might initially seem inconsistent with the low level of configural similarity, but as we discussed earlier, low joint differential saturation minimizes configural similarity's effect on φ. Thus, here, although configural similarity is weak, its effect on φ is small. Additionally, because the joint general saturation is so high, it overpowers the effect that the relatively low configural similarity would have on the congruence coefficient:

$$\varphi = \frac{(.16)(.11)(.13) + (.68)(.72)}{\sqrt{[(.11^2 + .68^2)(.13^2 + .72^2)]}}$$
$$\varphi = \frac{.002 + .490}{\sqrt{(.4745)(.5425)}}$$
$$\varphi = \frac{.492}{.507}$$
$$\varphi = .98$$

The psychological implication of this is that the two solutions are in fact, quite similar (as φ indicates). Although their patterning of loadings differs, they are both highly saturated with the latent variable and there is little variation in the loadings within each set.

Consider a second example by revisiting the Table 2 comparison of typical females and adolescent females. Examining the profile-based components reveals that although configural similarity is perfect in this case, there is a greater dissimilarity in differential saturation ($s_x = .27$, $s_y = .03$, $|.27 - .03| = .24$). In other words, although the patterning of two sets of loadings is identical, the variability of the loadings differs within in each set: The typical female sample has more variable loadings than the adolescent females. Moreover, there is low joint differential saturation ($s_X s_y = .27 \times .03 = .008$), meaning that overall, the two sets of loadings have a low degree of variability. Finally, there is low joint general saturation (i.e., $\sqrt{\overline{X}\overline{Y}} = \sqrt{(.20)(.20)} = .20$, squared geometric mean = .04), meaning the two sets, overall, are not very saturated with the latent variable. In this case, the small degree of joint general saturation overpowers the effect of perfect configural similarity, which has been minimized by the small degree of joint differential saturation. The psychological implication of this is that the two sets, despite having loadings that share the same patterning, are in fact, quite different.

In summary, the congruence coefficient is an omnibus statistic that blends several separable components of factor similarity, all of which can be elucidated and understood by the profile-based approach we have already outlined.

The impact of profile-based components on expert judgments of factor similarity

The profile-based framework reveals components underlying factorial similarity in general and the congruence coefficient in particular. These components might have different psychological implications, and researchers might view some components as particularly relevant to gauging factorial similarity. Although previous research reveals that researchers' subjective judgments of factorial similarity correspond with the congruence coefficient (Lorenzo-Seva & ten Berge, 2006), we are aware of no work revealing what researchers actually attend to when gauging factorial similarity.

The profile-based framework might reveal precisely what researchers deem relevant to evaluating factorial similarity, even if they are not fully aware of the components that shape their evaluations. As we have mentioned, we suspect that researchers' evaluations of factorial similarity track configural similarity more closely than other components of similarity. However, it is quite possible that researchers' evaluations are heavily influenced by other components as much as, or even more than configural similarity. Moreover, it is possible that different researchers base their evaluations of factorial similarity on different components. Some might attend to configural similarity, whereas others might attend to, say, general saturation similarity.

Although expert judgments are not the gold standard of similarity, it is crucial to examine them for several reasons. First, they can reveal whether the profile-based approach has psychological resonance with researchers' subjective evaluations of factorial similarity. If the components are associated with judgments of similarity, this would support the meaningfulness of the framework. If not, however, then the framework (beyond its relevance for understanding φ) might have little psychological utility. Second, these judgments could explain why the congruence coefficient can be confusing. For example, if researchers' ratings of similarity are most influenced by configural similarity, this explains why confusion arises when Pearson's r diverges from φ (e.g., Gorsuch, 1983). Finally, these issues might reveal sources of disagreement when two or more researchers are gauging factorial similarity. If researchers differ in what they attend to, such differences can produce divergent conclusions. For example, if one researcher attends primarily to configural similarity, and another attends primarily to general saturation similarity, those researchers might disagree about overall factorial similarity in a given case (e.g., when configural similarity is high but general saturation similarity is low).

Participants and data set

We examined these issues in data provided by Lorenzo-Seva and ten Berge (2006). Participants ($N = 56$) or "judges" were researchers in the area of intelligence, personality, or social psychology "who apply EFA in their research" (p. 59). Each judge received eight pairs of factors (i.e., each pair made up of two sets of factor loadings), and they rated the similarity for each pair, on a 5-point Likert scale ranging from 1 (*very poor*) to 5 (*very good*). Judges were divided into 10 panels of 3 to 11 judges each, with each panel receiving a different set of eight factor pairs to rate. Thus, 80 separate pairs of factors were rated, with an average of 5.6 judges independently rating each, for a total of 448 ratings.

