Tableplot

A New Tool for Assessing Precise Predictions

Ernest Kwan[1], Irene R.R. Lu[1], and Michael Friendly[2]

[1]Carleton University, Ottawa, Canada, [2]York University, Toronto, Canada

Abstract. In the debate over null hypothesis significance testing, Paul Meehl strongly advocated appraising theories through the generation and evaluation of precise predictions (e.g., Meehl, 1978). The study of personality structure through the five-factor model (FFM; McCrae & John, 1992) is an important area of research where one encounters many precise predictions. Extant methods of assessing such predictions, however, do not allow researchers to examine the outcome of the predictions in great detail. That is, it may be difficult to determine how estimates fail to match predicted values. As Meehl argued, one must examine how a theory fails to predict in order to refine and improve the theory. To promote better theory appraisal in FFM research, we present a powerful new tool, called a *tableplot* (Kwan, 2008a), that can summarize and clarify factor-analytic results. Specifically, we illustrate how the tableplot enables detailed appraisal of precise predictions in the FFM.

Keywords: precise predictions, theory appraisal, five-factor model, graphical display, tableplot

In the debate over null hypothesis significance testing (NHST) Paul Meehl strongly criticized psychology's use of NHST for theory testing (e.g., Meehl, 1967, 1978, 1986, 1990, 1997). Meehl's main concern is that NHST does not require researchers to carefully generate or examine precise predictions derived from a substantive theory of interest. Meehl argued that unless one examines how empirical estimates fail to match a theory's precise predictions, one can do little to refine or improve the theory.

Despite Meehl's emphasis on precise predictions, he also noted that substantive theories in many areas of psychology cannot generate very precise predictions (e.g., Meehl, 1978). We consider an important exception: personality research through the use of factor-analytic techniques. Particularly, we refer to the study of personality structure through the five-factor model (FFM; McCrae & John, 1992) and the associated five-factor theory (FFT; McCrae & Costa, 1996, 1999). In the FFM there are many parameters with corresponding predicted values (e.g., Rolland, 2002). However, researchers do not adequately examine the extent to which these estimates match their predicted values (e.g., McCrae, Zonderman, Costa, Bond, & Paunonen, 1996). According to Meehl, this neglect is detrimental because it limits the advancement and refinement of FFT.

One clear reason behind the unsatisfactory assessment of precise predictions in FFM research is the large number of parameter estimates involved. The FFM has 150 parameters (e.g., Costa & McCrae, 1992). Extant methods of assessment include summary indices of fit and significance tests of such indices (e.g., McCrae et al., 1996; also see the next section of this paper). If one wishes to examine how parameter estimates deviate from predicted values, one must rely on a table of predicted and estimated values (see Table 1). As one can verify from Table 1, using a table to compare 300 values can be a challenge, even when supplemented by summary statistics. Indeed, several researchers have discussed the inadequacies of employing tables to interpret quantitative information (e.g., Friendly & Kwan, 2003; Wainer, 1997). Thus, even if one intends to carefully examine the outcome of precise predictions in the FFM, one may still find it difficult to do so.

The goal of this paper is to promote better assessment of precise predictions in FFM research. Recently Kwan (2008a) proposed a new graphical display, called a *tableplot*, for presenting factor analysis results. We illustrate how the tableplot can be used to help assess precise predictions. First we examine research on the FFM: We discuss the role of precise predictions, extant methods for assessing such predictions, and the limitations of these methods. We then introduce the tableplot and illustrate how the tableplot facilitates better assessment of precise predictions in the FFM. We provide a discussion and conclusion in the last section.

Precise Predictions in Studies of Personality

The FFM is commonly operationalized by the Revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992). The NEO PI-R comprises 240 items that measure 30 facets. These facets relate to five domains (factors) of personality: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Factor analysis (FA) estimates the relationship between a facet and a domain; these

DOI 10.1027/0044-3409.217.1.38

Table 1. Comparing the factor patterns of NEO PI-R of two samples: Normative and Shona

	Normative sample					Shona sample					
	Factors					Factors					
Facets	N	E	O	A	C	N	E	O	A	C	ϕ
N1	**81**	2	–1	–1	–10	**66**	–4	–4	1	–14	99
N2	**63**	–3	1	**–48**	–8	**53**	–18	6	–39	–16	96
N3	**80**	–10	2	–3	–26	**60**	–17	15	1	**–40**	94
N4	**73**	–18	–9	4	–16	**58**	–11	13	13	–19	94
N5	**49**	35	2	–21	–32	**58**	20	–3	–34	–36	96
N6	**70**	–15	–9	4	–38	**57**	5	–26	3	**–46**	93
E1	–12	**66**	18	38	13	–24	**61**	3	39	21	96
E2	–18	**66**	4	7	–3	–14	**59**	–1	39	–7	89
E3	–32	**44**	23	–32	32	**–51**	22	27	–26	14	88
E4	4	**54**	16	–27	**42**	–15	35	5	1	**42**	85
E5	0	**58**	11	–38	–6	–2	35	**52**	–4	–30	59
E6	–4	**74**	19	10	10	4	**65**	30	25	9	96
O1	18	18	**58**	–14	–31	27	25	28	**–54**	–7	69
O2	14	4	**73**	17	14	1	**40**	**53**	10	16	83
O3	37	**41**	**50**	–1	12	**47**	32	22	–8	21	90
O4	–19	22	**57**	4	–4	–3	15	38	9	14	89
O5	–15	–1	**75**	–9	16	–22	–7	**73**	4	21	98
O6	–13	8	**49**	–7	–15	–18	0	**59**	18	–2	87
A1	–35	22	15	**56**	3	**–46**	22	7	**46**	–2	97
A2	–3	–15	–11	**68**	24	–13	–4	–10	**61**	38	96
A3	–6	**52**	–5	**55**	27	–34	28	25	26	39	72
A4	–16	–8	0	**77**	1	–12	20	–27	**61**	31	78
A5	19	–12	–18	**59**	–8	10	–32	12	**61**	9	82
A6	4	27	13	**62**	0	10	24	25	**54**	**42**	82
C1	**–41**	17	13	3	**64**	–29	22	10	–9	**57**	97
C2	–4	6	–19	1	**70**	–12	3	10	27	**72**	86
C3	–20	–4	1	29	**68**	–25	13	6	**44**	**65**	96
C4	–9	23	15	–13	**74**	–6	14	10	31	**72**	84
C5	–33	17	–8	6	**75**	–29	10	–4	30	**67**	95
C6	–23	–28	–4	22	**57**	–29	–2	15	34	**63**	89
ϕ						94	89	80	83	93	89

Note. The Shona factor pattern (Piedmont et al., 2002) was rotated to match the normative pattern (Costa & McCrae, 1992) by Procrustes rotation. Loadings ≥ 0.40 in absolute value are in **boldface**. Values are in the hundredth decimal.

