University of Groningen

Commentaries on the Ten Most Highly Cited Psychometrika Articles from 1936 to the Present

Kiers, Henk; McArdle, John; Lewis, Charles; Hubert, Lawrence; Heiser, Willem; Takane, Yoshio; Meulman, Jacqueline; Sijtsma, Klaas; Kohn, Hans-Friedrich

Published in:
Psychometrika

DOI:
[10.1007/s11336-016-9540-y](https://doi.org/10.1007/s11336-016-9540-y)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Kiers, H., McArdle, J., Lewis, C., Hubert, L., Heiser, W., Takane, Y., Meulman, J., Sijtsma, K., & Kohn, H-F. (2016). Commentaries on the Ten Most Highly Cited Psychometrika Articles from 1936 to the Present. *Psychometrika*, 81(4), 1177-1211. <https://doi.org/10.1007/s11336-016-9540-y>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

COMMENTARIES ON THE TEN MOST HIGHLY CITED PSYCHOMETRIKA ARTICLES FROM 1936 TO THE PRESENT

The ten most frequently cited articles appearing in *Psychometrika* since its establishment in 1936 are highlighted in a series of ten commentaries. They are grouped in three themes, with chronological ordering within these groups. The first group regards some characteristic problems in factor analysis, the second group is about the analysis of proximities, and the third group addresses psychometric concerns in multivariate analysis.¹

1. Solutions for Problems in the Factor Analysis Model

KAISER, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200. (5237 citations according to Google Scholar as of April, 2016)

Applications of Principal Components Analysis or Exploratory Factor Analysis in the behavioral sciences are usually followed by rotation aiming at simple structure. And the rotation method used is very often Kaiser's normal varimax. What makes varimax so popular?

Varimax rotation is certainly not the only simple structure rotation method, nor was it the first:

The first analytic criterion for determining psychologically interpretable factors was presented in 1953 by Carroll.

[Kaiser thereby ignored earlier proposals to partially use analytic rotation criteria along with graphical methods (e.g., Tucker, 1944), probably because he strongly disliked graphical methods]

[...] before the advent of computers, rotations were always carried out on a subjective graphical basis. Scientifically, of course, this was nonsense, and perhaps led, more than anything else, to a bad name for factor analysis among professional mathematicians and statisticians (Kaiser, 1960, pp. 146–147).

Kaiser considered Carroll's proposal a breakthrough: the first method to optimize a single criterion operationalizing Thurstone's (1947) rules for simple structure. On the other hand, he made it clear that there was room for improvement, to say the least:

In the light of later developments, Carroll's criterion should probably be relegated to the limbo of 'near misses'; however, this does not detract from the fact that it was the first attempt to break away from an inflexible devotion to Thurstone's ambiguous, arbitrary, and mathematically unmanageable qualitative rules for his intuitively compelling notion of simple structure (p. 188).

¹Contributions from: Willem Heiser, Leiden University; Lawrence Hubert, University of Illinois; Henk Kiers, University of Groningen; Hans-Friedrich Köhn, University of Illinois; Charles Lewis, Fordham University; Jacqueline Meulman, Leiden University; John McArdle, University of Southern California; Klaas Sijtsma, Tilburg University; Yoshio Takane, University of Victoria.

Almost simultaneously, Saunders (1953), Neuhaus and Wrigley (1954), and Ferguson (1954) proposed differently formulated, but equivalent methods for orthogonal simple structure rotation. Kaiser (1958) elegantly and briefly showed them to be equivalent to Carroll's method. Following Neuhaus and Wrigley, the method was called quartimax.

Kaiser (1958) observed that quartimax has a bias. To explain what tends to go wrong, he showed that quartimax was equivalent to maximizing yet another criterion, that is, the sum of row-wise variances of squared loadings. He wrote

[...] its aim is to simplify the description of each row, or test, of the factor matrix. It is unconcerned with simplifying the columns, or factors, of the factor matrix (probably the most fundamental of all requirements for simple structure). The implication of this is that the quartimax criterion will often give a general factor. [...] there is no reason why a large loading for each test may not occur on the same factor (p. 190).

And from these sentences, almost automatically the suggestion to maximize the varimax criterion emerged: if simplicity of the columns is what is most important, then one should maximize the sum of column-wise variances of the squared loadings, that is, the varimax criterion. The introduction of varimax in Kaiser's 1958 paper was less dramatic than it might seem: Kaiser simply identified the sum of column-wise variances of the squared loadings as the (raw) varimax criterion he had presented earlier at the 1955 APA conference, while referring to its abstract (Kaiser, 1955). Also, in 1955, the varimax criterion had been proposed in his unpublished PhD dissertation (see Carroll, 1957, or Kaiser 1979). The 1958 paper, however, is the first journal publication on varimax.

Kaiser (1958) offered quite a bit more than just the varimax criterion. He started out by comparing varimax rotation to quartimax rotation on two examples. He demonstrated the superiority of varimax over quartimax in that varimax did not yield such a strong general factor as quartimax did. Nevertheless, he observed in the second example that also after varimax rotation, the variance contributions of the four factors differed considerably (though less than after quartimax). Kaiser guided us in his search for what caused this difference:

It seems reasonable to attribute the systematic bias [...] to the divergent weights which implicitly are attached to the tests by their communalities. When one deals with fourth-power functions of factor loadings, a test with communality 0.6, for example, would tend to influence the rotations four times as much as a test whose communality was 0.3. [...] There seems no rational basis for choosing among different weighting schemes. Let us then make the agnostic confession of ignorance which pervades any form of correlational analysis. For the purpose of rotation, weight the tests equally, in the sense that the lengths of the common parts of the test vectors have equal length. (The author is indebted to Dr. D.R. Saunders for this suggestion.) (pp. 192–193).

This rescaling leads to the currently prevalent "normal varimax" criterion. Kaiser demonstrated on an example that this procedure indeed had the effect of reducing the differences between the variance contributions of the rotated factors.

After his ad hoc introduction of the weighting scheme, Kaiser came with an interesting additional rationale. He proved mathematically, for the case of two factors, that, when tests in a cluster of tests all are collinear as far as their common parts are concerned, then "[...] the normal varimax solution is invariant under changes in the composition of the test battery" (p. 195). Such invariance does not hold for (raw) varimax or quartimax. It looks like a very strong property, and even though he admitted that his result holds only in a special case, he goes on to write that "The principle of simple structure may probably be considered incidental to the more fundamental concept of factorial invariance" (p. 195). He next verified what happens for a real data example, when adding variables (tests) one by one, and observed a high degree of stability of loadings on factors that are sufficiently well represented. However, one example of course does not prove a

general invariance property. In fact, it is easy to demonstrate numerically that his result for the two-factor case, does not generalize to cases of more than two clusters of collinear variables (and hence, to cases of more than two factors). Moreover, it can be shown that his proven result is simply a consequence of the fact that, in his special case, variances are considered of vectors with only two different values, and such variances do not depend on how often such values occur. Hence, his invariance property holds for optimizing any criterion using column-wise variances, even for the opposite method of minimizing rather than maximizing the varimax criterion. So it seems that the factorial invariance property of normal varimax, even in this special case, is of no special value.

The paper essentially closes by a brief section on the oblique case. In these fourteen lines, Kaiser proposed an alternative criterion for oblique rotation (which was dubbed covarimin by Carroll 1957 even before Kaiser published his paper), and suggested how to optimize it. It has by no means become as popular as varimax, but it is remarkable that in this paper on orthogonal simple structure rotation, he offered, almost in passing, a new criterion for oblique rotation as well.

The actual last pages of the paper concern a technical appendix, which describes a planar rotation procedure for iteratively maximizing the varimax criterion. Surprisingly, the paper does not finish with a discussion section, which might have given suggestions for future research or further developments. In hindsight, we can conclude that such a suggestion would have been superfluous. The usefulness of varimax apparently was self-evident, and many future developments have taken Kaiser's varimax as a starting point. One of the earliest such developments was Harris and Kaiser's (1964) successful "orthoblique" approach, where orthogonal varimax rotation was, in an ingenious way, used to find simple loadings for oblique factors. As another early follow-up, the varimax solution was proposed to be used for setting up a simple target for oblique target rotation. The ensuing Promax method (Hendrickson and White, 1964) is also one of the more successful approaches to oblique simple structure rotation. Since then, various further developments along these lines have been proposed.

The varimax criterion has also inspired some authors to set up families of simple structure criteria, of which quartimax and varimax were the most prominent members (e.g., Clarkson and Jennrich, 1988). Furthermore, special procedures for application of varimax in other situations than that of a single ordinary factor loading matrix have been proposed. For instance, for the situation with several loading matrices for the same set of variables, Bloxom (1968) proposed a method for rotation of these loading matrices to maximum similarity and maximum simple structure in the varimax sense.

To conclude, with varimax, Kaiser has offered a tool which has been and still is very successful. Its success not only becomes apparent in the practice of exploratory factor analysis, but also in the development of new rotation methods, in which varimax was used as a basis.

