Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity

Urbano Lorenzo-Seva[1] and Jos M. F. ten Berge[2]

[1]Rovira i Virgili University, Spain
[2]University of Groningen, the Netherlands

Abstract. When Tucker's congruence coefficient is used to assess the similarity of factor interpretations, it is desirable to have a critical congruence level less than unity that can be regarded as indicative of identity of the factors. The literature only reports rules of thumb. The present article repeats and broadens the approach used in the study by Haven and ten Berge (1977). It aims to find a critical congruence level on the basis of judgments of factor similarity by practitioners of factor analysis. Our results suggest that a value in the range .85–.94 corresponds to a fair similarity, while a value higher than .95 implies that the two factors or components compared can be considered equal.

Keywords: congruence coefficient, factor congruence, factor similarity, factor interpretation, exploratory factor analysis, principal component analysis, procrustes rotation

Introduction

Component or factor analysis studies that involve the same variables, applied to subjects from different populations or in different experimental conditions, often require factor interpretations to be compared. Multigroup confirmatory factor analysis (CFA) is the best way of testing hypotheses of equivalence of factors. This is done by constraining factors across groups by specifying identical constraints and then specifying remaining pattern loadings to be equal across the groups. This method has serious limitations: First of all, when the sample size is large, any hypothesis of equal factors will systematically be rejected. Moreover, the available software for CFA often fails to converge to a solution (Lorenzo-Seva & Ferrando, 2003).

As some authors have noted (Church & Burke, 1994; Ferrando & Lorenzo, 2000; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996), exploratory factor analysis (EFA) might be more appropriate as a basis for factor comparisons than the CFA approach in most applications, especially in large multidimensional solutions that do not approach very simple structures. The most popular tool for such comparisons was first suggested by Burt (1948) and became popular as Tucker's congruence coefficient (Tucker, 1951). This index is typically computed after one of the factor loading matrices has been transformed to fit another loading matrix in the least squares sense by a Procrustes rotation. This approach of carrying out a Procrustes rotation, followed by evaluation of Tucker's index, is still encountered in methodological journals (see, e.g., Chan, Ho, Leung, Chan, & Yung, 1999; Lorenzo-Seva & Ferrando, 2003; Lorenzo-Seva, Kiers, & ten Berge, 2002), as well as in applied journals (see, e.g., Chico, Tous, Lorenzo-Seva, & Vigil-Colet, 2003; Hendriks, Hofstee, & De Raad, 1999; Rodríguez-Fornells, Lorenzo-Seva, & Andrés-Pueyo, 2001).

The congruence coefficient is the cosine of the angle between the two vectors, and can be interpreted as a standardized measure of proportionality of elements in both vectors. It is evaluated as

$$\phi(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \quad (1)$$

where x_i and y_i are the loadings of variable i on factor x and y, respectively, $i = 1, \ldots, n$. Usually the two vectors are columns of a pattern matrix. However, they could also be columns of a structure matrix. The popularity of the congruence coefficient can be attributed to the following properties:

First, $\varphi(x, y)$ is insensitive to scalar multiplication of x and y. This implies that it measures factor similarity independently of the mean absolute size of the loadings: It can be high when loadings are near zero and vice versa. This is desirable because factor interpretations on the one hand, and the explained variance of factors (reflected in the sum of the squared loadings) on the other, are entirely different and unrelated concepts, which should not be confused. If a researcher is interested in taking the sizes of loadings obtained in two samples into account, the root mean square can be computed (see, e.g., Harman, 1960). However, this index is not often used in applied research mainly because it confounds factor interpretations and the explained variance of factors.

Second, $\varphi(x, y)$ is sensitive to additive constants. This is also desirable because factor interpretations are indeed sensitive to additive constants. For instance, the loadings (.4

DOI 10.1027/1614-1881.2.2.57

.4 0) would give rise to quite a different interpretation than the loadings (.1 .1 –.3); see also Korth (1973, p. 58).

Third, $\varphi(x, y)$ is insensitive to a change in the sign of any pair (x_i, y_i), which reflects a change in the sign of variable i.

Fourth, $\varphi(x, y)$ is mathematically attractive since it is a continuous function of x_i and y_i.