150 relationships are summarized as a 30 by 5 matrix where the regression coefficient of facet *i* on domain *j* appears in cell (*i*, *j*). In FA terminology this matrix is the *factor pattern*, *loadings* matrix, or Λ, each regression coefficient is a *pattern coefficient*, *loading*, or λ (for simplicity, we omit hat notation to denote estimates).

A critical prediction of FFT is that the FFM can be a model of personality across cultures (e.g., Allik & McCrae, 2002). In other words, the NEO PI-R factor pattern based on the normative sample (Costa & McCrae, 1992) should be replicable in any culture. This is a precise prediction and many studies over the past decade and a half have provided evidence to assess this prediction (e.g., Rolland, 2002).

The popular approach for appraising the cross-cultural replicability of the FFM is to develop and evaluate versions of the NEO PI-R in different cultures (e.g., Rolland, 2002). Given the initially estimated factor pattern from a particular culture, researchers transform this estimate to match the normative factor pattern as closely as possible (e.g., McCrae et al., 1996). This transformation entails a *Procrustes rotation*, which minimizes the squared deviations between the normative factor pattern and the factor pattern of the other culture (e.g., Browne, 1972). For example, Table 1 contains the normative factor pattern and the Procrustes-rotated factor pattern from Piedmont, Bain, McCrae, and Costa's (2002) study of the Shona (a native language of Zimbabwe) NEO PI-R. Researchers then assess the degree of similarity of the rotated factor pattern to the normative factor pattern; the greater the similarity, the stronger the evidence to support FFT's precise prediction.

Extant Methods to Assess the Similarity of Two Factor Patterns

There are several popular methods to assess the similarity between two factor patterns. One could use a *congruence coefficient* (Burt, 1948; Tucker, 1951; Wrigley & Neuhaus, 1955), which ranges from –1 to 1, inclusively. We denote this index as ϕ. ϕ = 1 for two sets of identical numbers[1]; ϕ = –1 for two sets of identical numbers with opposite signs. For example, there are 30 loadings for Neuroticism in the normative factor pattern and 30 corresponding loadings in the Shona factor pattern; the ϕ for Neuroticism is calculated from these 30 pairs of loadings and it is equal to 0.94. One could calculate ϕ for each factor to summarize the degree of factor-similarity between the normative and Shona factor patterns. We include these five ϕs in the row below the Shona factor pattern in Table 1.

McCrae et al. (1996) proposed applying ϕ to also measure *variable* and *total* similarity between two factor patterns. For example, one calculates ϕ from the five pairs of loadings for every facet (variable) in Table 1; these 30 facet ϕs appear in the column to the right of the Shona factor pattern. One also calculates ϕ for all 150 pairs of loadings; this total ϕ appears in the bottom right cell in Table 1.

Because the sampling distribution of ϕ is unknown, researchers have relied on some rules of thumb to interpret the magnitude of ϕ (e.g., Lorenzo-Seva & ten Berge, 2006). For example, if we impose 0.90 as a cut-off for acceptable similarity, then Extraversion, Openness, and Agreeableness would be considered poorly replicated in the Shona factor pattern (Table 1).

The unknown distributional properties of ϕ are problematic because one cannot derive a formal significance test for ϕ (e.g., McCrae et al., 1996; Chan, Ho, Leung, Chan, & Yung, 1999). Paunonen, Jackson, Trzebinski, and Forsterling (1992) proposed one approach to approximate a significance test: One uses computer simulation to calculate ϕs of randomly matched factors. A large collection of such ϕs can be interpreted as a sampling distribution of ϕ under a null hypothesis of zero factor similarity. If an observed ϕ (e.g., 0.94 of Neuroticism in Table 1) falls in an extreme region of this sampling distribution (e.g., tail 5%), one regards this as evidence that the true ϕ is not zero. McCrae et al. (1996) generalized this method of testing factor similarity to assess facet and total similarity. Chan et al. (1999) also proposed a method for testing if ϕ could be regarded as statistically different from 1. We refer to all such procedures as "ϕ significance tests."

Although ϕ significance tests do provide a more formal method for assessing similarity between factor patterns, such tests have a few limitations. Of course, some criticisms of NHST also apply to ϕ significance tests. These include, for example, the arbitrariness in the critical level of significance used to reject hypotheses (e.g., Rosnow & Rosenthal, 1989) and the appropriateness of using a decision-making procedure for scientific inference (e.g., Rozeboom, 1960). Another limitation is that ϕ significance tests (and ϕs) do not reveal the nature of dissimilarity. Indeed, recent conjectures have led to predictions with regards to the relative success of replicating subsets of loadings in the NEO PI-R (e.g., McCrae, 2001; Piedmont et al., 2002). The evaluation of these further predictions could lead to refinements of FFT. ϕ and its significance test are functions of loadings, but it would be difficult for a researcher to use these aggregates to explicitly evaluate the extent to which each loading misses its predicted value (e.g., McCrae et al., 1996).

The limitations of ϕ and ϕ significance tests stand out even more if one intends to compare how several cultures replicate the normative factor pattern. Table 2 presents the rotated factor patterns from a Portuguese (Lima, 2002) and a Marathi (Lodhi, Deo, & Belhekar, 2002) version of the NEO PI-R, along with factor, facet, and total ϕs. For example, Openness is consistently the worst replicated factor in every culture based on the factor ϕs (see Tables 1 and 2). Is the nature of misfit similar across cultures or, do different cultures differ from the normative loadings in specific ways? In fact, the Shona Openness has the lowest ϕ amongst all factors and cultures; what particular loadings of the Shona Openness contribute to this substantial misfit? What is unique about Shona compared to the other two cultures with regards to the Openness loadings? The answers to these questions would offer a rich basis of further research; but ϕs or their significance tests are not helpful at providing such answers.

Beyond ϕs and their significance tests, researchers also rely on visual inspection of tables as a method of assessing similarity of factor patterns. As we already noted, tables are far from ideal for understanding quantitative information. Researchers commonly use boldface to highlight those |loadings| ≥ 0.40 to improve their tables (e.g., Floyd & Widaman, 1995). We adopted this convention in Tables 1 and 2. Although this enhancement distinguishes loadings into "high" vs. "low," it does very little to show the actual magnitude of the loadings (e.g., how much higher is one boldfaced value over another?). Once we increase the number of cultures to compare, the inadequacies of tables quickly escalate. For example, how

[1] If two sets of numbers differ by a positive multiplicative constant, ϕ is also 1 (ϕ is –1 for a negative multiplicative constant). A more stringent index of similarity is an *identity coefficient*, which is equal to 1 if ,and only if, the two sets of numbers are identical (e.g., see van de Vijver & Leung, 1997). We use ϕ in our illustrations because it is more commonly used in factor-analytic research (e.g., Lorenzo-Seva & ten Berge, 2006).

