Optimal Data Analysis
Vol. 3 (March 29, 2014), 14-16

2155-0182/10/$3.00

UniODA *vs.* Kappa: Evaluating the Long-Term (27-Year) Test-Retest Reliability of the Type A Behavior Pattern

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

This 27-year follow-up investigated long-term stability of Type A behavior (TAB) for 1,180 surviving participants in the Western Collaborative Group Study. The kappa statistic was used to assess reliability among and between self- and Structured Interview-based TAB assessments. Results indicated fair temporal reliability for self-assessments (kappa=0.39), moderate temporal reliability for Interview-assessments (0.24), and low parallel-forms reliability for self and Interview assessments at intake (0.16) or follow-up (0.11). In contrast, for these data UniODA found relatively strong temporal reliability for self-assessments (ESS=50.2), and weak effects for all other estimates (ESS≤23.3).

Table 1 presents the 27-year test-retest cross-classification table obtained for TAB assessments (Type A or Type B) based on the Structured Interview (SI).[1] As seen, 32% of Type As became Type Bs, and 45% of Type Bs became Type As. Although criticized on several grounds[2] the kappa statistic was used to assess temporal reliability. Here kappa=0.24, reflecting "moderate" reliability (estimated $p<0.0001$).

Discussion and examples of application of UniODA in reliability analysis is presented elsewhere.[2] A confirmatory UniODA analysis was conducted presently to test the *a priori* hypothesis that TAB assessments are consistent across time—that is, ratings fell into the major diagonal running from the upper left-hand corner to the lower right-hand corner of the test-retest cross-classification table.[3] The UniODA model achieved relatively weak[2] ESS of 23.3 (exact $p<0.0001$), indicating relatively weak test-retest reliability.

Table 1: Stability of Structured Interview-based assessments of TAB over 27-year test-retest interval (Carmelli et al., 1991)

Second Rating	Initial Rating Type A	Type B
Type A	372	284
Type B	175	349

Note: Tabled are frequency counts.

Table 2 gives the 27-year test-retest cross-classification table for TAB self-

2155-0182/10/$3.00

assessments.[1] As seen, 56% of Type As became Type Bs, and 3% of Type Bs became Type As. For these data, kappa=0.39, reflecting "fair" reliability (estimated $p<0.0001$). A confirmatory UniODA model achieved relatively strong[2] ESS of 50.2 (exact $p<0.0001$), indicating relatively strong test-retest reliability. Here the stability is attributable to Type Bs. The instability of the Type As underscores the success of efforts to modify TAB behavior at the study outset.[1]

Table 2: Stability of TAB self-assessments over 27-year retest interval (Carmelli et al., 1991)

Second Rating	Initial Rating: Type A	Initial Rating: Type B
Type A	180	8
Type B	227	272

Note: Tabled are frequency counts.

Finally, Table 3 gives the parallel-forms reliability[2] cross-classification table for the SI and self-assessment, at study intake and also at study follow-up.[1]

Table 3: Agreement of Structured-Interview and TAB self-assessments (Carmelli et al., 1991)

Study Intake

Self Rating	SI-Based Rating: Type A	SI-Based Rating: Type B
Type A	209	95
Type B	199	185

Study Follow-Up

Self Rating	SI-Based Rating: Type A	SI-Based Rating: Type B
Type A	118	67
Type B	243	243

Note: Tabled are frequency counts.

For intake data kappa=0.16, and for follow-up data kappa=0.11: both results indicate "low" inter-method concordance and have estimated $p<0.01$.

For intake data the confirmatory UniODA model yielded ESS=16.9 (exact $p<0.0001$), and for follow-up data UniODA achieved ESS=13.8 (exact $p<0.002$): both indicate relatively weak parallel-forms reliability.

In contrast to kappa, UniODA tests the *a priori* hypothesis that data are consistent across time, or that different methods agree; UniODA provides an exact Type I error rate; and the index of UniODA performance is normed—for *every* UniODA analysis the ESS which is expected by chance is 0, and errorless, prefect classification is 100. There thus is no apparent rational reason to use kappa in applications such as the present.

References

[1]Carmelli D, Dame A, Swan G, Rosenman R (1991). Long-term changes in Type A behavior: A 27-year follow-up of the Western Collaborative Group Study. *Journal of Behavioral Medicine*, *14*, 593-606.

[2]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books. For a discussion of kappa see p. 124.

[3]UniODA analysis of data in Table 1 was accomplished using the following code: commands are indicated in red.[2]

```
OPEN DATA;
CATEGORICAL ON;
OUTPUT carmelli.out;
TABLE 2;
CLASS ROW;
DIRECTIONAL < 1 2;
MCARLO ITER 25000;
```

2155-0182/10/$3.00

```
DATA;
372 284
175 349
END DATA;
GO;
```

Author Notes

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., Suite 163
Chicago, IL 60646

eMail: Journal@OptimalDataAnalysis.com