Tucker (1951) made the following comment on the practical use of congruence: "For practical purposes, it may be desirable to set up some value of φ less than unity which will be regarded as acceptable for indicating the identity of the factors in the two studies. However, no guiding line values have yet been developed, and it seems proper to delay specifying any minimally acceptable value of the coefficient of congruence until adequate experience in the application of the method has been gained" (p. 43).

Five decades later, this index is still very popular (see, e.g., McCrae et al., 1996). However, hardly any research seems to have been reported about these guiding line values. Haven and ten Berge (1977) carried out an empirical study, and reported that congruence values above .85 could be seen as indicative of equal factor interpretations. However, other rules of thumb can also be encountered. For instance, Horn, Wanberg, and Appel (1973, pp. 153–154) apply a threshold value of .80 for congruence. That is, they consider factors as identical if the congruence between them is .80 or higher. Other authors, like Mulaik (1972, p. 355), Bentler and Bonett (1980), and Van de Vijver and Leung (1997), adopt the more stringent threshold value of .90. MacCallum, Widaman, Zhang, and Hong (1999) follow the guidelines given by Tucker: .98 to 1.00 = *excellent*, .92 to .98 = *good*, .82 to .92 = *borderline*, .68 to .82 = poor, and below .68 = *terrible*. In fact, the possible subjectivity in the interpretation of the index has already been criticized (Davenport, 1990). The purpose of the present study is to gather explicit empirical evidence on the threshold problem for congruence, by repeating the much-cited Haven and ten Berge study (1977), and extending its design to include different domains of variables.

The literature on congruence also contains studies that give statistical baselines for congruence: for example, Schneewind and Cattell (1970); Nesselroade and Baltes (1970); Nesselroade, Baltes, and Labouvie (1971); Korth (1973, 1978); Korth and Tucker (1975); Cattell (1978); and Bentler and Bonett (1980). These statistical baseline values can be used to decide whether congruence is significantly different from zero given the particular optimizing rotation that has been applied. Chan et al. (1999) presented a study to decide whether congruence is significantly different from one, by means of bootstrap (see also Lorenzo-Seva & Ferrando, 2003). Statistical significance is a necessary condition for taking congruence values seriously. It is by no means sufficient for inferring that factors have the same interpretation (Davenport, 1990; Korth, 1978). Therefore, these studies have no bearing on the question of what level of congruence is necessary or sufficient to infer equality of factor interpretations Other papers study the effect of data characteristics (loading values, size of sample, number of factors, and number of variables per factor) related to the value of the congruence index (Broadbooks & Elmore, 1987; Guadagnoli & Velicer, 1991; Paunonen, 1997). However, these studies again have no bearing on the main problem addressed in the present article, which is to ascertain how subjective assessments of factor similarity are reflected by the numerical values of the congruence index.

As has been said before, the present study aims to extend the study presented by Haven and ten Berge (1977). In their study, 20 pairs of columns obtained from an empirical study were evaluated by 20 judges. The pairs of vectors were obtained from a single applied research and involved 18 variables; congruence between pairs ranged from zero to one. Half of the judges received labels of the 18 variables, while the others evaluated pairs of columns without any information on the nature of the variables involved. We shall extend this study to include more experimental conditions and a larger number of judges. But we narrow down the range of congruence values to be examined to .62–.97, because Haven and ten Berge's study has shown that, if there is a threshold value above which congruence indicates factor identity, it has to be in this range.

Method

We combined columns from real-life loading matrices with artificially constructed columns to obtain pairs of columns with specific congruence values. These pairs were submitted to judges who assessed the subjective similarity within each pair. Below, the procedure is described in detail.

Real-Life Loading Matrices

We selected six loading matrices from applied research in three different fields: two related to personality (Arrindell et al., 2001; Steer, Rissmiller, & Beck, 2000), two related to intelligence (Acton & Schroeder, 2001; Miller & Vernon, 1996), and two related to social psychology (Korf & Malan, 2002; Valk & Karu, 2001). Three of them (one from each field) pertained to 15 variables, and the others (also one from each field) involved 20 variables.