Table 2. The NEO PI-R factor pattern from a Portuguese and Marathi sample

	Portuguese sample						Marathi sample					
Facets	Factors						Factors					
	N	E	O	A	C	ϕ	N	E	O	A	C	ϕ
N1	**77**	5	–4	17	14	93	**81**	1	–1	15	–3	98
N2	**65**	–17	–10	–34	–1	95	**64**	–14	–4	**–41**	–2	98
N3	**78**	–11	–18	6	–17	96	**71**	–12	–6	7	–36	98
N4	**71**	–8	–8	6	–19	99	**69**	–8	–12	–1	–26	98
N5	39	33	26	–21	–34	93	**45**	21	–2	–28	–39	97
N6	**65**	–10	–14	–3	**–42**	99	**72**	1	–7	10	–39	98
E1	–10	**65**	9	**41**	20	99	–11	**73**	6	29	5	97
E2	–15	**66**	3	7	–10	99	0	**69**	–15	30	–13	88
E3	–32	34	17	**–47**	17	95	–37	37	27	**–42**	22	97
E4	6	33	20	–24	34	98	–18	29	10	–32	**41**	89
E5	–2	**60**	25	–24	–3	96	2	**41**	19	–34	–21	94
E6	–11	**65**	35	–2	4	95	7	**62**	34	18	19	95
O1	17	34	**61**	–7	–15	95	34	11	**50**	–5	–14	93
O2	12	21	**64**	15	16	97	26	19	**62**	29	17	95
O3	17	**47**	**53**	–10	17	95	20	35	**56**	3	25	95
O4	–16	11	**55**	4	–7	98	–31	11	**43**	–5	–19	90
O5	–5	22	**69**	–17	8	93	–22	19	**48**	–17	35	86
O6	–7	5	**71**	8	–8	94	–35	–29	**55**	2	–20	80
A1	–21	34	–8	**54**	1	91	–28	15	6	**51**	–1	99
A2	–2	–26	3	**70**	7	95	15	–24	15	**61**	15	89
A3	–3	**42**	6	**60**	28	98	–15	33	24	**51**	28	89
A4	–11	–3	–21	**71**	8	95	–18	1	–7	**70**	–6	98
A5	13	–28	11	**66**	5	87	–3	–26	–10	38	–6	87
A6	16	12	20	**55**	13	93	29	22	–21	**50**	26	73
C1	–26	28	5	6	**63**	97	–34	21	17	–2	**70**	99
C2	7	3	2	2	**69**	94	–1	–4	–15	22	**70**	95
C3	–3	4	2	**40**	**71**	96	–16	–1	11	30	**67**	99
C4	–6	25	11	–4	**74**	99	–16	16	2	6	**74**	95
C5	–31	3	10	12	**73**	96	–34	8	–3	4	**73**	99
C6	–22	–14	–27	26	**56**	93	–12	–5	3	19	**69**	92
ϕ	98	94	90	97	97	95	95	91	89	95	96	94

Note. The Portuguese (Lima, 2002) and Marathi (Lodhi et al., 2002) factor patterns were rotated to match the normative pattern (Costa & McCrae, 1992) by Procrustes rotation. Loadings ≥ 0.40 in absolute value are in **boldface**. Values are in the hundredth decimal.

much effort is needed to determine from Tables 1 and 2 how the Shona Openness loadings differ from those of the other three cultures?

Recent studies have accumulated a vast amount of information on the performance of the NEO PI-R in many cultures (e.g., Rolland, 2002). Indeed, the precise prediction at the core of FFT entails that researchers compare the resulting factor patterns in terms of how they resemble or differ from the normative factor pattern (e.g., Allik & McCrae, 2002). Beyond the extant procedures for comparing factor patterns (i.e., ϕ, ϕ significance tests, inspection of tables), a new tool is needed to help answer the many interesting questions that could arise out of this research.

Introduction to the Tableplot

In an effort to improve applications of FA, Kwan (2008a) developed the tableplot as a tool for representing FA results. A tableplot is a graphical display that supplements each cell of a table with a symbol proportionate to the value in the cell (e.g., Figure 1 is a tableplot of the normative factor pattern in Table 1; we transpose the tableplots in this paper to facilitate captioning beneath each plot). We illustrate in the next section that a tableplot can be a powerful tool for appraising FFT. That is, one can use a tableplot for detailed diagnosis of the fit between predicted and estimated factor patterns. We first

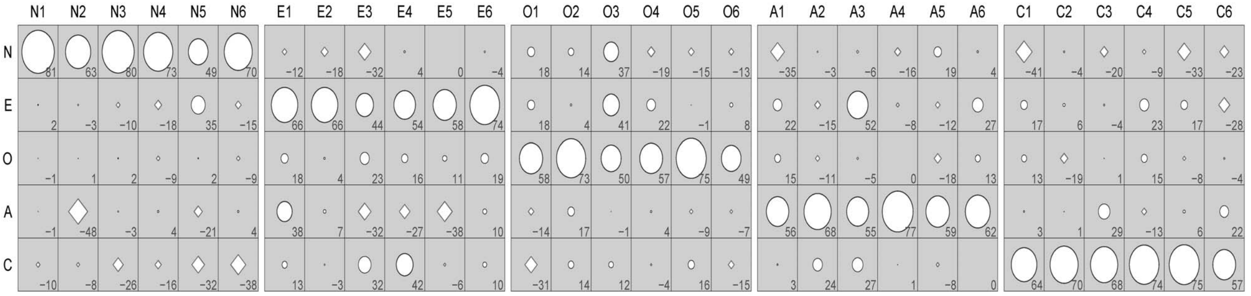

Figure 1. Tableplot of the NEO PI-R factor pattern based on the normative sample (Costa & McCrae, 1992). Symbols scaled to maximum of 1; cell labels in hundredth decimal.

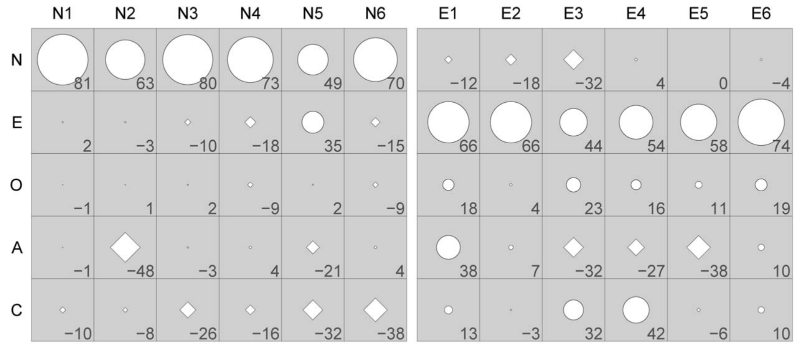

Figure 2. Tableplot of the first 12 facets from the normative NEO PI-R factor pattern (Costa & McCrae,1992). Symbols scaled to maximum of 1; cell labels in hundredth decimal.