UNIVERSITY OF GRONINGEN

Henk Kiers

References

- Bloxom, B. (1968). Factorial rotation to simple structure and maximum similarity. *Psychometrika*, 33, 237–247.
- Carroll, J. B. (1953). An analytic solution for approximating simple structure in factor analysis. *Psychometrika*, 18, 23–38.
- Carroll, J. B. (1957). Biquartimin criterion for rotation to oblique simple structure in factor analysis. *Science*, 126, 1114–1115.
- Clarkson, D. B., & Jennrich, R. I. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, 53, 251–259.

- Ferguson, G. A. (1954). The concept of parsimony in factor analysis. *Psychometrika*, 19, 281–290.
- Harris, C. W., & Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29, 347–362.
- Hendrickson, A. E., & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17, 65–70.
- Kaiser, H. F. (1955). An analytic rotational criterion for factor analysis. *American Psychologist*, 10, 438 (abstract).
- Kaiser, H. F. (1960). The application of electronic computer to factor analysis. *Educational and Psychological Measurement*, 20, 141–151.
- Kaiser, H. F. (1979). Citation classic commentary. *Current Contents/Social and Behavioral Sciences*, number 24.
- Neuhaus, J. O., & Wrigley, C. (1954). The quartimax method: An analytic approach to orthogonal simple structure. *British Journal of Mathematical and Statistical Psychology*, 7, 81–91.
- Saunders, D. R. (1953). An analytic method for rotation to orthogonal simple structure. Research Bulletin, RB 53-10. Princeton, NJ: Educational Testing Service.
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Tucker, L. R. (1944). A semi-analytical method of factorial rotation to simple structure. *Psychometrika*, 9, 43–68.

HORN, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. (3362 citations in Google Scholar as of April, 2016)

Who would have thought that a paper published over 50 years ago would still have an impact today. But this one still does! Such longevity must be due to the method being used—the simulation of the random sampling base in factor analysis. Over the years, the journal *Psychometrika* has been a haven for what some might consider mathematical minutiae, but this paper is definitely not one of them. It has enjoyed almost 3400 citations as of April 2016; the question of “why” comes to mind.

The solution to the burning question of the correct number of common factors in a matrix of correlations is easy. Henry Kaiser used to say that “it was so simple that he solved it everyday before breakfast” (from J.L. Horn, Personal Communication, June 1977). Horn recognized that a solution was needed; but he thought he was proposing just a temporary fix, put in place for others to develop more formal statistical procedures (which they soon did! For a full review, see Lawley and Maxwell 1963; Horn and Engstrom 1979; Jöreskog and Sörbom 1979). In Horn (1965), it was suggested that Guttman’s latent-root-one lower bound estimate for the rank of a correlation matrix be accepted as a psychometric upper bound, following the proofs and arguments of Kaiser and Dickman. Kaiser showed that for a “principal component to have positive KR-20 internal consistency, it is necessary and sufficient that the associated eigenvalue be greater than one” (Kaiser, 1960, p. 6); this was the source of the root-one criterion. This rule, however, was for a population, so Horn tried to add some sampling considerations. In the statistical calculation of the correlations and the roots, the rank for a sample matrix should be estimated by subtracting out the component in the latent roots that can be attributed to sampling error and the least squares “capitalization” on this error. Horn promoted a procedure, termed “Parallel Analysis” (PA) by later authors, based on the generation of random variables for estimating the component that needs to be subtracted.

Horn became an expert in the classical techniques of common factor analysis. He generally asked why any elegant mathematical-statistical theory should be based on specific assumptions when we know these key assumptions are wrong and untestable. Specifically, Horn suggested that the number of common factors should not be determined simply by using the well-known “eigenvalues greater than one” criterion defined by one of his favorite advisors, Henry Kaiser (1960). Instead, Horn proposed the use of “computer simulation” techniques, mainly based on model assumptions that were random, and advocated for their use whenever possible (see Horn and McArdle, 1980).

Horn (1965) determined the number of common factors by selecting the number of the eigenvalues of a correlation matrix that were greater than or equal to those provided by computer-simulated data with known characteristics. In this very simple idea, all that was needed was to generate “random data of similar size”; one could then calculate the latent roots and vectors of these random data to provide a criterion tailored to the particular dataset being analyzed. As an example, Horn (1965) implemented this procedure for 297 people measured on 65 ability variables; he found evidence for 16 common factors by the Kaiser criterion but only 9 by the PA criterion.

Of course, we now recognize the need to consider statistical fluctuations in these latent roots, but it is noteworthy that these kinds of calculations were done some 50 years ago! Horn’s random variable approach has more recently been found to be the most accurate for determining the number of unrotated common factors (e.g., Ledesma and Valero-Mora, 2007; Montanelli and Humphreys, 1976; Velicer, Eaton, and Fava, 2000; Zwick and Velicer, 1986; see also Dinno 2009; Hayton 2009; Courtney 2013 for recent evaluation and hearty affirmation). This success also marks the beginning of Horn’s fascination with the use of computer-simulated data to solve the most complex problems in mathematical statistics (see Horn and McArdle, 1980).

John Horn (1928–2006) was a pioneer in multivariate thinking and the application of multivariate methods to research on intelligence and personality. As a leader in multivariate methodology, Horn tried to reach the incredible heights of his well-known mentor, Raymond Cattell. John believed strongly in a multivariate scientific approach, and questioned the typical use of unweighted sum scores as if they represented the best scores of the psychological constructs of interest. On a substantive basis, John believed that there were important individual differences among adults within the domains of cognition and personality (see Horn and Knapp, 1974; Horn and Donaldson, 1977).

Some of Horn's early comments on the methods of factor analysis are worth repeating, especially the central concept of a "functional unity" (Horn, 1972, pp. 161–162). He applied the ideas about a multivariate meta-theory to data on cognitive abilities, to create various testable hypotheses. This type of reasoning provides the basis for arriving at several substantive results on cognitive abilities, perhaps the most important being that Horn expanded on this initial work of Cattell (Horn and Cattell, 1966, 1967) to identify additional functional unities of primary mental abilities.

John was also my mentor, teacher, colleague, and friend. I think that his major contributions to psychology and psychometrics will continue to be ahead of their time. Through his research and teaching, he forced people to question popular assumptions, evaluate all data available, and consistently challenged us to think longer, harder, and better. For a full review of his work, the interested reader is referred to McArdle and Hofer (2014).²

UNIVERSITY OF SOUTHERN CALIFORNIA

John McArdle

References

- Courtney, M. G. R. (2013). Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2.0 to make more judicious estimations. *Practical Assessment, Research and Evaluation*, 18(8), 1–14.
- Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, 44, 362–388.
- Hayton, J. C. (2009). Commentary on "Exploring the sensitivity of Horn's Parallel Analysis to the distributional form of random data." *Multivariate Behavioral Research*, 44, 389–395.
- Horn, J. (1972). State, trait and change dimensions of intelligence. *British Journal of Educational Psychology*, 42, 159–185.
- Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57, 253–270.
- Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, 26, 107–129.
- Horn, J. L., & Donaldson, G. (1977). Faith is not enough: A response to the Baltes–Schaie claim that intelligence does not wane. *American Psychologist*, 32, 369–373.
- Horn, J. L., & Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research*, 14, 283–300.
- Horn, J. L., & Knapp, J. R. (1974). Thirty wrongs do not make a right: Reply to Guilford. *Psychological Bulletin*, 81, 502–504.

²Acknowledgements: We thank the National Institute on Aging (AG0713720 to the author) for support of this research and to Carol Prescott (University of Southern California) for helpful comments on the draft. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health.

- Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical and statistical model building (MASMOB) in research on aging. In L. Poon (Ed.), *Aging in the 1980's: Psychological issues* (pp. 503–541). Washington, DC: American Psychological Association.
- Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151.
- Lawley, D. N. & Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London: Butterworth.
- Ledesma, R., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation*, 12(1), 1–11.
- McArdle, J. J., & Hofer, S. M. (2014). Fighting for intelligence: A review of the academic research of John L. Horn. *Multivariate Behavioral Research*, 49, 1–16.
- Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika*, 41, 341–348.
- Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In Goffen, R. D. & Helms, E., (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at Seventy* (pp. 41–71). New York: Springer.
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432–442.

TUCKER, L. R. & LEWIS, C. (1973). A reliability coefficient for maximum-likelihood factor analysis. *Psychometrika*, 38, 1–10. (4094 citations according to Google Scholar as of April, 2016)

Ledyard R Tucker's most frequently cited scientific article, and one of the most frequently cited *Psychometrika* articles, is the paper that he and I wrote in 1973, proposing a goodness-of-fit index for factor analysis. This index (and important generalizations introduced by Bentler and Bonett, 1980) is currently referred to as the Non-Normed Fit Index (NNFI) or the Tucker–Lewis Index (TLI). Presumably, the large number of citations of our article was influenced by the appearance of subsequent publications (especially those of Peter Bentler and his colleagues) that generalized our work and evaluated its usefulness in a variety of settings. This research, in turn, presumably led to popular structural equation modeling computer programs (such as AMOS, EQS, LISREL, and Mplus) including a version of the index in their output. When a researcher uses one of these programs, and includes the index in their published results, and wishes to give a source for the index—voilà! Another citation.