Data Construction

From each of the six real-life loading matrices, one column was selected to be presented to judges, in combination with one out of eight artificially constructed other columns. The latter columns were artificially constructed so that the congruence values between the selected column and the eight artificial columns were .62, .66, .72, .79, .85, .90, .94, and .97, respectively. So, if $\mathbf{a}$ was the selected column, and $\mathbf{c}_i$ was one of the artificial columns, we obtained eight pairs of columns as

$$\mathbf{P}_i = [\mathbf{a}|\mathbf{c}_i]. \quad (2)$$

The whole set $\mathbf{P}_1, \ldots, \mathbf{P}_8$ was submitted to judges who assessed the degree of subjective similarity between columns. The judges received the matrices $\mathbf{P}_i$ in a random order with respect to the congruence values. Each judge

evaluated only one set of eight matrices. We carried out this procedure for each of the six loading matrices obtained from applied research, so we constructed six sets of eight matrices $\mathbf{P}_i$.

Evaluation of Pairs of Columns

As already pointed out, each set of eight $\mathbf{P}_i$ matrices was submitted to judges experienced in factor analysis: They were university professors who apply EFA in their research. Each judge received a set built from a loading matrix related to his or her research field, and some of them also received the corresponding labels of the variables in the selected loading matrix. All judges scored the similarity of the pair of columns on the following five-point scale: 1 = *very poor*, 2 = *poor*, 3 = *fair*, 4 = *good*, and 5 = *very good*.

We obtained responses from 56 judges. The number of judges related to research in personality, intelligence, and social psychology were 28, 21, and 7, respectively. Of these, 32 also received variable labels, while 24 merely received the set of pairs of columns with no information on the nature of the variables. Finally, 36 judges received a set of pairs of columns that involved 20 variables, while 20 judges received a set of pairs of columns that involved 15 variables. As each judge evaluated 8 pairs of columns from the same set, and the number of judges was 56, a total of 448 pairs of columns were evaluated.

Results

Overall Study

Table 1 contains the congruence (φ), score frequencies, total score (T), score averages ($\bar{X}$), and standard deviations (S_x) for the 48 pairs of factors evaluated by the 56 judges. The averages and their corresponding 95% confidence interval are shown in Figure 1.

The results show a strong linear relationship (r = .974) between φ and the subjective scores. So, overall, subjective judgments of similarity were strongly related to real congruence. However, individual subjective scores show that some judges were very strict in their judgments: For example, two judges indicated that the match was *poor*, when actually the real congruence between columns was as high as .97. These two judges were evaluating pairs of columns of 15 variables related to personality research, and one of them received the label of the variables. As a matter of fact, they indicated that the match was either *poor* or *very poor* for all pairs of columns evaluated. Also, five other judges that evaluated pairs of columns (two of 15 labeled variables, and three of 20 labeled variables) related to personality (two judges) and intelligence (three judges) research indicated that match was either *poor* or *very poor* when actually the real congruence between columns was .90. None of these five judges indicated a *very good* match for any pair of columns. The rankings of subjective scores reported by 44 judges corresponded to the ordering of the underlying congruence coefficients. The worst rankings corresponded to two judges who reported rankings in which two scores did not correspond to the underlying ordering. These judges evaluated pairs of columns of 20 variables.

Subjective scores reported by judges showed a wide disparity. However, the average of these scores was well related to factor similarity. This result is important, because it shows that subjective judgments should not be considered in applied research unless a large number of judges evaluate the factor solution and the average of the judgments can be computed.

The main question is how to establish a threshold value for φ above which factors can be considered equal. If a score average of 3, corresponding to *fair* factor similarity, is required in order for factors to have the same interpretations, then the threshold value seems to be near .85. If, however, an average of 4, corresponding to *good* factor similarity, is required in order for factors to be considered equal, then the threshold value seems to be .95.