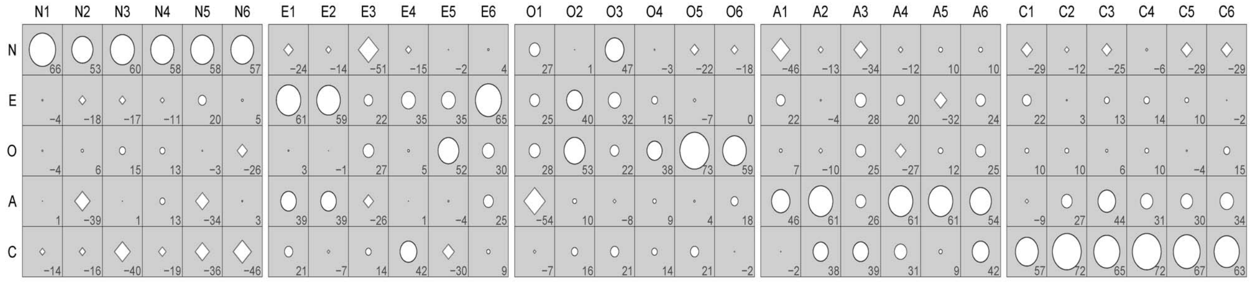

Figure 3. Tableplot of the NEO PI-R factor pattern based on a Shona sample (Piedmont et al., 2002). Symbols scaled to maximum of 1; cell labels in hundredth decimal.

explain how the tableplot is drawn to clarify its interpretation.

To describe the tableplot's construction, we refer to a simpler tableplot of only the first 12 facets in the normative factor pattern (Figure 2). Let the loading in cell (i, j) be denoted λ_{ij}. Note that each cell of a tableplot is a square. If the width (height) of this square is w, the diameter of the circle in row i column j is $|\lambda_{ij}|$ - $m \times w$, where m is the largest possible loading (in absolute value). Because the factors in the NEO PI-R are extracted by principal components (Costa & McCrae, 1992), m = 1. Thus, the circle of cell (i, j) has diameter $|\lambda_{ij}| \times w$ and the largest possible circle occurs if |loading| = 1.

One could use a different plot symbol/color to distinguish negative cell values. We use a diamond with a red outline, in contrast to a circle with blue outline for positive values. All plots in this paper follow this general color scheme (Figures 1, 2, 3, and 8 only appear in grey scale; the color versions can be found at *http://sprott.carleton.ca/ekwan*). We scaled the diagonals of a diamond the same way as the diameter of a circle (e.g., in Figure 2 the diamond in cell A-E3 has a diagonal equal to the diameter of the circle in cell C-E3); thus, for tableplots of the NEO PI-R the largest possible diamond occurs if |loading| = 1.

Because of the size of the NEO PI-R factor pattern, we omit leading zeros and decimal points in cell labels of our tableplots to reduce visual clutter; units in Figure 2 represent the hundredth decimal. Figure 2 also illustrates the use of a gap between columns to visually distinguish groups of facets. We review further graphical considerations in the Discussion.

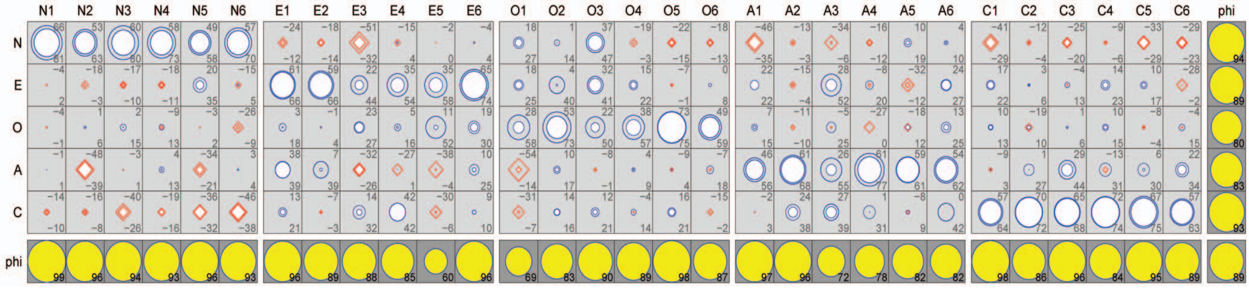

Figure 4. Superimposed tableplot of the normative (McCrae & Costa, 1992) and Shona (Piedmont et al., 2002) NEO PI-R factor patterns, augmented by congruence coefficients (phi) from Table 1. Symbols scaled to maximum of 1; cell labels in hundredth decimal.

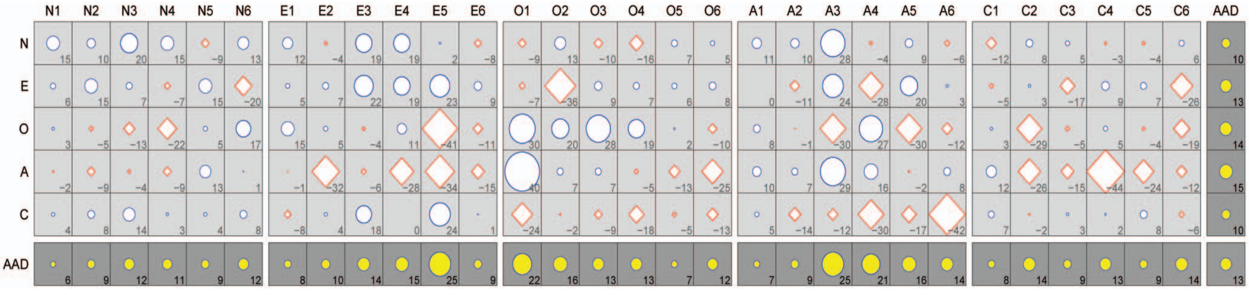

Figure 5. Shona residual tableplot, scaled to observed maximum and augmented by average absolute differences (AAD). Residuals defined as the difference obtained by subtracting the Shona factor pattern (Piedmont et al., 2002) from the normative factor pattern (Costa & McCrae, 1992). Symbols scaled to maximum of 0.44; cell labels in hundredth decimal.