To see how the numbers of these citations have distributed themselves over the years since our article's publication, I made (mis)use of Google Scholar. The results are summarized in Table 1. The increase in the number of citations over the years is difficult to interpret without additional data, but in any event the numbers show no sign of decreasing.

Tuck made many important contributions to our field, most of which are represented by some 30 articles in *Psychometrika*. They include “Some mathematical notes on three-mode factor analysis” (Tucker, 1966), “An inter-battery method of factor analysis” (Tucker, 1958a), “An individual differences model for multidimensional scaling” (Tucker and Messick, 1963), “Determination of parameters of a functional relation by factor analysis” (Tucker, 1958b), and “Evaluation of factor analytic research procedures by means of simulated correlation matrices” (Tucker, Koopman, and Linn, 1969), to list only his most frequently cited papers (in descending order).

As far as I know, Tuck never cited our paper in his subsequent writings, and neither Tuck nor I ever followed up on the idea of a goodness-of-fit index for factor analysis. In a 2003 interview, he gave to Neil Dorans, Tuck said “Measures of goodness-of-fit can be helpful but they can also be misleading – especially the global ones” (Dorans, 2004, p. 10). I agree.

At the end of our 1973 article, Tuck writes: “The number of factors to accept appears to depend on size of loadings and meaningfulness of factoring results.” He concludes with the sentence: “Any accepted solution should have a high coefficient of reliability” (p. 9). In other words, only a limited role is proposed for the index, with the emphasis placed on the meaningfulness of the results.

TABLE 1.
Number of citations for the Tucker and Lewis (1973) article in 5-year intervals.

Years	Number of citations
1971–1975	5
1976–1980	26
1981–1985	56
1986–1990	87
1991–1995	222
1996–2000	303
2001–2005	489
2006–2010	941
2011–2015	1703

Given this lukewarm attitude toward the just-proposed measure, one might wonder why the paper was written at all. Part of the answer is that it was produced as a response to an idea that Tuck found highly unsatisfactory, namely using significance testing to help select an appropriate number of factors. This proposal was made by (among others) Karl Jöreskog in his 1967 *Psychometrika* article on maximum-likelihood factor analysis.

The idea of using significance testing to evaluate a statistical model must have reminded Tuck of two articles that had appeared in *Psychometrika* several years previously regarding Thurstone's method of paired comparisons. An article written in 1951 by Fred Mosteller begins with the sentence "It would be useful in Thurstone's method of paired comparisons to have a measure of the goodness-of-fit of the estimated proportions to the observed proportions" (Mosteller, 1951, p. 207). It goes on to develop a test of the hypothesis that the model is correct for the population from which the data have been sampled, and seems to say that a failure to reject this hypothesis should be treated as providing support for the model.

A response to this idea appeared in a 1958 *Psychometrika* article written by Harold Gulliksen and John Tukey with the title "Reliability for the Law of Comparative Judgment." (It may be noted that the authors thanked Ledyard Tucker and Frederic Lord for "valuable suggestions.") The paper begins with the observation that significance tests have a well-known weakness, namely their sensitivity to sample size. Cochran and Fisher are identified among those who have also discussed this issue.

As an alternative approach to the problem of evaluating the fit of a model, Harold and John proposed a "reliability" measure. Essentially, they defined this reliability as the ratio of the variance associated with the model to the total variance. The analogy between their measure and the ratio of true score variance to observed score variance presumably accounts for their use of the term "reliability."

When evaluated for an example given by Mosteller that had a non-significant χ^2 , the value obtained for the estimated reliability was only .73. A second example, based on a larger dataset, yielded a significant chi-square value ($\chi^2_{28} = 126.76$, $p < 0.0001$) combined with a reliability of about .95.

Our paper begins with a brief review of developments in maximum-likelihood factor analysis, concluding with the Jöreskog (1967) article. It then reviews the shortcomings of significance testing in the context of likelihood ratio tests applied to factor analysis. Next, the work of Mosteller (1951) and the response of Gulliksen and Tukey (1958) are described. It is concluded that "An analogous reliability type coefficient is needed for factor analysis" (p. 3). I hope that this background may help to explain why our paper was written in the first place. It may also help to explain where the term "reliability coefficient" in our title comes from. Happily, subsequent researchers abandoned this label in favor of the more appropriate "goodness-of-fit index."

The likelihood ratio statistic (denoted here by X_m^2) for an unrestricted factor model with m factors has an asymptotic null distribution that is χ^2 with degrees-of-freedom denoted by df_m . Arguing by analogy, in our paper, we treat X_m^2 as though it were a sum of squared deviations of observations from their fitted values. When divided by its degrees-of-freedom, and by $N - 1$,³ it is analogous to a mean square, and is denoted by M_m :

$$M_m = \frac{X_m^2}{(N - 1)df_m}.$$

Its expectation is written as the sum of two components

$$E(M_m) = \delta_m + \epsilon_m.$$

³We actually used $n'_m = N - 1 - (n + 5)/6 - 2m/3$, where n is the number of variables, but all subsequent work uses $N - 1$, so that is what I do here as well.

Here δ_m represents a component representing systematic deviations of the model from the population (proportional to a non-centrality parameter) and ϵ_m is a component associated with sampling variation. For the null case when $\delta_m = 0$, we have

$$E(X_m^2 | \delta_m = 0) = df_m,$$

so

$$E(M_m | \delta_m = 0) = E\left(\frac{X_m^2}{(N-1)df_m} | \delta_m = 0\right) = \frac{1}{N-1}.$$

This allows us to write

$$\epsilon_m = \frac{1}{N-1}.$$

We may also obtain an unbiased estimator for δ_m :

$$\hat{\delta}_m = M_m - \frac{1}{N-1}.$$

Rewriting this estimator gives

$$\hat{\delta}_m = \frac{X_m^2}{(N-1)df_m} - \frac{1}{N-1} = \frac{X_m^2 - df_m}{(N-1)df_m}.$$

Note that the square root of this quantity is the familiar RMSEA introduced by Steiger and Lind (1980).

The idea that Tuck had in our paper was to compare the fit for a factor model with m factors to the fit for a factor model with zero factors (i.e., the model assuming independence among the variables). In our article, the population goodness-of-fit index for a model with m factors is denoted by ρ_m . It may be defined as

$$\rho_m = \frac{\delta_0 - \delta_m}{\delta_0}.$$

We may write the estimator proposed in our paper as

$$\hat{\rho}_m = \frac{\hat{\delta}_0 - \hat{\delta}_m}{\hat{\delta}_0} = \frac{M_0 - M_m}{M_0 - 1/(N-1)} = \frac{X_0^2/df_0 - X_m^2/df_m}{X_0^2/df_0 - 1}.$$

In effect, this index compares the mean squared errors of approximation (MSEA) for the factor models with zero factors and with m factors, expressed as a proportional reduction in MSEA. As our paper says in the concluding paragraph, “It [the index] does not appear to provide a criterion as to how many common factors to accept [...] The number of factors to accept appears to depend on the size of loadings and the meaningfulness of factoring results” (p. 9).

Finally, I would like to add a personal note. I came to the University of Illinois at Urbana-Champaign in the summer of 1970. One of my primary motivations for joining Quantitative Psychology was to have the chance to work with Ledyard Tucker. Sometime that fall, Tuck told me that he had been working on a paper on maximum-likelihood factor analysis. He invited me to join him in this project. I was only too happy to accept, but wondered what I might contribute. Tuck's answer was that, since I had learned about maximum-likelihood factor analysis in a course taught at Princeton by Karl Jöreskog, perhaps I could critically discuss Tuck's ideas from the perspective of someone with that training. In addition, I believe that he wanted to help me get a start on my career in psychometrics. This gesture was typical of the kindness and generosity Tuck showed to all those who knew him. Working with and getting to know Tuck was a wonderful experience. Little did he (or I) know that our article would take on such a life of its own!

FORDHAM UNIVERSITY

Charles Lewis

References

- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Dorans, N. J. (2004). A conversation with Ledyard R. Tucker. Princeton, NJ: ETS.
- Gulliksen, H., & Tukey, J. W. (1958). Reliability for the law of comparative judgment. *Psychometrika*, 23, 95–110.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32, 443–482.
- Mosteller, F. (1951). Remarks on the method of paired comparisons, III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, 16, 207–218.
- Steiger, J. H., & Lind, J. M. (1980). Statistically-based tests for the number of common factors. In *Paper presented at Psychometric Society meeting*, Iowa City, IA.
- Tucker, L. R. (1958a). An inter-battery method of factor analysis. *Psychometrika*, 23, 111–136.
- Tucker, L. R. (1958b). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, 23, 19–23.
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279–311.
- Tucker, L. R., & Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika*, 28, 333–367.
- Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, 34, 421–459.