Number of Variables

As already pointed out, 36 judges gave responses for a set of pairs of columns that involved 20 variables (19, 13, and 4 related to research in personality, intelligence, and social psychology, respectively), while 20 judges gave responses

Table 1. Congruence (φ), score frequencies, total score (T), score averages ($\bar{X}$), and standard deviations (S_x) for 48 pairs of factors evaluated by 56 judges

	Score Frequencies							
φ	1	2	3	4	5	T	$\bar{X}$	S_x
.97	0	2	1	21	32	251	4.48	0.707
.94	0	2	12	30	12	220	3.93	0.753
.90	2	5	25	22	2	185	3.30	0.822
.85	2	13	22	19	0	170	3.04	0.844
.79	7	20	25	4	0	138	2.46	0.801
.72	14	26	15	1	0	115	2.05	0.766
.66	22	25	7	2	0	101	1.80	0.789
.62	26	24	6	0	0	92	1.64	0.666

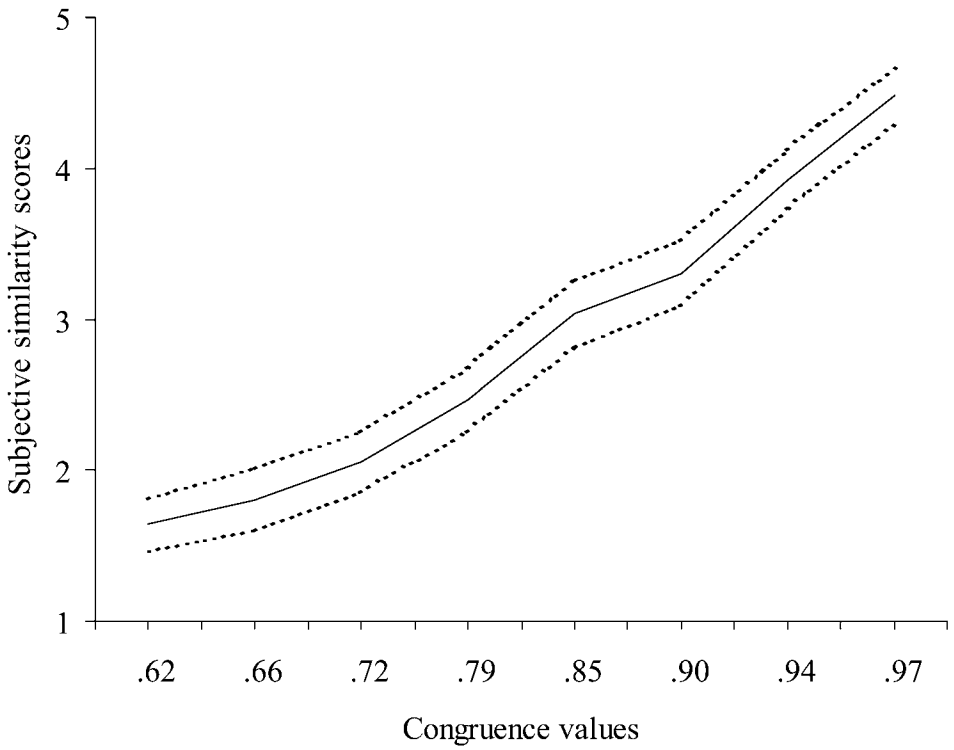

Figure 1. Subjective score averages (solid line) and the corresponding 95% confidence intervals (dotted lines) for 48 pairs

for a set of pairs of columns that involved 15 variables (9, 8, and 3 related to research in personality, intelligence, and social psychology, respectively). It seems interesting to see whether or not the number of variables in the columns affected the subjective perception on the similarity between columns. The averages for each condition are presented in Figure 2.

The results indicate that judges who evaluated pairs of columns with 15 variables gave lower levels of similarity than judges who evaluated pairs with 20 variables, for factors displaying at least a *fair* similarity (congruence equal to or higher than .85). Repeated measured analysis of variance showed that these differences were significant ($F = 2.10$, $P = .043$, power $= .803$).

Labels of Variables

A total of 32 judges gave responses for the set of pairs of columns with variable labels (13, 13, and 6 related to research in personality, intelligence, and social psychology, respectively), while 24 judges only had the set of pairs of columns with no information on the nature of the variables (15, 8, and 1 related to research in personality, intelligence, and social psychology, respectively). It was presumed that the judges who did have the variable labels would be less sensitive to fluctuations in individual loadings because of the additional information, and that they would assign higher scores than the other judges. The averages for each condition are presented in Figure 3.

As Figure 3 shows, the judges did not assign higher scores when they had the labels as additional information. In fact, in comparison to judges who had no additional information, they revealed a slightly smaller similarity when the congruence was between .90 and .94. Actually, repeated measured analysis of variance showed that these differences were not significant ($F = 0.976$, $P = .448$).