For each factor pair, we calculated the profile-based components discussed earlier (general saturation similarity, configural similarity, etc.). Although factor pairs had not been created to vary in terms of these components, each component varied to some degree (see Table 4 for ranges of each component, across factor pairs). However, some components varied more than others. For example, configural similarity varied from –.54 to .98, but general saturation similarity varied only from .00 to .13. Moreover, the components were correlated with each other to various degrees across factor pairs. For example, factor pairs that had high levels of configural similarity also tended to have high levels of general saturation similarity. Our goal was to identify the connections between these components and experts' ratings of similarity for each factor pair.

Results and discussion

We conducted preliminary analyses evaluating the data's nested structure. A cross-classified unconditional multilevel model with the 448 similarity ratings as the outcome suggested that judge and factor pair should be modeled as random effects. Thus, all subsequent analyses are based on cross-classified multilevel models with these variables as random effects.

For main analyses, each model predicted the judges' 448 subjective ratings of the factor pairs' similarity from one or more profile-based components derived from each pair's actual factor loadings. First, we examined each component separately in its own model, estimating both a fixed effect and random effect (centered within judge). As shown in Table 4, each component was significantly associated with subjective ratings of factorial similarity. For example, configural similarity's slope of 2.72 ($p < .001$) indicates that (according to the average judge) a 1-unit difference in configural similarity (e.g., going from $r_{XY} = .00$ to $r_{XY} = 1.0$) produced a difference of 2.72 in subjective ratings of similarity. The other slopes are significant and larger, with the differences in size likely related to the vast difference in ranges of the components and the fact that they are reflecting different phenomena. In terms of a more standardized effect size, reduction in unexplained variance, configural similarity has the largest effect, although again this likely reflects the fact that configural similarity varied much more than any other

Table 4. Fixed effects of components of factor similarity on judges' subjective ratings of factor similarity.

		Model			
		Each component separately		All components Together	
Component (predictors)	Range	B	% Var exp	B	% Var exp
General saturation similarity	.00–13	-50.72^{***}	11%	-29.49^{***}	8%
Differential saturation similarity	.00–.22	-27.81^{***}	7%	8.45	0%
Configural similarity	–.54–.98	2.72^{***}	26%	2.34^{***}	21%
Joint general saturation	.06–.18	-82.06^{**}	15%	−24.55	0%
Joint differential saturation	.03–.20	46.45^{**}	11%	13.41	0%

Note. % var exp = % of variance explained. For the model that includes all components together, the % Var Exp value for a given component was estimated by comparing the residual variance of the full model (with all five components) to the residual variance of a model that included the four components aside from the given component. These values estimate the variance explained by each component, over and above all other components.
$^{**}p < .01$. $^{***}p < .001$.

component. In general, expert ratings of similarity seem to be driven by high levels of configural similarity, low levels of general saturation (dis)similarity, and low levels of differential saturation (dis)similarity. Ratings also appear linked to low levels of joint general saturation and high levels of joint differential saturation, although these effects might be by-products of correlations among predictors, which we examine shortly.

Interestingly, the random effects (i.e., variance components) presented in Table 5 suggest that judges differ in their attention to most components. The effects were significant across judges for all components except differential saturation similarity. Thus, some judges' ratings were more highly connected to, say, configural similarity or general saturation similarity, than were other judges' ratings. This might reflect disagreement in expert judges' thinking about factorial similarity.

Because stimuli were not created to orthogonally differentiate the components from each other, they were correlated with each other across factor pairs. Thus, the apparent effect of a given component might be a spurious by-product of its association with another component that directly affects similarity ratings. To evaluate this, we entered all components as predictors of similarity ratings in a single model. As shown in Table 4, only configural similarity and general saturation similarity were significant at $p <$ .001 (Bs = –29.49 and 2.34, respectively). Thus, these two components seem to have the most direct and unique effects on judges' ratings of factor similarity. This also suggests that the apparent effects of the other three components are likely due to overlap between those components and either configural similarity or general saturation similarity (or both).

Finally, we examined whether differential saturation moderates the effect of configural similarity on subjective ratings of similarity. Again, Figure 1a reveals this moderating effect on the congruence coefficient, but it is not clear that expert judges also weight the effect of configural similarity by the degree of differential saturation. For this analysis, we entered configural similarity, joint differential saturation, and their product as predictors of subjective ratings. Of main interest, the product's slope was positive and significant ($B =$ 24.63, $p = .049$). This indicates that as joint differential saturation increases, expert judges place more weight on configural similarity in shaping their evaluations of factorial similarity.