KAISER, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–36. (6043 citations as of April, 2016)

Simple structure is one of the central themes in multiple factor analysis. But the problem of how to measure simple structure explicitly for any given matrix of factor loadings had not been addressed until the appearance of this paper. Kaiser started with the following definition of factorial simplicity:

After a factor analysis has been completed, it is of interest to assess how good the solution is, in the sense of how simple – and thus how interpretable – the final pattern matrix is. The ideal solution, most investigators would agree, is one that is unifactorial, i.e., a solution for which each row of the pattern matrix has one, and only one, non-zero loading (p. 31).

Acceptance of that ideal radically reduced the ambiguity present in earlier discussions of simple structure.

From his conceptual description of simplicity, it followed naturally that he was going to propose an index which measures a tendency toward unifactoriality for each row of the factor matrix separately, and for the matrix as a whole. He asserted that the so-called quartimax criteria for rotation of a factor matrix would be most suitable for this purpose. Note what a startling move it was for Kaiser to return to quartimax after he had rejected precisely this criterion for rotation in Kaiser (1958), on the grounds that it often does not simplify the loadings of each factor and even tends to give a general factor. That was the reason why he introduced in his 1958 article the varimax criterion for rotation, which requires a bimodal distribution of loadings for each column of the factor matrix separately and for the matrix as a whole.⁴ After having dismissed quartimax for rotation, he now simply invented a new use for it.

First, Kaiser demonstrated rigorously that the different rotation criteria developed by Carroll (1953), Neuhaus and Wrigley (1954), and Saunders (1953) become equivalent if they are scaled to lie between zero and one. Adopting the name “quartimax” proposed by Neuhaus and Wrigley, he remarked: “It should be pointed out that, while we develop our index from the quartimax viewpoint, our results are applicable to any factor pattern matrix” (p. 31). Hence the resulting index would be useful both for comparing the simplicity of different rotations of one basic factor solution, as well as for characterizing the simplicity of (rotated) solutions for different datasets.

The resulting index for factorial simplicity (IFS) of a matrix of loadings with elements v_{js} , with $j = 1, \dots, p$ indexing variables, and $s = 1, \dots, q$ indexing factors, looks as follows:

$$\text{IFS} = \sqrt{\frac{\sum_{j=1}^p \left[q \sum_{s=1}^q v_{js}^4 - \left(\sum_{s=1}^q v_{js}^2 \right)^2 \right]}{\sum_{j=1}^p \left[(q-1) \left(\sum_{s=1}^q v_{js}^2 \right)^2 \right]}}.$$

Here the square root is taken to diminish the influence of the fourth-power functions, and to make the index “more consistent with other statistical measures, e.g., the correlation ratio” (pp. 34–35). The worst possible value of IFS is zero; it occurs when all loadings are equal in absolute value for all rows (variables). The best possible value of IFS is one: it occurs if only one loading is non-zero for all variables.

Next, Kaiser realized that it would be useful to calibrate the IFS scale. To achieve a reasoned answer, he considered the special case of a row of the factor matrix with q elements, c of which

⁴For more details, see the commentary on Kaiser (1958).

are non-zero and equal in absolute value, and $(q - c)$ are zero. Here, c refers to the complexity of the variable. He then obtained the following result for the squared simplicity of row j :

$$\lim_{q \rightarrow \infty} (\text{IFS}(j))^2 = \lim_{q \rightarrow \infty} \frac{q - c}{c(q - 1)} = \frac{1}{c}.$$

So for a large number of factors, IFS converges to 0.50 if all variables have complexity 4, which he regarded unacceptably far from unifactoriality. He also offered a table of IFSs calculated for different values of q and c , to help the calibration.

By referring to his experience with IFSs for a substantial number of factor analyses “from the real world”, Kaiser reached the following—admittedly still subjective—verbal evaluation for the levels of his index of factorial simplicity:

in the .90s, marvelous;
in the .80s, meritorious;
in the .70s, middling;
in the .60s, mediocre;
in the .50s, miserable;
below .50, unacceptable.

This verbal calibration concludes the paper, but there is an unexpected caveat of considerable historical importance. For all articles that have appeared in *Psychometrika* since its first volume in 1936, this highly cited paper by Henry Kaiser is second only to Cronbach’s massively cited 1951 article on coefficient α . However, it appears that almost all but a very small handful of the citations to Kaiser (1974) are in error. Some possible reasons for these misattributions are laid out in what follows.

In his Psychometric Society Presidential Address, Kaiser (1970) added a few embellishments to his well-known approach to factor analysis that Chet Harris had disdainfully labeled “Little Jiffy”; this approach could be characterized as “principal components with associated eigenvalues greater than one followed by normal varimax rotation”. One such addition in this 1970 paper was a “measure of sampling adequacy” (MSA) that was intended to reflect whether it was reasonable to proceed with a factor analysis in the first place. Kaiser attributed this MSA to work he was doing at the time with Professors Meyer at Loyola (Chicago) and Olkin at Stanford. It is now commonly referred to as the Kaiser–Meyer–Olkin Measure of Sampling Adequacy (KMO-MSA). It is calculated routinely, for example, in the heavily used SPSS and SAS factor analysis programs; also, functions for KMO-MSA appear in R, such as in the `psych` package developed by Bill Revelle.

The particular KMO-MSA computed by SPSS and various R functions, is not the exact same measure given in Kaiser (1970); instead, it is a modification meant to improve stability given in Kaiser and Rice (1974). In this latter article, Kaiser and Rice suggest that the same verbal scale used for the index of factorial simplicity in Kaiser (1974), also be used “as is” for KMO-MSA.⁵ So, apparently what has happened over the last four or so decades, is that when the value of the KMO-MSA index is reported (usually taken directly from SPSS output), the reference given for it is most often Kaiser (1974), the factorial simplicity paper, and not the correct citation of Kaiser and Rice (1974), which is the Version IV Little Jiffy article.

The reason for the citation mixup and allowing Kaiser (1974) to become the inappropriately highly cited paper it has developed into, may be due in part to an article by Charles Dziuban and

⁵It must have been intended as a quick fix, because the rationales for KMO-MSA and IFS are quite different, and only the label “unacceptable” for a value of 0.50 or less was substantiated (albeit with a different argument). Further attempts to calibrate the KMO-MSA scale came later (see Meyer, Kaiser, Cerny, and Green, 1977, or Cerny and Kaiser, 1977), but too late to have much impact.

Edwin Shirkey that appeared in *Psychological Bulletin* in 1974. In discussing KMO-MSA, the following sentence appears (p. 359): “Kaiser’s (1974) present calibration of the [KMO-MSA] index is as follows:”; what then follows are the exact same verbal labels given earlier and used by Kaiser (1974) to characterize his index of factorial simplicity. No reference, however, appears for Kaiser (1974) in the bibliography for Dziuban and Shirkey (1974), but one is given for an “in press” piece by Kaiser and Rice. So, what apparently has happened when a KMO-MSA value is reported, the reference for it is incorrectly given as Kaiser (1974) instead of the correct reference of Kaiser and Rice (1974)—an article, by the way, that did not appear in *Psychometrika*.⁶

The mixup of using Kaiser (1974) for Kaiser and Rice (1974) may be a miscitation phenomenon that is difficult to correct. Two popular SPSS-related user manuals, for example, make this citation error: the *SPSS Survival Manual* by Julie Pallant, and *Discovering Statistics Using SPSS* by Andy Field; the same citation error is also made in other books by Andy Field, such as in *Discovering Statistics Using SAS*. In hindsight, it is a bit surprising that such a miscitation wasn’t caught earlier by one of the authors of these secondary SPSS user manuals. Not a single statistical package (SPSS, SAS, SYSTAT, Matlab, or R) computes Kaiser’s index of factorial simplicity, irrespective of how good or bad the index might be. Several statistical packages, as noted before however, do compute KMO-MSA (SPSS, SAS, and Bill Revelle’s `psych` package for R).

When Bill Revelle was asked why he didn’t include a function for Kaiser’s index of factorial simplicity in his `psych` R package, he said that he instead included a function for what he thought was a better index (of factorial complexity) due to Hofmann (1978). He also noted that he was never tempted to use Kaiser (1974) in reference to his KMO-MSA function; he gave the three (correct) citations for it of Kaiser (1970), Kaiser and Rice (1974), and Dziuban and Shirkey (1974).

UNIVERSITY OF ILLINOIS
LEIDEN UNIVERSITY

Lawrence Hubert
Willem Heiser

References

- Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, 18, 23–38.
- Cerny, B. A., & Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research*, 12, 43–47.
- Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin*, 81, 358–361.
- Hofmann, R. J. (1978). Complexity and simplicity as objective indices descriptive of factor solutions. *Multivariate Behavioral Research*, 13, 247–250.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200.
- Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika*, 35, 401–415.
- Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, 34, 111–117.
- Meyer, E. P., Kaiser, H. F., Cerny, B. A., & Green Jr., B. F. (1977). MSA for a special Spearman matrix. *Psychometrika*, 42, 153–156.
- Neuhauss, J. O., & Wrigley, C. (1954). The quartimax method: An analytical approach to orthogonal simple structure. *British Journal of Statistical Psychology*, 7, 81–91.
- Saunders, D. R. (1953). An analytic method for rotation to orthogonal simple structure. *American Psychologist*, 8, 428 (Abstract).