Research Field

Finally, it may be interesting to assess whether the judges from different research fields responded differently. In our study, 13 judges in personality and 13 judges in intelligence received a set of pairs of columns with variable labels. The score averages given by researchers in the personality and intelligence fields are presented in Figure 4.

As Figure 4 shows, there was no difference when congruence was high. However, when congruence was low, those judges involved in intelligence research revealed higher factor similarity than those in personality research. Repeated measured analysis of variance showed that these differences were significant ($F = 2.82$, $P = .007$, power $= .918$). Apparently, personality factors are more easily perceived as different than intelligence factors. This may be due to the fact that all intelligence tests share "general intelligence" to some extent.

Discussion

Our results showed a strong linear relationship ($r = .974$) between φ and the subjective similarity scores. This im-

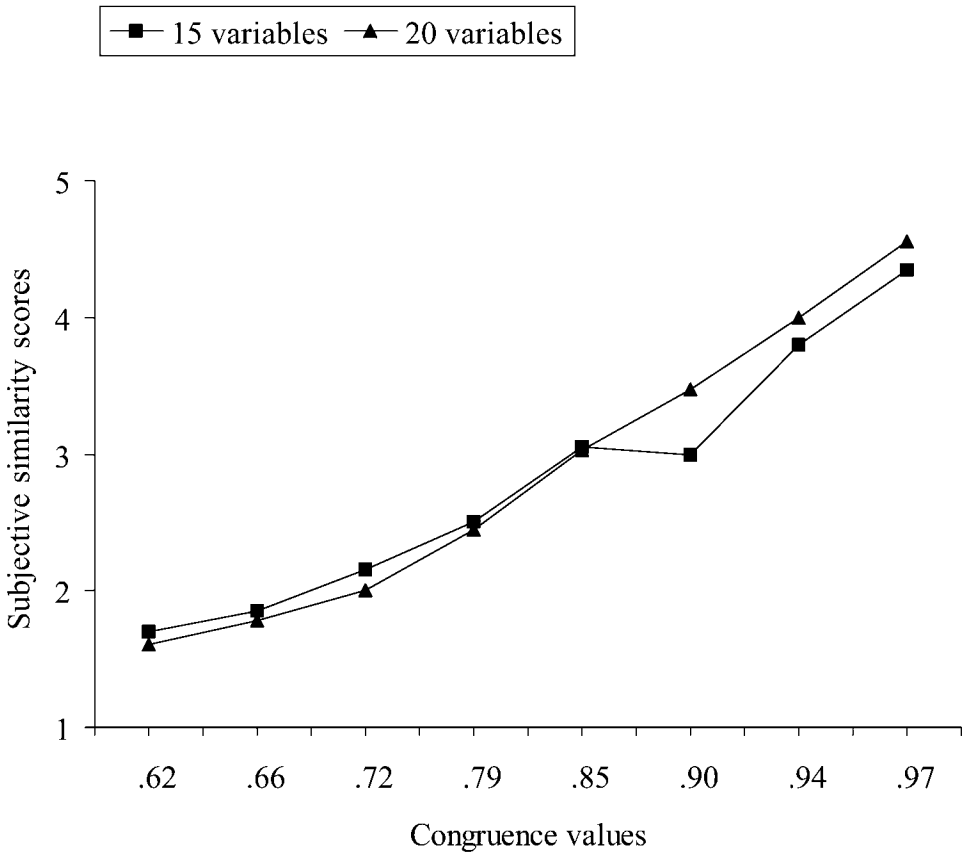

Figure 2. Subjective score averages for 48 pairs of factors evaluated by 56 judges. Squares are for columns of 15 variables,

plies that φ is indeed an excellent measure of factor similarity. However, even if the average of the subjective scores reported by judges was well related to factor similarity, these scores showed a wide disparity. This is important, because it shows that rules of thumb discussed in the introduction are quite hazardous in applied research: They are based on the subjective judgment of just a single researcher. Note that, while none of these rules are based in any kind of empirical study, our study obtains a threshold from the average of 56 researchers.