In sum, these findings suggest that configural similarity and general saturation similarity are the components most fundamental to experts' evaluations of factorial similarity. If two sets of loadings have a similar pattern of high vs. low loadings or similar general level of magnitude, expert judges see those sets as being similar. Moreover, expert judges weight configural similarity by the degree of differentiation among the loadings. When the loadings vary dramatically, judges weight the patterning of loadings highly. Controlling for these effects, no other components are significantly related to judges' evaluations of similarity.

These findings partly explain why confusion might arise when researchers interpret congruence coefficients. Researchers do not seem to directly attend to joint general saturation (at least not after controlling for its overlap with other components); however, joint general saturation has a direct and potentially robust effect on φ (see Equation 2 and Figure 1d). Thus, the expert raters "agree" with φ, in terms of the importance of configural similarity, and even in terms of the moderating effect of joint differential saturation. However, they seem to disagree with the congruence coefficient's weighting of joint general saturation. This disagreement can lead to instances where the congruence coefficient seems to indicate a level of similarity that (if based primarily on joint general saturation) researchers might see as counterintuitive.

It is important to reiterate that the stimuli from Lorenzo-Seva and ten Berge (2006) were not designed to examine components of factorial similarity, and there happened to be far more variability in configural similarity than in other components. This might lead to underestimation of effects for those other components, at least in terms of standardized effect sizes. Thus, although novel and suggestive, these results should not be viewed as conclusive until replicated with stimuli that are orthogonally manipulated with greater variability across all components.

Table 5. Random effects of components of factor similarity on judges' subjective ratings of factor similarity.

	Model			
	Each component separately		All components Together	
Component (predictors)	Est.	Wald Z	Est.	Wald Z
Judge	—	—	.34***	4.89
Factor pair	—	—	.00	.73
General saturation similarity	196.42*	2.16	150.32*	2.00
Differential saturation similarity	54.57	1.63	—	—
Configural similarity	.741**	2.61	.620**	2.68
Joint general saturation	6428.19*	2.37	—	—
Joint differential saturation	925.65*	1.94	—	—

Note. Est. = estimate.
*$p < .05$. **$p < .01$. ***$p < .001$.

Conclusions and recommendations

This work articulates a profile-based framework for understanding factorial similarity in the context of EFA. We believe this framework offers insight into the ways in which factorial similarity can be conceptualized, enhancing the precision and flexibility with which researchers can examine factorial structure. Indeed, it might open new opportunities for discovery when examining factorial similarity.

This framework also enhances researchers' ability to understand and interpret the most commonly used index of factorial similarity in EFA—the congruence coefficient (φ). To our knowledge the congruence coefficient is rarely (if ever) articulated in a way that clarifies its meaning or that outlines precisely what makes it larger or smaller. Such lack of clarity produces paradoxical situations in which researchers might find the congruence coefficient to be ambiguous, inconsistent with their intuitions, or even "ridiculous" (Gorsuch, 1983, p. 285). The profile-based framework resolves such ambiguity by demonstrating that the congruence coefficient emerges clearly and straightforwardly from a set of intuitive components.

To reiterate, we have discussed five components (summarized in Table 1), three of which are directly related to similarity: general saturation similarity, differential saturation similarity, and configural similarity. It is worth considering whether the two remaining components that we have discussed should be considered relevant to evaluating factorial similarity: joint general saturation—the degree to which the measured variables in both groups tend to be connected to the latent variable—and joint differential saturation—the degree to which the measured variables are differentially connected to the latent variable in both groups. Neither directly reflects similarity; rather, they reflect qualities as aggregated across factors (or groups). For example, we could find joint general saturation of, say, .16 when two sets of loadings have identical general saturation (i.e., $\overline{XY} = .40 \times .40$) or when they have dramatically dissimilar general saturation (e.g., $.80 \times .20$). Therefore, joint general saturation provides no direct information about the similarity between sets of loadings, and our analyses suggest that experts do not view it as directly relevant to evaluating factorial similarity. In the same way, joint differential saturation itself is not directly relevant to similarity, but it seems to indirectly affect experts' ratings via moderating the effect of configural similarity. Thus, expert researchers do seem to view (consciously or otherwise) joint differential saturation as being indirectly relevant to evaluating factorial similarity.

The fact that the congruence coefficient is directly affected by joint general saturation might reveal a key logical and practical problem with φ. Logically, it is not directly related to similarity, and experts seem not to consider it relevant. Thus, practically, it might be the primary reason that researchers sometimes find φ to be ambiguous, erroneous, or ridiculous.