⁶According to Google Scholar, Kaiser and Rice (1974) has 1578 citations as of 4/1/2016. This is a substantial number but nowhere near the 6043 citations for Kaiser (1974), almost all of which should be reallocated to Kaiser and Rice (1974).

AKAIKE, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317–332. (3852 citations in Google Scholar as of April, 2016)

In 1986, as the President elect of the Psychometric Society, I was given \$1000 as a President's discretionary fund to support the Annual Meeting of the Society to be held in Toronto. Without hesitation, I decided to spend the money to bring Professor Akaike from Japan, who was then the Director General of the Institute of Statistical Mathematics in Tokyo. He was already famous internationally for his ground breaking idea of using AIC (Akaike Information Criterion) for model selection, and I thought it was opportune to invite him to the meeting as the keynote speaker. Ham Bozdogan (University of Virginia) and I (McGill University) also organized a symposium on the topic, in which several more AIC-related papers were presented. After the meeting, Ivo Molenaar, then the Editor of *Psychometrika*, asked us if we could put together the paper by Akaike and those presented in the symposium as a special section of the journal. The paper by Akaike referred to in the title of this note is the leading paper in the special section published in the September 1987 issue of *Psychometrika*. It has turned out that this paper is one of the top 10 most highly cited papers ever published in the journal. It is interesting to point out that this paper comes as only the third most highly cited paper among all AIC-related papers by Akaike; Akaike (1973, 1974) are the top two, and have attracted four to seven times as many citations as the *Psychometrika* paper. It is indeed phenomenal that papers dealing with statistical theory have attracted so many citations. The first paper on AIC (Akaike, 1973) was subsequently selected for the volume, *Breakthroughs in Statistics*, edited by Kotz and Johnson as one of the most influential papers ever written in statistics. Incidentally, another paper published in the special section (Bozdogan, 1987) has also attracted quite a large number of citations (over 2000). This paper emphasized foundational aspects of AIC, while Akaike's paper focussed on applications of AIC to a particular problem in psychometrics.

The high citation rates of these papers are undoubtedly due to the popularity of AIC as a model selection criterion, which in turn is related to its simplicity in use. AIC is defined as

$$\text{AIC}_\pi = -2 \log(L_\pi) + 2k,$$

where π indicates a fitted model and L_π is the likelihood of model π maximized over k parameters. It can be readily calculated, once the maximum likelihood of the fitted model is calculated. AIC is evaluated for each model in the set of competing models, and the model associated with the smallest value of AIC is selected as the best-fitting model. The first term of AIC measures the badness-of-fit of the model to the current dataset (the dataset used to estimate k parameters), which is supplemented by the second term measuring the complexity of the fitted model. In general, the badness-of-fit of the model can be reduced by merely employing more complicated models, which has to be discouraged to increase the predictability of the model for future observations. AIC is deemed to hit a good balance between the two opposing demands. This goal is achieved by measuring the closeness of the estimated model to the true model. The model chosen as the best-fitting model according to the minimum AIC criterion is thus the one closest to the true model among all competing models. See Takane (1994) for an illustration of how the exact form of the penalty term, $2k$, emerges in AIC.

In his *Psychometrika* paper, Akaike proposed to use AIC for determining the number of factors in factor analysis. Factor analysis is considered one of the most important contributions of psychometrics to statistics. It was quite natural that Akaike chose this problem to demonstrate the usefulness of AIC. (Akaike, in fact, admitted that he initially thought of AIC in the context of factor analysis.) At surface, this seemed like a straightforward application of AIC: simply calculate AIC with the number of factors systematically varied, and choose the minimum AIC solution. It turned out, however, that the problem was not as easy as first expected. As he increased the number of factors, he soon encountered improper solutions, in which negative estimates of unique variances

were obtained. To overcome the difficulty, Akaike proposed a maximum penalized likelihood (MPL) estimation, in which the (log) likelihood to be maximized was appended by a penalty term to keep the estimates of variances positive. This strategy was successful to some extent. The penalty term, however, had a parameter that regulated the amount of driving force that kept the variance estimates positive. This parameter was not treated as part of the optimization framework. Akaike only tried a few values for the penalty parameter and informally observed the consequences. It left some arbitrariness in his proposal.

There was also another problem in employing the MPL estimation. The penalty introduced in the MPL estimation affects the definition of AIC itself. It is generally believed that the effective number of parameters in the model has to be altered depending on the penalty term. A couple of years after Akaike's paper, Shibata (1989) proposed another criterion, called RIC (Regularized Information Criterion), that took into account the penalty term, and allowed an optimal choice of the penalty parameter. Introducing RIC seemed like a perfect solution for Akaike's problem. However, the closed-form evaluation of RIC required the penalty term defined by a summation indexed in the same way as the (log) likelihood term. When this was not the case, Shibata (1997) recommended the use of a bootstrap estimate of RIC. Unfortunately, this procedure required quite a bit of additional computations.

Akaike himself never produced user-friendly software for his proposed method, which nowadays is a must to popularize a statistical method. The high citation rate of his *Psychometrika* paper is probably due to the popularity of AIC in social science research in general, and not to the specific proposal he made in his paper. AIC has been proven useful in many situations other than determining the number of factors. For example, virtually all commercial software for structural equation modeling reports the value of AIC as one of the goodness-of-fit indices.

Akaike did not preclude the possibility of other information criteria. In fact, he originally used the acronym AIC to stand for "An Information Criterion," implying that there could be other criteria based on different rationales. Indeed, a host of other information criteria have subsequently been proposed, following Akaike's lead. Among those, perhaps the strongest competitor has been BIC (Bayesian Information Criterion) proposed by Schwarz (1978), defined by

$$\text{BIC}_\pi = -2 \log(L_\pi) + (\log N)k,$$

where N is the sample size. BIC was derived as an approximation to (minus twice) the log posterior probability of model π under noninformative prior probabilities of the model (Neath and Cavanaugh, 2001). As in AIC, the model corresponding to the smallest value of BIC is chosen as the best model. BIC tends to choose a more compact model than AIC because its penalty term is usually larger than that of AIC. As N increases, it is bound to choose the correct model with probability 1, if the correct model is included among the competing models. That is, BIC is consistent, while AIC has no such property. A large number of studies have been conducted to compare the relative performance of the two criteria (e.g., Vrieze, 2012). When applied to factor analysis with small to moderate sample sizes, BIC tends to underextract factors. Although the chance of missing a big factor is not high, if it happens, it can have serious detrimental effects on estimated parameters (Fava and Velicer, 1996). AIC, on the other hand, tends to overextract, although its effect is known to be relatively minor (Fava and Velicer, 1992).

AIC's tendency to overextract is still a nuisance to be avoided as much as possible. Ogasawara (2016) recently proposed a correction of AIC for its higher-order asymptotic bias. This statistic, called CAIC (Corrected AIC), is defined as

$$\text{CAIC}_\pi = -2 \log(L_\pi) + 2k + 4k^2/p(N-1),$$

where p is the number of observed variables. The additional term in CAIC tends to discourage overextraction of factors. A preliminary simulation study indicates that CAIC is more accurate than the original AIC in determining the correct number of factors.

UNIVERSITY OF VICTORIA

Yoshio Takane

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *The second international symposium on information theory*, (pp. 267–281). Budapest: Akademiai Kiado. (Also in S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics, Vol. 1. Foundations and basic theory* (pp. 610–624). New York: Springer, 1992.)
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345–370.
- Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research*, 27, 387–415.
- Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction in factor and component analysis. *Educational and Psychological Measurement*, 56, 907–929.
- Neath, A. A., & Cavanaugh, J. E. (2011). The Bayesian information criterion: Background, derivation, and applications. *WIREs Computational Statistics* 2012, 4, 199–203.
- Ogasawara, H. (2016). Bias correction of the Akaike information criterion in factor analysis. *Journal of Multivariate Analysis* 149, 144–159.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Shibata, R. (1989). Statistical aspects of model selection. In J. C. Willems (Ed.), *From data to models*, (pp. 501–522). Berlin: Springer.
- Shibata, R. (1997). Bootstrap estimate of Kullback–Leibler information for model selection. *Statistica Sinica*, 7, 375–394.
- Takane, Y. (1994). A review of applications of AIC in psychometrics. In H. Bozdogan (Ed.), *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach*, (pp. 379–403). Amsterdam: Kluwer Academic Publishers.
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17, 228–243.

2. Analysis of Proximities

KRUSKAL, J. B. (1964). Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika*, 29, 1–27; Nonmetric multidimensional scaling: A numerical method. 29, 115–129. (5773 citations in Google Scholar as of April, 2016)

Although the assignment was to write a commentary about the famous, highly cited Kruskal 1964 papers, it would hardly be fair if the topic was not described in the context of two other papers by Roger Shepard (1962), with 2309 citations in Google Scholar as of April 2016. The development of what is called *nonmetric multidimensional scaling* started at Bell Labs, Murray Hill, NJ. Before going into more detail, some of the earlier history of multidimensional scaling should be addressed as well, to place the breakthrough from the early 1960s in a proper context.