The main goal of this study was to find a critical congruence level on the basis of judgments of factor similarity by practitioners of factor analysis. Two threshold points can be established:

1. A value in the range .85–.94 means that the two factors compared display *fair* similarity. This result should prevent congruence below .85 from being interpreted as indicative of any factor similarity at all.
2. A value higher than .95 means that the two factors or components compared can be considered equal. That is what we have called a *good* similarity in our study.

In addition, subjective judgments appeared to be sensitive to particular characteristics of the data evaluated: The number of variables and the judge's knowledge of labels of variables did affect his or her judgments. When there were only 15 variables, and when variable labels were given, the judges reported lower congruence. These results are interesting in two ways. First, an effect was expected to show up, but in the opposite direction: When judges have fewer loadings to compare or when they have substantive additional information, they were in fact expected to report higher subjective similarity. Second, this effect appeared only when factor similarity was *fair* or *good* (i.e., congruence between columns was equal to or above .85). It seems that under these particular circumstances subjective judgments become more conservative.

Finally, we studied whether researchers' theoretical background affected the subjective similarity reported by judges. Specifically, we found that those judges involved in intelligence research revealed higher factor similarity than those in personality research. This effect might be explained by considering the theoretical assumptions in intelligence research: As all intelligence tests share general intelligence to some extent, they are more easily perceived as similar. This means that subjective interpretation of factor similarity in applied research, where substantive additional information is always available, could be systematically affected by the theoretical background related to the particular research field at hand.

Acknowledgments

This research was partially supported by a grant from the Spanish Ministry of Science and Technology (SEJ2005-09170-C04-04/PSIC).

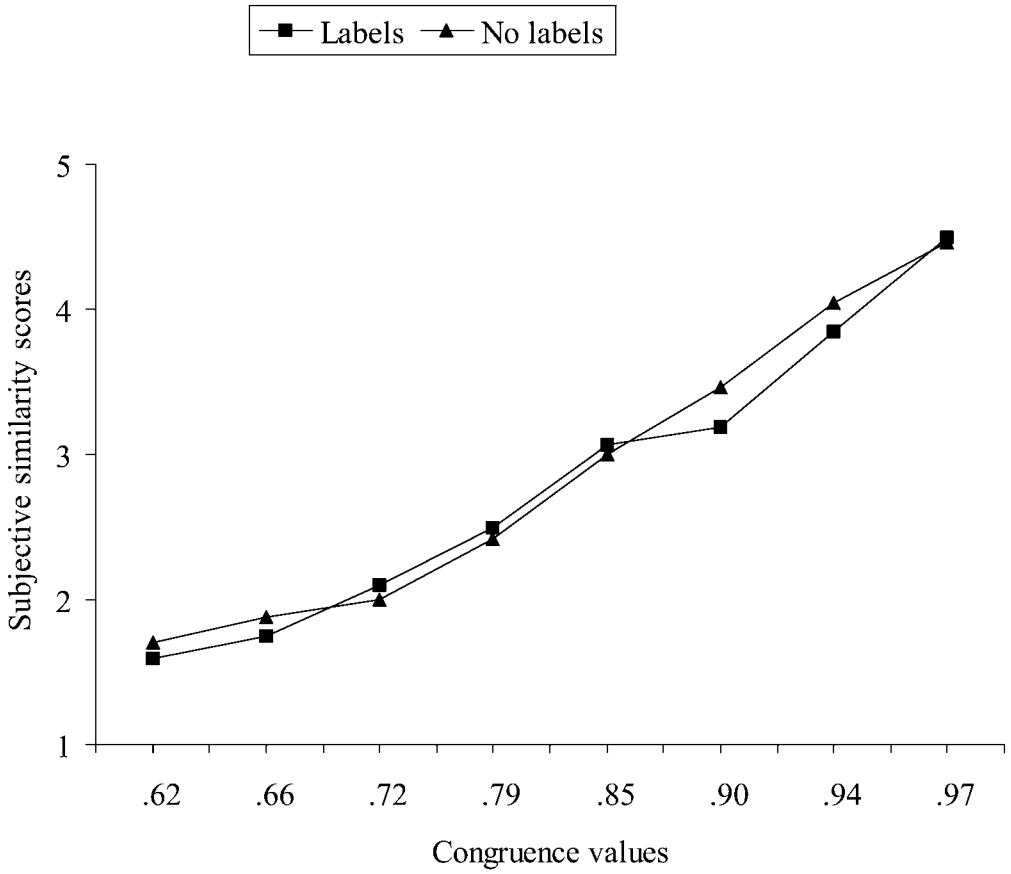

Figure 3. Subjective score averages for 48 pairs of factors evaluated by 56 judges. Squares are for columns provided with

References

Acton, G. S., & Schroeder, D. H. (2001). Sensory discrimination as related to general intelligence. *Intelligence, 29,* 263–271.