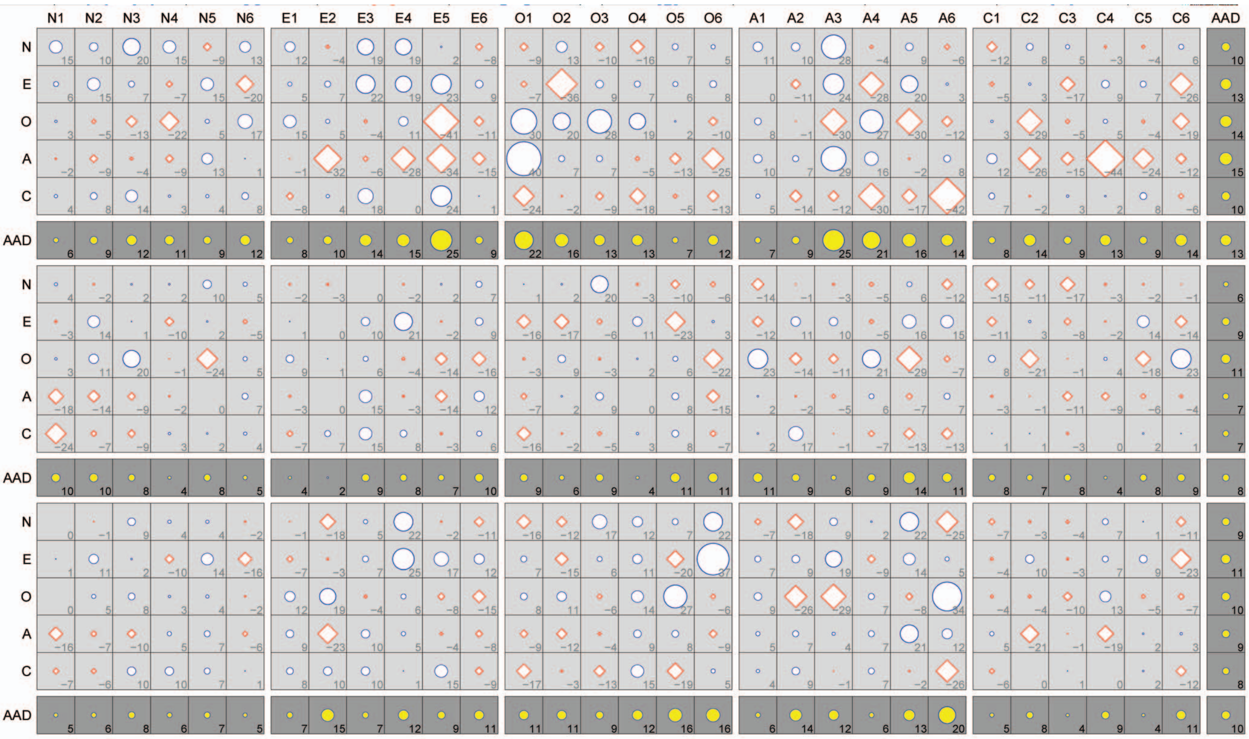

Figure 6. Comparing residuals of the Shona (Piedmont et al., 2002), Portuguese (Lima, 2002), and Marathi (Lodhi et al., 2002) factor patterns. The top, middle, and bottom tableplots are the Shona, Portuguese, and Marathi residual tableplots, respectively. The Portuguese and Marathi residuals are analogously defined as the Shona residuals. Symbols are scaled to observed maximum of 0.44; cell labels in hundredth decimal.

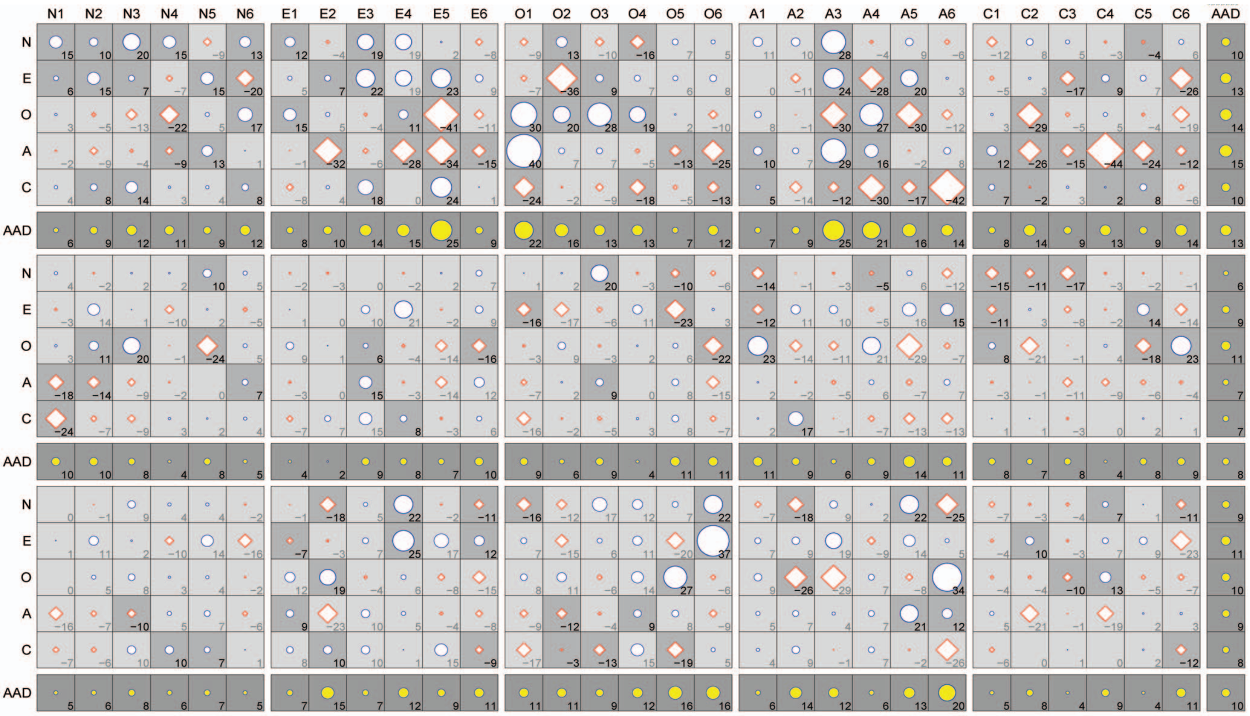

Figure 7. Emphasizing the worst residuals. Each cell of the NEO PI-R factor pattern appears three times; the one that contains the uniquely largest residual has a darkened cell background and label. The top, middle, and bottom tableplots are the Shona (Piedmont et al., 2002), Portuguese (Lima, 2002), and Marathi (Lodhi et al., 2002) residual tableplots, respectively. Symbols scaled to observed maximum of 0.44; cell labels in hundredth decimal.