Multidimensional scaling (MDS) is the generic name of a class of techniques that aim to find a low-dimensional space, in which the distances between N points resemble as closely as possible a given or derived set of dissimilarities or proximities between N objects. Multidimensional scaling methods can be divided into two major approaches. Prior to the Shepard–Kruskal approach, MDS was mostly associated with Torgerson (1952, 1958), and an approach now called *Classical* MDS. As Shepard wrote in his 1980 paper in *Science* (p. 390):

Proposals that stimuli be modeled by points in a space in such a way that perceived similarity is represented by spatial proximity go back to the suggestions of Isaac Newton [...] However, little progress was made toward the development of data-analytic methods for the construction of such spatial representations on the basis of psychological data until the efforts of a group of psychometricians, beginning in the late 1930s at Chicago and subsequently moving to Princeton, culminated in the 1952 development by Torgerson of the first fully workable method of metric multidimensional scaling.

The prime incentive of the Shepard–Kruskal approach was the notion of nonmetricity. The interest in nonmetricity came from several perspectives, the most important one being viewed as a property that could be attributed to data, especially psychological data such as preference rank orders. The central aspect of the so-called nonmetric breakthrough is the accomplishment of a *metric* representation, such as configuration of points, from essentially *nonmetric* information, such as rank orders for pairs of objects. Shepard described his work as the merging of what used to be two phases (i.e., first obtaining tentative distances from empirically given data by some predetermined distance function, and then applying classical MDS to obtain a configuration of points) into one phase. The innovative step was to make no assumptions about the form of the distance function, except that its values had to be monotonic with the data. Shepard observed that forces tending to flatten the configuration were counteracted by forces tending to maintain monotonicity. To implement these ideas, Shepard developed a FORTRAN program to carry out an iterative process. His procedure seeks a low-dimensional configuration of points for N objects while working in an $N - 1$ dimensional space, where monotonicity is exactly realized. In the final low-dimensional space, however, departure from monotonicity could arise. From the Shepard (1980) *Science* paper (pp. 391–392):

At the Bell Telephone Laboratories, I began in 1960 to explore a new approach to multidimensional scaling, called “analysis of proximities,” that proved capable of overcoming the limitations of the earlier approaches. I used a one-stage iterative method (i) to adjust the positions of points in a space until the rank order of the interpoint distances was as nearly as possible the inverse of the rank order of the corresponding similarities, and (ii) to find the space of the smallest number of dimensions for which the residual departure from a perfect inverse ranking was acceptably small

[...] Following a few adjustments, on 17 March 1961 the iterative process with which I had been experimenting finally converged to its first stationary configuration (at just 2.33 p.m. EST, according to the computer log). From then on, results of surprising precision were regularly obtained.

The fact that departure from monotonicity did not play a central part in Shepard's procedure, inspired his colleague from Bell Labs, Joe Kruskal, to display MDS as a problem of statistical fitting, where a configuration of points is sought whose distances fit a nonmetric function of the dissimilarities best. The genius of the two highly cited 1964 *Psychometrika* papers is that they took Shepard's intuitive ideas and turned them into a reliable means of carrying out a non-metric multidimensional scaling. Kruskal began by defining an explicit least squares loss function to minimize, called *stress*, proposed a gradient method to obtain successor configurations, and introduced monotone regression of distances upon dissimilarities to obtain perfect monotonicity.

As is often the case in major scientific developments, similar objectives were pursued elsewhere, most notably by Guttman. Guttman's important 1968 paper postulates the rank-image principle as the fundamental mechanism guarding nonmetricity, while at the same time introducing his correction-matrix method to obtain successor configurations. At the risk of oversimplifying, disregarding the very interesting discussion on the differences between the Kruskal and the Guttman approach, we could say that Guttman's nonmetric transformation involves the permutation of the distances so they become monotonic with the dissimilarities, while Kruskal's transformation involves a monotone regression of the distances upon the dissimilarities.

The major contribution by Kruskal and Shepard to the development of MDS, and which turned the technique into the elegant data analysis method as it is appreciated nowadays, has been primarily recognized as transforming multidimensional scaling into a *nonmetric* method compared to the Classical *metric* MDS method. What is often not greatly acknowledged is that both Kruskal and Guttman proposed at the same time an alternative procedure for the *metric* multidimensional scaling problem. In Classical MDS, the approximation of dissimilarities by distances is channeled through scalar products, and involves an eigenvalue decomposition of a matrix $\mathbf{S} \equiv -\frac{1}{2}\mathbf{J}\mathbf{\Delta}^{(2)}\mathbf{J}$. Here, $\mathbf{\Delta}^{(2)} \equiv \{\delta_{ij}^2\}$ is a matrix with elements equal to the squares of the given or derived dissimilarities, δ_{ij} , between N objects, and $\mathbf{J}$ is a centering operator, which renders the resulting matrix $\mathbf{S}$ double-centered. From the eigenvalue decomposition of $\mathbf{S} = \mathbf{K}\mathbf{\Lambda}\mathbf{K}'$, a low-dimensional configuration $\mathbf{X}$ is found by taking $\mathbf{X} = \mathbf{K}_p\mathbf{\Lambda}_p^{1/2}$, with $\mathbf{K}_p$ the first p eigenvectors and $\mathbf{\Lambda}_p^{1/2}$ the square root of the first p eigenvalues, ordered from largest to smallest. Thus, $\mathbf{S}$ is approximated by the scalar product matrix $\mathbf{X}\mathbf{X}'$:

$$\mathbf{S} \equiv -\frac{1}{2}\mathbf{J}\mathbf{\Delta}^{(2)}\mathbf{J} \approx \mathbf{X}\mathbf{X}'.$$

This procedure is also known as the Young–Householder (1938) process. Since the method was proposed by Gower (1966) independently from Torgerson, Heiser coined the acronym of YOHOTOGO scaling. An important property of the scalar product approximation approach is that the resulting distances $d_{ij}(\mathbf{X})$ in low-dimensional space will always be smaller than their associated dissimilarities δ_{ij} .

By contrast, instead of using scalar products, Kruskal introduced a least squares loss function defined on dissimilarities and distances directly; it can be written (in its raw, squared, form) as

$$\text{STRESS} = \min_{\{\delta_{ij}\} \in \Gamma} \sum_{i=1}^N \sum_{j=1}^N (\delta_{ij} - d_{ij}(\mathbf{X}))^2,$$

where Γ denotes the set of all monotone transformations of the original dissimilarities $\{\delta_{ij}\}$. Finding the optimal transformations, usually called disparities and indicated by $\hat{d}_{ij}$, is done by *inner minimization*, while the *outer minimization* solves the metric scaling problem by minimization over $\mathbf{X}$ for given dissimilarities δ_{ij} .

The first Kruskal FORTRAN program was called MDSCAL. However, as Shepard (1980, note 21) writes:

A variety of computer programs of this general type have subsequently been developed [...] Perhaps the currently most versatile such program is the one, available from the Bell Telephone Laboratories, named KYST (after Kruskal, Young, Shepard, and Torgerson, from whose earlier programs it derives).

What is also noteworthy, is that the nonmetric breakthrough inspired a whole succession of computer programs that applied the same idea of monotonic transformation to the *variables* in multivariate data (instead of to the *dissimilarities* in proximity data). According to Doug Carroll, at Shepard's suggestion, Kruskal introduced his MFIT monotone regression algorithm into the first procedure fitting a two-way ANOVA model with no interaction terms but with optimal monotonic transformation of the dependent variable (Kruskal, 1965). This application was followed by nonmetric factor analysis in Kruskal and Shepard (1974), and in the Alternating Least Squares–Optimal Scaling system by Young, De Leeuw, and Takane (1980) that includes multiple regression, principal components analysis, and canonical correlation analysis. Monotonic regression was replaced by monotonic regression splines (see Ramsay, 1988), and the principle was applied independently, and outside the psychometric literature, in the Alternating Conditional Expectations algorithm (ACE) by Breiman and Friedman (1985) in JASA.

Last, but not least, we have to mention the Gifi system, developed by “the Dutch school of nonlinear multivariate analysis” (Buja, 1990), and described extensively in Gifi (1990). Later, nonlinear techniques were successfully extended and implemented in software for nonlinear data analysis in the IBM/SPSS Statistics package CATEGORIES. The latter package also includes very elaborate extensions of the original Kruskal et al. ideas for nonmetric multidimensional scaling and unfolding, which shows that the influence of the 1960s nonmetric breakthrough still carries on.⁷

LEIDEN UNIVERSITY

Jacqueline Meulman

References

- Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, 80, 580–619.
- Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, 1032–1069.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.
- Gower, J. (1966). Some distance properties of latent roots and vector methods used in multivariate analysis. *Biometrika*, 53, 325–338.
- Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 33, 469–506.
- Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society, Series B*, 27, 251–263.
- Kruskal, J. B., & Shepard, R. N. (1974). A nonmetric variety of linear factor analysis. *Psychometrika*, 39, 123–157.

⁷This note was written while the author was Visiting Professor at the Department of Statistics, Stanford University, in June–July 2016.