Arrindell, W. A., Bridges, K. R., Van der Ende, J., Lawrence, J. S., Gray, S. L., Harnish, R., Rogers, R., & Sanderman, R. (2001). Normative studies with the Scale for Interpersonal Behaviour (SIB): II. US students. A cross-cultural comparison with Dutch data. *Behaviour Research & Therapy, 39,* 1461–1479.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analyses of covariance structures. *Psychological Bulletin, 88,* 588–606.

Broadbooks, W. J., & Elmore, P. B. (1987). A Monte Carlo study of the sampling distribution of the congruence coefficient. *Educational & Psychological Measurement, 47,* 1–11.

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

Cattell, R. B. (1978). Matched determines vs. factor invariance: A reply to Korth. *Multivariate Behavioral Research, 13,* 431–448.

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

Chico, E., Tous, J. M., Lorenzo-Seva, U., & Vigil-Colet, A. (2003). Spanish adaptation of Dickman's impulsivity inventory: Its relationship to Eysenck's personality questionnaire. *Personality & Individual Differences, 35,* 1883–1892.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality & Social Psychology, 66,* 93–114.

Davenport, E. C. (1990). Significance testing of congruence coefficients: A good idea? *Educational & Psychological Measurement, 50,* 289–296.

Ferrando, P. J., & Lorenzo, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica, 21,* 301–323.

Guadagnoli, E., & Velicer, W. F. (1991). A comparison of pattern matching indices. *Multivariate Behavioral Research, 26,* 323–343.

Harman, H. H. (1960). *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press.

Haven, S., & ten Berge, J. M. F. (1977). Tucker's coefficient of congruence as a measure of factorial invariance: An empirical study. *Heymans Bulletin* 290 EX, unpublished report by the Department of Psychology, University of Groningen.

Hendriks, A. A. J., Hofstee, W. K. B., & De Raad, B. (1999). The Five-Factor Inventory (FFPI). *Personality & Individual Differences, 27,* 307–325.

Horn, J. L., Wanberg, K. W., & Appel, M. (1973). On the internal structure of the MMPI. *Multivariate Behavioral Research, 8,* 131–172.