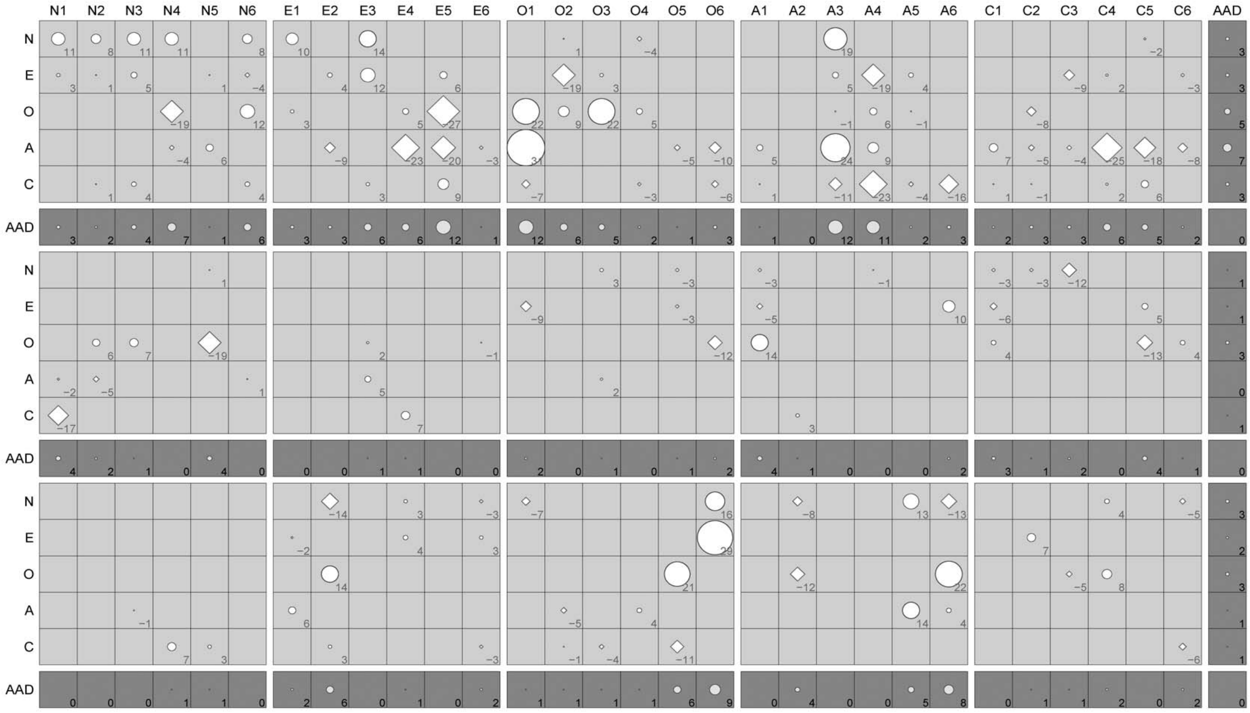

Figure 8. Showing the unique misfit in each culture. The worst residuals identified in Figure 7 are adjusted by subtracting the magnitude of the next largest residual for that cell. The top, middle, and bottom tableplots are the Shona (Piedmont et al., 2002), Portuguese (Lima, 2002), and Marathi (Lodhi et al., 2002) residual tableplots, respectively. Symbols are scaled to observed maximum of 0.31; cell labels in hundredth decimal.

Using Tableplots to Assess Precise Predictions

Comparing the Normative to an Estimated Factor Pattern

Returning to the appraisal of FFT, we first consider some new ways of assessing the similarity between the Shona and normative factor pattern. One approach is to examine the two corresponding tableplots. Figure 3 is the tableplot of the Shona factor pattern. Unlike Table 1, Figures 1 and 3 more clearly reveal the nature of each factor pattern for easy comparison. For example, in the Shona factor pattern A4 and A6 have a high positive loading on Conscientiousness, while C2, C4, and C5 have a high positive loading on Agreeableness; in contrast, these loadings are mostly very small in the normative factor pattern.

As a more detailed assessment of misfit between factor patterns, we overlay the tableplots of Figures 1 and 3 to create the superimposed tableplot in Figure 4 (the additional cell label appears in the upper right corner of each cell). This superimposed tableplot reveals misfitting loadings more clearly than two separate tableplots. For example, the previously noted discrepancies in Agreeableness and Conscientiousness facets stand out in Figure 4 as symbols with little overlap. One can also see further instances of bad fit (e.g., E5 and O1) that may have been less obvious from looking at Figures 1 and 3.

Certain features could enhance a superimposed tableplot so that it more effectively shows misfit. In cells with superimposed symbols, we only fill in (using white) the interior of the |smallest symbol| (see Figure 4); this facilitates seeing discrepancy between superimposed symbols. Because ϕ summarizes the misfit of many loadings, it could help pinpoint sources of misfit in a tableplot; this is especially helpful for large factor patterns with many rows and columns. We, thus, include the factor, facet, and total ϕs in Figure 4 to provide a visual summary of the other 150 pairs of symbols. Accordingly, we use a darker cell background and a yellow interior to distinguish the ϕ cells from the rest of the tableplot in Figure 4. (The ϕ symbols are scaled to a |maximum| of 1.) For example, by looking at the facet ϕs in Figure 4, E5 easily jumps out as the worst replicated facet; one could then study the loadings of E5 in more detail to understand how the misfit occurred.

Another approach for seeing misfit is a "residual" tableplot comprised of the difference between a predicted and observed factor pattern. Figure 5 is the residual tableplot obtained by subtracting the Shona factor pattern from that of the normative sample. Analogous to the ϕs in Figure 4, we show the factor, facet, and total average absolute differences (AADs) in Figure 5. For example, the AAD of N1 is 6; this is the average of the five residuals (differences) for N1 in absolute value. The AADs in Figure 5 serve the same purpose as the ϕs in Figure 4. Residuals have a theoretical |maximum| of 2; this occurs if the predicted loading is 1 and the observed loading is –1, or vise versa. Because it is very unlikely for the theoretical |maximum| to occur, we scaled all the cells in Figure 5 to the observed |maximum| of 0.44. Cell labels in Figure 5 represent the hundredth decimal.

Figure 5 clearly exposes the sources and degree of misfit between the two factor patterns. For example, E5, O1, O2, A3, A4, A5, and C2 appear to be the worst predicted facets in that they each have at least two very large residuals. On the other hand the facets of Neuroticism as a group appears to be the best replicated in comparison to facets of other factors.