- Ramsay, J. O. (1988). Monotonic regression splines in action. *Statistical Science*, 3, 425–441.
- Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. II. *Psychometrika*, 27, 125–140; 219–246.
- Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210, 390–398.
- Torgerson, W. S. (1952). Multidimensional scaling: 1. Theory and method. *Psychometrika*, 17, 401–419.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3, 19–22.
- Young, F. W., De Leeuw, J., & Takane, Y. (1980). Quantifying qualitative data. In E. D. Lantermann and H. Feger (Eds.), *Similarity and choice* (pp. 150–179). Bern: Huber.

JOHNSON, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32, 241–254. (3755 citations in Google Scholar as of April, 2016)

One of the 10 most-cited *Psychometrika* articles is by Stephen Johnson, who was a young computer scientist working at Bell Labs at the time the article was written. Interestingly, this piece is Johnson's sole published work in the area of statistics and/or psychometrics. It was obviously influenced heavily by other Bell Labs personnel at the time, including several future Psychometric Society presidents. The introductory footnote to the article reads: "I am indebted to R. N. Shepard and J. D. Carroll for many stimulating discussions about this work, and for aid in preparing this paper."

The Johnson paper in *Psychometrika* is nicely written and accessible, and provides for the first time in the literature the key concept of an ultrametric for characterizing all hierarchical clustering schemes. Johnson's inspiration for developing the correspondence between ultrametrics and hierarchical clustering schemes was from the work of Joe Ward, a Personnel Research Psychologist working in the 1950s and 1960s on occupational analyses for the US Air Force. To quote Johnson (p. 242):

The notion of a hierarchical clustering scheme, the central idea of this paper, was abstracted from examples given by Ward [1963]. We first consider such schemes, and develop a correspondence between hierarchical clustering schemes and a certain type of metric.

To characterize more formally the basic problem posed by hierarchical clustering, suppose S is a set of n objects, $\{O_1, \dots, O_n\}$, and between each pair of objects, O_i and O_j , a symmetric proximity measure, p_{ij} , is available (that might possibly be constructed from a more basic object by variable data matrix). It is assumed that the proximity measure has a dissimilarity interpretation so larger proximity values correspond to more dissimilar objects. These proximity values are collected into an $n \times n$ proximity matrix, $\mathbf{P} = \{p_{ij}\}_{n \times n}$. Any hierarchical clustering strategy produces (using Johnson's terminology) a hierarchical clustering scheme; the latter is a sequence or hierarchy of partitions of S , denoted $\mathcal{P}_0, \mathcal{P}_1, \dots, \mathcal{P}_{n-1}$, from the information present in $\mathbf{P}$. In particular, the (disjoint) partition, $\mathcal{P}_0$, contains all objects in separate classes (Johnson's "weak" clustering); $\mathcal{P}_{n-1}$ (the conjoint partition) consists of one all-inclusive object class (Johnson's "strong" clustering); and $\mathcal{P}_{k+1}$ is defined from $\mathcal{P}_k$ by uniting a single pair of subsets in $\mathcal{P}_k$.

Generally, the two subsets chosen to unite in defining $\mathcal{P}_{k+1}$ from $\mathcal{P}_k$ are those that are "closest," with the characterization of this latter term specifying the particular hierarchical clustering method being used. We mention three of the most common options for this notion of closeness:

- (a) *complete-link*: the maximum proximity value attained for pairs of objects within the union of two sets [thus, the maximum link (or the subset "diameter") is minimized];
- (b) *single-link*: the minimum proximity value attained for pairs of objects where the two objects from the pair belong to the separate classes (thus, we minimize the minimum link); and
- (c) *average-link*: the average proximity over pairs of objects defined across the separate classes (thus, the average-link is minimized).

From the time of Johnson's 1967 paper (which mainly emphasized the single-link and complete-link criteria), it has been generally accepted that the complete-link criterion should be the default selection for the task of hierarchical clustering when done in the traditional agglomerative way that starts from $\mathcal{P}_0$ and proceeds step-by-step to $\mathcal{P}_{n-1}$. A reliance on the single-link criterion tends to produce "straggly" clusters that are not very internally homogenous or substantively interpretable. To quote Johnson (p. 252):

... the results of the Maximum Method [that is, complete-link] have appeared to be the more meaningful or interpretable. That is, the search for compact clusters (of small over-all “diameter”) have proved more useful than the search for internally “connected” but potentially long chain-like clusters.

As noted by many users over the years, the average-link choice seems to produce results that are the same as or very similar to the complete-link criterion but relies on more information from the given proximities; the complete-link criterion (or for that matter, the single-link criterion) depends only on the rank order of the proximities. Given the Bell Labs context in which the Johnson paper was produced (with Kruskal and Shepard as senior colleagues to Johnson and the recent development of *nonmetric* multidimensional scaling) it should not be surprising that Johnson emphasized only single-link and complete-link clustering. To quote (p. 253):

Sokal and Sneath [1963, p. 190] have pointed out that, in methods like our Minimum or Maximum Methods [that is, single- or complete-link], the merging of two clusters depends upon a single similarity value (viz., the least or greatest in the appropriate set). They suggest that, for greater robustness of the solution, it may sometimes be desirable to use some sort of average value instead. As we have already noted, *to base such a procedure upon averages of the more obvious types is to lose the invariance, sought here, under monotone transformation of the similarity values.* [emphasis added]

Johnson notes that the FORTRAN program he wrote to carry out both single-link and complete-link clustering (called `hiclust.f`, and still available at www.netlib.org) could be easily modified to include other clustering criteria such as average-link. Again, we give a relevant quote (p. 254):

Nevertheless, when this seems desirable, the methods described here can be (and, indeed, have been) modified to yield solutions intermediate between those obtained by these two extreme methods. J. D. Carroll (personal communication) has suggested an average method based upon medians which, of course, do have the desired property of monotone invariance.

As noted earlier, the seminal contribution of Johnson’s paper and the reason for its continued popularity is the characterization of a hierarchical clustering scheme in terms of what is called an ultrametric. We now turn to a more formal definition.

Given the partition hierarchies from any of the three criteria mentioned (complete-, single-, or average-link), suppose the values determined when the new subsets were formed (that is, the maximum, minimum, or average proximity between the united subsets) are placed into an $n \times n$ matrix, $\mathbf{U}$. In general, there are $n - 1$ distinct nonzero values that define the levels at which the $n - 1$ new subsets are formed in the hierarchy; thus, there are typically $n - 1$ distinct nonzero values present in an appropriately row and column reordered matrix $\mathbf{U}$ that characterizes the identical blocks of matrix entries between subsets united in forming the hierarchy.

Based on a matrix such as $\mathbf{U}$, the partition hierarchy can be retrieved immediately along with the levels at which the new subsets were formed. In fact, any (strictly) monotone (that is, order preserving) transformation of the $n - 1$ distinct values in such a matrix $\mathbf{U}$ would serve the same retrieval purposes. Thus, as one illustration, the $n - 1$ distinct values in $\mathbf{U}$ could be replaced by the simple integers, $(1, 2, 3, \dots, n - 1)$, and the partitions of the hierarchy could still be reconstructed easily. Generally, a matrix $\mathbf{U}$ that can be used to retrieve a partition hierarchy in this way is called an ultrametric (matrix).

A matrix $\mathbf{U}$ is called an ultrametric (matrix) if for every triple of subscripts, i , j , and k , $u_{ij} \leq \max(u_{ik}, u_{kj})$; or equivalently (and much more understandably), among the three terms, u_{ij} , u_{ik} , and u_{kj} , the largest two values are equal.

The last paragraph of Johnson's paper is very prescient about where his introduction of the ultrametric concept will lead. This last paragraph begins (p. 254):

Finally, a different kind of possible extension will be briefly indicated here, in the form of a presently unsolved problem. In Section I we saw that the construction of an HCS is equivalent to finding a metric which satisfies the ultrametric inequality. Given a similarity measure d , we would in general like to find the closest metric D which satisfies the ultrametric inequality — various measures of closeness could be used. ...To the author's knowledge, this problem is unsolved.⁸

So, as Johnson foreshadowed, it is now common in the literature to reformulate this hierarchical clustering task as that of locating a best-fitting ultrametric matrix, say, $\mathbf{U}^* = \{u_{ij}^*\}$, to the given proximity matrix, $\mathbf{P}$, such that the least squares criterion

$$\sum_{i < j} (p_{ij} - u_{ij}^*)^2,$$

is minimized. Nevertheless, Johnson's paper is a true classic, as is evidenced by the fact that about one-fifth of its citations were received in the most recent 4 years since its appearance almost 50 years ago.

UNIVERSITY OF ILLINOIS

Lawrence Hubert

References

- Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman.
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.

⁸This comment about being “unsolved” is somewhat of an understatement. The task of finding the “closest” ultrametric falls within the class of optimization problems now known as NP-hard, which includes all the old combinatorial chestnuts, such as the traveling salesman problem. Although Johnson was a computer scientist, he cannot be faulted for not proving or knowing this in 1967. The notion of a problem being NP-hard was not even introduced into the computer science literature until the early 1970s.