These observations suggest two recommended alternatives for examining factorial similarity in an EFA context. One alternative would be to avoid φ as a unitary index of factorial similarity, and adopt a more differentiated framework. By blending several separable and meaningful components, φ is an "omnibus" index that might not reliably or clearly reflect any single component. This is potentially problematic particularly because the components themselves are generally orthogonal, except at the extremes. That is, general saturation similarity has no statistical or conceptual overlap with configural similarity or differential saturation similarity (again, except at the extremes). Indeed, indexes that blend multiple orthogonal components are often avoided (e.g., Cronbach, 1955; Nunnally, 1962). Moreover, as just discussed, φ also includes at least one component that does not directly reflect similarity.[6]

However, avoiding φ altogether is likely a less viable option, as it is likely the most widely used index of factorial similarity in EFA, and researchers might hesitate to trade its simplicity. Thus, the most practical alternative to avoiding φ is to supplement it with examination of the profile-based components. Separately examining these components can provide researchers with a fuller, more coherent understanding of their data, and helps detect whether φ obscures an important finding. If this examination reveals something potentially problematic or "ambiguous," we encourage researchers to confront it directly and resolve the meaning and implications of the findings. The profile-based framework provides a way of doing just that.

As it requires evaluating several pieces of information rather than only one, a profile-based approach might be slightly more complex than simply relying on the congruence coefficient. However, this cost is likely outweighed by the benefits, in terms of clarity, precision, and avoidance of the ambiguity that can—and often does—accompany the congruence coefficient. We hope this profile-based framework enhances work done by researchers who examine factorial similarity in an EFA context. We believe it can clarify findings, resolve ambiguities, and offer insights that might be missed by solely relying on the most common alternative index.

[6] As we noted earlier, although φ is the most widely used index of factorial similarity in EFA, there are others (R_V coefficient, KHB index, SVS index). Although it is beyond the scope of this article, such indexes also merit examination in terms of a profile-based framework.

Acknowledgments

We would like to thank Urbano Lorenzo-Seva for his generosity in sharing his data set with us.

Funding

This project and publication was made possible through the support of a grant from the Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Templeton World Charity Foundation.

References

Abdi, H. (2007). RV coefficient and congruence coefficient. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 849–853). Thousand Oaks, CA: Sage.

Baker, B. O., & Block, J. (1957). Accuracy of interpersonal predictions as a function of judge and object characteristics. *Journal of Abnormal and Social Psychology*, *54*, 37–43.

Bellmann, M. (2016). Factor analysis of geometric figures with four attributes: A comparison of PCA, varimax and varimin. *Personality and Individual Differences*, *90*, 326–331.

Burt, C. (1948). The factorial study of temperamental traits. *British Journal of Statistical Psychology*, *1*, 178–203.

Cattell, R. B. (1949). rp and other coefficients of pattern similarity. *Psychometrika*, *14*, 279–298.

Cattell, R. B., & Baggaley, A. R. (1960). The salient variable similarity index for factor matching. *British Journal of Statistical Psychology*, *13*(1), 33–46.

Chan, W., Ho, R., Leung, K., Chan, D., & Yung, Y. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods*, *4*, 378–402.

Chmielewski, M., & Watson, D. (2008). The heterogeneous structure of schizotypal personality disorder: Item-level factors of the Schizotypal Personality Questionnaire and their associations with obsessive–compulsive disorder symptoms, dissociative tendencies, and normal personality. *Journal of Abnormal Psychology*, *117*, 364.

Cooke, D., & Michie, C. (2001). Refining the construct of psychopathy: Towards a hierarchical model. *Psychological Assessment*, *13*, 171–188.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52(3)*, 177.

Cronbach, L., & Gleser, G. (1953). Assessing similarity between profiles. *Psychological Bulletin*, *50*, 456–473.

DeYoung, C. G., Weisberg, Y. J., Quilty, L. C., & Peterson, J. B. (2013). Unifying the aspects of the Big Five, the interpersonal circumplex, and trait affiliation. *Journal of Personality*, *81*, 465–475.

Fossati, A., Barratt, E. S., Borroni, S., Villa, D., Grazioli, F., & Maffei, C. (2007). Impulsivity, aggressiveness, and DSM–IV personality disorders. *Psychiatry Research*, *149*(1), 157–167.

Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality*, *76*, 1267–1316.

Furr, M. (2010). The double-entry intraclass correlation as an index of profile similarity: Meaning, limitations, and alternatives. *Journal of Personality Assessment*, *92*, 1–15.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: University of Chicago Press.