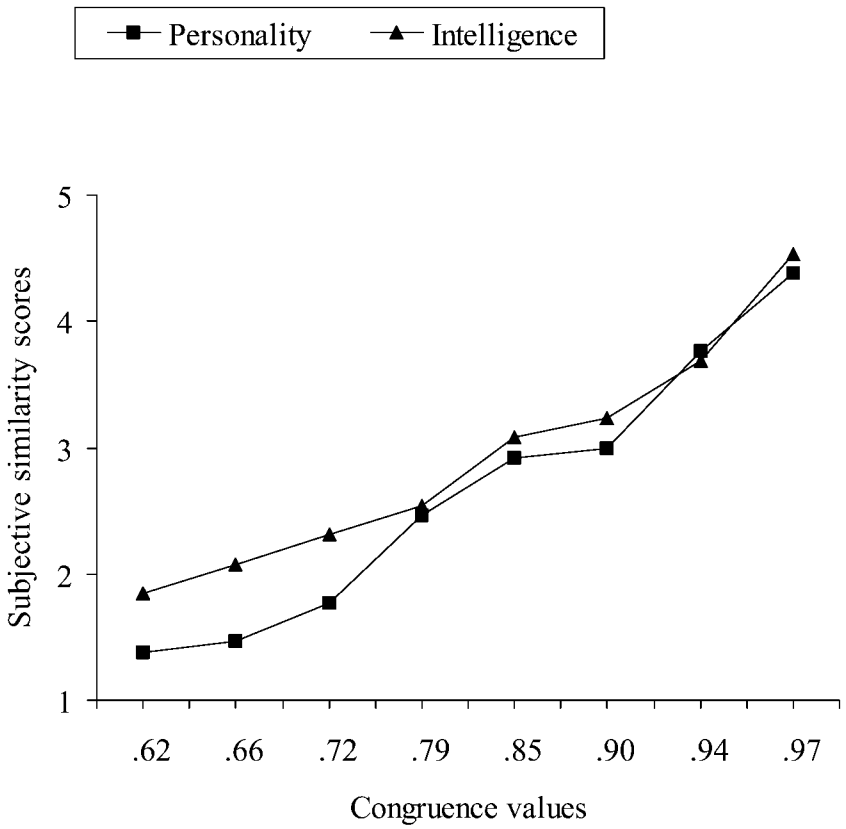

Figure 4. Subjective score averages for 32 pairs of factors evaluated by 26 judges. Squares are for columns related to

Korf, L., & Malan, J. (2002). Threat to ethnic identity: The experience of white Afrikaans speaking participants in postapartheid South Africa. *Journal of Social Psychology, 142,* 149–169.

Korth, B. A. (1973). Analytic and experimental analysis of factor matching methods. *Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign*.

Korth, B. A. (1978). A significance test for congruence coefficients for Cattell's factors matched by scanning. *Multivariate Behavioral Research, 13,* 419–430.

Korth, B. A., & Tucker, L. R. (1975). The distribution of chance congruence coefficients from simulated data. *Psychometrika, 40,* 361–372.

Lorenzo-Seva, U., & Ferrando, P. J. (2003). IMINCE: An unrestricted factor-analysis-based program for assessing measurement invariance. *Behavior Research Methods, Instruments, & Computers, 35,* 318–321.

Lorenzo-Seva, U., Kiers, H. A. L., & ten Berge, J. M. F. (2002). Techniques for oblique factor rotation of two or more loading matrices to a mixture of simple structure and optimal agreement. *British Journal of Mathematical & Statistical Psychology, 55,* 337–360.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the revised NEO personality inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality & Social Psychology, 70,* 552–566.

Miller, L. T., & Vernon, P. A. (1996). Intelligence, reaction time, and working memory in 4 to 6 year old children. *Intelligence, 22,* 155–190.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Nesselroade, J. R., & Baltes, P. W. (1970). On a dilemma of comparative factor analysis: A study of factor matching based on random data. *Educational & Psychological Measurement, 30,* 935–948.

Nesselroade, J. R., Baltes, P. W., & Labouvie, E. W. (1971). Evaluating factor invariance in oblique space: Baseline data generated from random numbers. *Multivariate Behavioral Research, 6,* 233–241.

Paunonen, S. V. (1997). On chance and factor congruence following orthogonal Procrustes rotation. *Educational & Psychological Measurement, 57,* 33–59.

Rodríguez-Fornells, A., Lorenzo-Seva, U., & Andrés-Pueyo, A. (2001). Psychometric properties of the Spanish adaptation of the Five Factor Personality Inventory. *European Journal of Psychological Assessment, 17,* 145–153.

Schneewind, K. A., & Cattell, R. B. (1970). Zum problem der Faktorenidentifikation: Verteilungen und Vertrauensintervalle von Kongruenzkoefficienten für Persönlichkeitsfaktoren im Bereich objektiv-analytischer Tests. [On the problem of the factor identification: Distributions and confidence intervals of congruence coefficients for personality factors in the area of objective analytic tests.] *Psychologische Beiträge, 12,* 214–226.

Steer, R. A., Rissmiller, D. J., & Beck, A. T. (2000). Use of Beck Depression Inventory II with depressed geriatric inpatients. *Behaviour Research & Therapy, 38,* 311–318.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Valk, A., & Karu, K. (2001). Ethnic attitudes in relation to ethnic pride and ethnic differentiation. *Journal of Social Psychology, 141,* 583–601.

Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London: Sage.

Urbano Lorenzo-Seva

Universitat Rovira i Virgili
Facultat de Psicologia
Carretera Valls s/n
43007 Tarragona
Spain
E-mail uls@urv.net