One could further study the nature of misfit between the Shona and normative factor pattern based on the tableplots of Figures 1, 3, 4, and 5. We do not give a full analysis here. Our goal is simply to point out that tableplots can give helpful insights for assessing precise predictions of factor patterns, especially insights that are not easily obtainable from extant methods of assessment. We also believe that each of the tableplots of Figures 1, 3, 4, and 5 is invaluable. Whereas Figures 1 and 3 clearly show the nature of the original factor patterns, Figure 4 shows how the misfit occurs with regards to the original factor patterns, and Figure 5 directly reveals the extent of misfit. A thorough diagnosis should involve all such tableplots.

Comparing the Normative to Several Estimated Factor Patterns

As a more elaborate appraisal of FFT, we evaluate the Shona factor pattern in conjunction with those for the Portuguese and Marathi samples. Previously we raised some interesting questions with regards to the misfit of Openness in these three cultures. We explore how tableplots could answer these questions, as well as address other points of interest for cross-cultural comparisons of the NEO PI-R.

To answer some of the earlier questions, we examine the Shona, Portuguese, and Marathi residual tableplots (i.e., normative factor pattern minus that of each culture) in Figure 6. Because we intend to compare the three residual tableplots, their symbols are scaled to the common observed |maximum|. Furthermore, given the large number of cells in Figure 6, we use a very light grey to label residuals in order to reduce visual clutter. In terms of the misfit of Openness, one can see from Figure 6 that there are noteworthy cultural differences and similarities. For example, the Shona O1 to O4 have very high positive residuals on Openness; the other two cultures show considerably smaller residuals in these cells. Figure 6 also reveals that in all three cultures the facets of Agreeableness have high residuals on Openness; with the exception of Marathi A6, the signs of these residuals are identical across cultures.

Of course, any consistent cultural differences would be disconfirming evidence for FFT but a detailed examination of this misfit could help refine FFT, not just refute it (e.g., perhaps different subsets of factors are replicable in different cultures). Thus, a pertinent concern may be what cultures have the worst fit and for what facets. One could apply the previous types of tableplots to study the misfit within and between cultures. However, these types of tableplots (e.g., Figures 4 and 6) do not easily reveal where the worst (or best) fit occurs. Accordingly, we propose several modifications to the residual tableplot.

To identify where the worst fit occurs in Figure 6, for a given facet one could emphasize the cells that contain the |largest| residual amongst the three cultures. For example, in the C-N1 cells the Shona, Portuguese, and Marathi residuals are 0.04, –0.24, and –0.07, respectively; we use a darker cell background and label for the Portuguese cell. In situations where the maximum is not unique (e.g., see the O-N1 cells), we do not add any emphasis. This modified version of Figure 6 appears in Figure 7.

Compared to Figure 6, Figure 7 more easily reveals that, generally speaking, the Shona factor pattern may be regarded as the worst fitting of the three cultures (i.e., more occurrences of largest residuals in the Shona pattern). There are, however, some instances where the fit is substantially worse in the other two cultures (the Portuguese N1 and A1; the Marathi O5, O6, A2, and A6).

It could also be of interest to see how the misfit may be unique in each culture. That is, to examine the worst residuals after adjusting for the magnitude of other residuals. For example, –0.07 is the second |largest| residual in the cells of C-N1; the |largest| residual of these cells is –0.24. We reduce the size of –0.24 by |–0.07| and obtain an adjusted residual of –0.17. Figure 8 is a modification of Figure 7 in that only the adjusted residuals are shown (e.g., –0.17 appears in the C-N1 cell of the Portuguese tableplot; the corresponding cells in the other two tableplots are blank). This adjusted residual is not defined if there is no |unique maximum|. For example, the residuals in the cells for O-N1 are 3, 3, and 0; the corresponding cells in Figure 8 are blank.

Adjusted residuals indicate the degree of "unique" misfit among the cultures. For example, Figure 7 indicates that the Shona A5 and C2 are among the worst predicted facets (i.e., they each have two or more very high residuals). Figure 8, however, reveals that the misfit of the Shona A5 and C2 is not that bad once we take into account the corresponding misfit in the other two cultures. In contrast, the Shona E5, O1, A3, and A4 stand out in Figure 8 as they have two or more large residuals, even after the misfit of other cultures has been accounted for.

Through Figure 7 one could assess the extent of disconfirming evidence for FFT. Figure 8 could suggest revisions of FFT by revealing how the FFM is uniquely misfitting in each culture. Thus, both types of residual tableplots would be helpful for comparing results from multiple samples. Beyond our illustrations, one could explore other definitions of residuals or of unique misfit.

Discussion and Conclusion

The tableplot is not the first instance of a graphic representation of a table.[2] Such displays, designed to highlight patterns and trends in numeric results, have a history that goes back at least as far as Lambert's (1779) semigraphic table of periodic variation in soil temperature. Among modern contributions, Bertin (1967, 1977) developed the idea of the *reorderable matrix*, a value-shaded display of a table of positive numbers, where the rows and columns can be permuted to show the table structure more clearly. For tables of integers or frequencies, Bachi (1968) used *graphical rational patterns*, where the symbol shown in a cell is both a visual and numeric mapping of the cell value.

Other, more specialized semigraphic tables have more recently appeared. For example, Friendly (2002) discusses various ways to render a correlation matrix (hue shading, pac-man pie symbols, etc.) and variable reordering to make the structure of correlations visually apparent. Friendly (1994) and others (e.g., Zeileis, Meyer, & Hornik, 2007) developed the use of the mosaic display, with residual-based shading schemes for visualizing multi-way contingency tables in relation to associated log-linear models.

The tableplot, however, has two unique features that render it especially useful for the applications we have proposed. Unlike other semigraphic displays, one can overlay tableplots; this feature offers an intuitive visualization to evaluate the similarity of two or more tables (e.g., Figure 4). Furthermore, a tableplot is simultaneously a table and a plot; so in addition to giving a graphical representation of a set of values, a tableplot also offers high accuracy in the interpretation by printing these values. That is, graphs facilitate seeing patterns whereas tables facilitate lookup to read a precise value; tableplots provide for both operations. The accessibility of the actual values is very helpful for our purposes because precise predictions must be evaluated with accuracy.

Like other graphical displays, choices in how to construct a tableplot can influence how effectively the tableplot conveys its intended message. For example, we decided to scale a set of residuals by their observed maximum absolute value; a different scaling would no doubt change the appearance of the tableplot (e.g., large residuals could appear smaller). Choosing an appropriate scaling is, thus, analogous to setting the range on the axes of a scatterplot, as the amount of blank space around the data cloud affects how large the bivariate relationship appears (e.g., Cleveland, Diaconis, & McGill, 1982).