Harpur, T. J., Hakstian, A. R., & Hare, R. D. (1988). Factor structure of the Psychopathy Checklist. *Journal of Consulting and Clinical Psychology*, *56*, 741.

Hawkins, A., Furr, M., Arnold, E., Law, M., Mneimne, M., & Fleeson, W. (2014). The structure of borderline personality disorder symptoms: A multi-method, multi-sample examination. *Personality Disorders: Theory, Research, and Treatment*, *5*, 380–389.
Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346.
Jensen, A. R. (1983). Effects of inbreeding on mental-ability factors. *Personality and Individual Differences*, *4*(1), 71–87.
Jensen, A. R. (1998). *The G factor: The science of mental ability*. Westport, CT: Praeger.
Kaiser, H., Hunka, S., & Bianchini, J. (1971). Relating factors between studies based upon different individuals. *Multivariate Behavioral Research*, *6*, 409–422.
Livesley, W. J., Jackson, D. N., & Schroeder, M. L. (1992). Factorial structure of traits delineating personality disorders in clinical and general population samples. *Journal of Abnormal Psychology*, *101*, 432.
Lorenzo-Seva, U., & ten Berge, J. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*(2), 57–64.
Louks, J., Hayne, C., & Smith, J. (1989). Replicated factor structure of the Beck Depression Inventory. *The Journal of Nervous and Mental Disease*, *177*, 473–479.
Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck and R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and new directions* (pp. 131–152). Mahwah, NJ: Erlbaum.
Nunnally, J. C. (1962). The analysis of profile data. *Psychological Bulletin*, *59*, 311–319.
Pinneau, S., & Newhouse, A. (1964). Measures of invariance and comparability in factor analysis for fixed variables. *Psychometrika*, *29*, 271–281.
Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment*, *25*, 1137.
Revelle, W. (n. d.). Constructs, components, and factor models. Retrieved from http://www.personality-project.org/r/book/Chapter6.pdf
Soto, C. J., & John, O. P. (2014). Traits in transition: The structure of parent-reported personality traits from early childhood to early adulthood. *Journal of Personality*, *82*, 182–199.
Teel, C., & Verran, J. A. (1991). Factor comparison across studies. *Research in Nursing & Health*, *14*(1), 67–72.
Tucker, L. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: U.S. Department of the Army.
Weiss, A., Staes, N., Pereboom, J. J., Inoue-Murayama, M., Stevens, J. M., & Eens, M. (2015). Personality in bonobos. *Psychological Science*, *26*, 1430–1439.
Yamagata, S., Suzuki, A., Ando, J., Ono, Y., Kijima, N., Yoshimura, K., … Kang, J. L. (2006). Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. *Journal of Personality and Social Psychology*, *90*, 987.

Appendix

By definition, the congruence coefficient is:

$$\varphi = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \tag{A.1}$$

This can be initially recast via the covariance between X and Y, and the variances of X and Y. The covariance between X and Y is

$$S_{XY} = \frac{\sum XY}{N} - \overline{XY}$$

Rearranging this produces the numerator of Equation A.1:

$$\sum XY = N\left(S_{XY} + \overline{XY}\right) \tag{A.2}$$

The variance of X is

$$S_X^2 = \frac{\sum X^2}{N} - \overline{X}^2$$

Rearranging this produces one element of Equation A.1′s denominator,

$$\sum X^2 = N\left(S_X^2 + \overline{X}^2\right) \tag{A.3}$$

Similarly, the variance of Y can be rearranged to produce the other element of Equation A.1′s denominator,

$$\sum Y^2 = N\left(S_Y^2 + \overline{Y}^2\right) \tag{A.4}$$

Replacing Equation A.1′s numerator with Equation A.2, and Equation A.1′s denominator with Equations A.3 and A.4,

$$\varphi = \frac{N\left(S_{XY} + \overline{XY}\right)}{\sqrt{N\left(S_X^2 + \overline{X}^2\right)N\left(S_Y^2 + \overline{Y}^2\right)}}$$

Simplifying,

$$\varphi = \frac{S_{XY} + \overline{XY}}{\sqrt{\left(S_X^2 + \overline{X}^2\right)\left(S_Y^2 + \overline{Y}^2\right)}}$$

Given that $S_{XY} = r_{XY}S_XS_Y$, the congruence coefficient can be framed in terms of correlation between X and Y, the standard deviations of X and Y, and the means of X and Y:

$$\varphi = \frac{r_{XY}S_XS_Y + \overline{XY}}{\sqrt{\left(S_X^2 + \overline{X}^2\right)\left(S_Y^2 + \overline{Y}^2\right)}}$$