The choice of circle (and diamond) as plot symbols is another decision that affects the perception of the plotted quantities. For a given positive loading, the diameter and area of the associated circle, and distance of that circle from its cell frame all indicate the loading's magnitude. (A diamond uses analogous visual cues.) Thus, the interpretation

[2] We thank a reviewer for pointing us to some alternatives.

of a tableplot invokes the elementary perceptual tasks of judging position along a common scale, position along a nonaligned scale, length, and area (e.g., Cleveland & McGill, 1984). Circle, diamond, or square as plot symbols are ideal if one intends to overlay symbols, but otherwise, one could consider symbols that do not have area (e.g., "+" or "ϕ"). Because area judgment is not as accurate as the other perceptual tasks (e.g., Cleveland & McGill, 1984), a tableplot that does not invoke area judgment might lead to more accurate interpretations.

Of course, the focus of this paper is not on how to construct tableplots. Thus, we do not offer any concrete recommendations on the graphical issues we have reviewed. A future paper will discuss how to use our forth-coming R package to construct various tableplots; these graphical issues will be more formally addressed in that paper. Particularly, we will examine how to optimally apply various features of the R package to produce tableplots.

We note that the tableplot can be useful in other contexts. Factor-analytic research of other models of personality structure could similarly benefit from a more rigorous and cross-cultural analysis (tableplots are not restricted to evaluating predictions of FFT). Indeed, one could improve FA applications, in general, by using tableplots to interpret and present FA results (Kwan, 2008b). Beyond FA, Friendly & Kwan (in press) also applied tableplots as a new approach to visualizing collinearity diagnostics.

In closing we provide two important clarifications. First, we are not proposing that tableplots should replace ϕ or ϕ significance tests. Consider the role of the scatterplot in the study of bivariate relationships. It is without doubt that the Pearson correlation and the NHST of this correlation are informative. However, to thoroughly understand a bivariate relationship, one must also examine its scatterplot. Analogously, we believe that to thoroughly assess the fit between predicted and observed factor patterns, researchers should examine various tableplots of these patterns, particularly to see the nature of the agreement or discrepancy. Relying only on summary statistics or significance tests is not enough.

Furthermore, as one reviewer reminded us, theory appraisal must involve more than just the evidence from a graph or statistic. A past criticism of NHST is that some researchers have come to regard theory appraisal as involving little else beyond establishing statistical significance (e.g., Bakan, 1966; Bolles, 1962; Carver, 1978; Meehl, 1978). We, thus, clarify that although the tableplot is a powerful tool for seeing patterns in tables, it only becomes a powerful tool for theory appraisal if researchers apply their knowledge, experience, and ingenuity to interpret what they see.

References

Allik, J., & McCrae, R.R. (2002). A five-factor theory perspective. In R.R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 303–322). New York: Kluwer Academic/Plenum.

Bachi, R. (1968). *Graphical rational patterns: A new approach to graphical presentation of statistics*. Jerusalem: Israel Universities Press.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

Bertin, J. (1967). *Sémiologie graphique: Les diagrammes, les réseaux, les cartes***[translation please]**. Paris: Gauthier-Villars.

Bertin, J. (1977). *La graphique et le traitement graphique de l'information***[translation please]**. Paris: Flammarion.

Bolles, R.C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, *11*, 639–645.

Browne, M.W. (1972). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology, 25*, 115–120.

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

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378–399.

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.

Cleveland, W.S., Diaconis, P., & McGill, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science*, *216*, 1138–1141.

Cleveland, W.S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531–554.

Costa Jr., P.T., & McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Floyd, F.J., & Widaman, K.F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286–299.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association, 89*, 190–200.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician, 56*, 316–324

Friendly, M., & Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics and Data Analysis*, *43*, 509–539.

Friendly, M., & Kwan, E. (in press). Where's Waldo: Visualizing collinearity diagnostics. *The American Statistician*.

Kwan, E. (2008a). *Improving factor analysis in psychology: Innovations based on the null hypothesis testing controversy*. Unpublished doctoral dissertation, York University, Toronto, ON.

Kwan, E. *Tableplot: A new display for factor analysis.* (in preparation).

Lambert, J.H. (1779). *Pyrometrie; oder, vom maasse des feuers und der wärme mit acht kupfertafeln***[translation please]**. Berlin: n.p.

Lima, M.P. (2002). Personality and culture: The Portuguese case. In R.R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 249–260). New York: Kluwer Academic/Plenum.

Lodhi, P.H., Deo, S., & Belhekar, V.M. (2002). The five-factor model of personality: Measurement and correlates in the Indian context. In R.R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 227–248). New York: Kluwer Academic/Plenum.

Lorenzo-Seva, U., & ten Berge, M.F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57–64.

McCrae, R.R. (2001). Trait psychology and culture: Exploring intercultural comparisons. *Journal of Personality*, *69*, 819–846.

McCrae, R.R., & Costa Jr., P.T. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J.S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 51–87). New York: Guilford.

McCrae, R.R., & Costa Jr., P.T. (1999). A five-factor theory of personality. In L.A. Pervin & O.P. John (Eds.), *Handbook of personality: Theory and research* (2 ed., pp. 139–153). New York: Guilford.

McCrae, R.R., & John, O.P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175–215.

McCrae, R.R., Zonderman, A.B., Costa Jr., 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 and Social Psychology, 70*, 552–566.

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Meehl, P. (1986). What social scientists don't understand. In D.W. Fiske & R.A. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 315–338). Chicago: University of Chicago Press.

Meehl, P. (1990). Appraising and amending theories: The strategy of Lakatosian defense and the two principles that warrant it. *Psychological Inquiry, 1*, 108–141.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Mahwah, New York: Erlbaum.

Paunonen, S.V., Jackson, D.N., Trzebinski, J., & Forsterling, F. (1992). Personality structure across cultures: A multimethod evaluation. *Journal of Personality and Social Psychology*, *62*, 447–456.

Piedmont, R.L., Bain, E., McCrae, R.R., & Costa Jr., P.T. (2002). The applicability of the five-factor model in a sub-Saharan culture: The NEO PI-R in Shona. In R.R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 155–173). New York: Kluwer Academic/Plenum.

Rolland, J. (2002). Cross-cultural generalizability of the five-factor model of personality. In R.R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 7–28). New York: Kluwer Academic/Plenum.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416–428.

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

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

Wainer, H. (1997). Improving tabular displays, with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*, 1–30.

Wrigley, C.C., & Neuhaus, J.O. (1955). The matching of two sets of factors. *American Psychologist*, *10*, 418–419.

Zeileis, A., Meyer, D., & Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, *16*, 507–525.

Ernest Kwan

Sprott School of Business
Carleton University
1125 Colonel By Drive
Ottawa, Ontario K1S 5B6
Canada
Tel. +1 613 520-2600 Ext. 3007
Fax +1 613 520-4427
E-mail ernest_kwan@carleton.ca