CARROLL, J. D. & CHANG, J. -J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart–Young” decomposition. *Psychometrika*, 35, 283–319. (3320 citations in Google Scholar as of April, 2016)

No doubt this paper was Doug Carroll’s most influential contribution. It introduced a method for estimating all parameters of the weighted Euclidean model for multidimensional scaling, in which the weights provide a simple way to describe individual differences in similarity data. That model had been originally proposed by Horan (1969), but after publication of the Carroll and Chang paper it became widely known as the INDSCAL model.⁹ Their method for finding parameter estimates for the weighted Euclidean model relied on the new CANDECOMP model and algorithm, which stands for CANonical DECOMPosition of N-way tables. Jih-Jie Chang had done all the programming.

It is important to appreciate the distinction between model and algorithm, both carrying the same name here, because they brought two distinct innovations. The CANDECOMP model is a generalization of the Eckart–Young decomposition mentioned in the title which was published in the first volume of *Psychometrika* (Eckart and Young, 1936), while the CANDECOMP algorithm is an extension of the iterative estimation scheme for principal components analysis, proposed by the econometrician Herman Wold (1966) under the name Nonlinear Iterative Least Squares (NILES). These novel ideas were presented in full generality, and they energized the emerging field of three-way analysis. That is why the Carroll and Chang paper is not only cited by quantitative psychologists, but also by a much larger three-way community.

In multidimensional scaling (MDS), the interest is to analyze the mutual similarity among stimuli in terms of a few basic dimensions; if we have m individuals, each judging the similarity of pairs formed out of n stimuli, the data can be collected in an $n \times n \times m$ three-way array. In earlier approaches, such three-way similarity data were usually averaged across individuals, on the assumption that the same basic dimensions would be operative in each of them, and that individual differences would be merely random variation, which could be ignored. One notable exception had been the approach proposed by Tucker and Messick (1963), which first factor analyzed the m similarity matrices (strung out into vectors) to form p factors of individuals called “Points of View” (PoV), and then performed p separate multidimensional scaling (MDS) analyses on the weighted average similarities (with the factor loadings as weights and the mean similarities reassembled into a matrix for each PoV).

Carroll and Chang started by clearly voicing earlier criticisms of PoV analysis: “Perhaps the most cogent criticism is that the method is little more powerful than doing separate scalings on the individual subjects—and it makes no explicit assumptions about possible or probable communality of the dimensional structures for different real or idealized individuals. It would be very surprising if the various configurations had no structure in common” (Carroll and Chang, p. 284). Then, they assumed again that one common set of r dimensions would be potentially operative in each individual, but now each dimension may have a different relative salience, or importance, for different people. Suppose x_{jt} denotes the coordinate value of the j th stimulus on the t th dimension of the common space ($j = 1, \dots, n$ and $t = 1, \dots, r$). Then, the “modified” Euclidean distance for the i th individual was defined as

$$d_{ijk} = \sqrt{\sum_{t=1}^r w_{it} (x_{jt} - x_{kt})^2}. \quad (1)$$

⁹The Horan paper was submitted to *Psychometrika* in September 1964, but because the author was killed in a road accident before the completion of his PhD work, his supervisor John Ross took care of finalizing its publication.

Here, the weights w_{it} represent the salience of dimension t for individual i (with $i = 1, \dots, m$). It is easily seen that if we define

$$y_{ijt} = w_{it}^{1/2} x_{jt}, \quad (2)$$

then the ordinary Euclidean distances based on the y -values defined in (2) produce the individual distances defined in (1). So, each individual is assumed to perceive distances in a space where the common dimensions have been (differentially) expanded or contracted, and ignored if $w_{it} = 0$.

The next step was to apply the classic Young–Householder–Torgerson transformation (Young and Householder, 1938; Torgerson, 1952) to the distances defined in (1); that is, for each individual separately, the row- and column-means of the matrix with elements $-\frac{1}{2}d_{ijk}^2$ are subtracted from its entries (called double centering). This transformation gives m matrices of scalar products with elements b_{ijk} related to the parameters of the model as

$$b_{ijk} = \sum_{t=1}^r y_{ijt} y_{ikt} = \sum_{t=1}^r w_{it} x_{jt} x_{kt}, \quad (3)$$

by using (1) and substituting (2). It is worth noting that Horan (1969) had already obtained result (3), and that he had used it for arguing that the average of the scalar products across individuals would result in an $n \times n$ matrix $\mathbf{B}$, of which the eigenvector–eigenvalue decomposition could be used to estimate the stimulus coordinates x_{jt} .

It was here that Carroll and Chang proceeded differently, using a clever maneuver. Instead of (3), they considered the more general model

$$z_{ijk} = \sum_{t=1}^r a_{it} b_{jt} c_{kt}, \quad (4)$$

which is the CANDECOMP model already mentioned. Any method to estimate the parameters of the model (4) could be used to get estimates of the parameters of the INDSCAL model, by just identifying $a_{it} = w_{it}$, $b_{jt} = x_{jt}$, and $c_{kt} = x_{kt}$. Presumably, the symmetry of the data would carry over to the fitted expected values having the property $\hat{z}_{ijk} = \hat{z}_{ikj}$ for all i, j and k , and that in turn would guarantee that the estimates of the matrices $\{b_{jt}\}$ and $\{c_{kt}\}$ coincide to yield one estimate of $\{x_{jt}\}$.

To switch from (3) to (4) was a stroke of genius, and the most original contribution of the paper. Note that in (3), there are terms that are quadratic in the x 's for $j = k$, while all terms in (4) are trilinear, implying that equation (4) is a linear model in the a 's when the b 's and c 's are fixed, linear in the b 's when the other two sets of parameters are fixed, and likewise for the c 's. Given some initial estimates of the three sets of parameters, the CANDECOMP algorithm iteratively solves for the parameters of one set given temporarily fixed values of the other two sets by linear regression, and continues with partial solutions until the process stabilizes. Then the INDSCAL algorithm is just the CANDECOMP algorithm with the identification stated before, with certain normalization constraints.

About the origin of the idea of the CANDECOMP model and algorithm, Carroll and Chang gave the following credits:

This method of “canonical decomposition” was suggested to us by a paper by Herman Wold (1966) [...], in which a related method of decomposition of two-way tables was discussed [...], involving what Wold calls a NILES (for “Nonlinear Iterative Least Squares”) procedure. In the same paper [...] Wold suggested the more general three-way model discussed here, but did not describe a computational scheme for this model (except for the special case of one dimension) (p. 312).

It may have occurred to the reader that in the one-dimensional case, least squares estimation under model (4) of the a 's when the b 's and c 's are fixed amounts to a series of simple linear regressions, while in the more-dimensional case one needs a series of multiple linear regressions—not quite a big conceptual leap, I would say. In addition, Carroll and Chang appeared to be unaware of other work that Wold and his co-workers had done, reported in Wold and Lyttkens (1969). They considered a whole range of estimation procedures under the new name NIPALS (Nonlinear Iterative Partial Least Squares), which was later abbreviated to PLS. Among these procedures, the full-blown CANDECOMP algorithm was described by Alf Israelson in a section called “Three-way (or second order) component analysis” (Wold and Lyttkens, 1969, pp. 32–33).¹⁰ It also became soon apparent that the CANDECOMP model had been conceived independently by Richard Harshman (1970) under the name PARAFAC (for parallel factors), for which Bob Jennrich had developed an algorithm that was identical to CANDECOMP (Wold and Lyttkens, p. 32).

The larger part of the Carroll and Chang paper consists of the following:

- (a) Extensive descriptions of two illustrative datasets collected by their colleagues Sandy Pruzansky and Mike Wish from Bell Labs.
- (b) Persistent discussions of the strong points of the INDSCAL model (unique and meaningful orientation of the axes, thus obviating the rotational problem; easier interpretation of individual differences).
- (c) Detailed comparisons with other work in three-way MDS and three-way component analysis, and several possible modifications of INDSCAL and CANDECOMP to accommodate these.

This emphasis on strong properties and flexibility of the methods, as well as on new perspectives that three-way data may bring over and above two-way data makes it a prototypical programmatic paper. It generated a lot of follow-up work about theoretical and technical aspects of three-way models and their extensions. But the high citation count is especially due to the fact that INDSCAL and CANDECOMP became the methods of choice in a wide variety of applications, not only in the behavioral sciences and psychometrics, but notably also in chemometrics and signal processing (see Kroonenberg, 2014).

LEIDEN UNIVERSITY

Willem Heiser

References

- Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 211–218.
- Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1–84.
- Horan, C. B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, 34, 139–165.
- Kroonenberg, P. M. (2014). History of multiway component analysis and three-way correspondence analysis. In J. Blasius & M. J. Greenacre (Eds.), *Visualization and verbalization of data* (pp. 77–94). Boca Raton, FL: Chapman and Hall/CRC.
- Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17, 401–419.
- Tucker, L. R., & Messick, S. (1963). An individual difference model for multidimensional scaling. *Psychometrika*, 28, 333–367.

¹⁰I am indebted to Pieter Kroonenberg for pointing out this reference to me.

- Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (pp. 391–420). New York: Academic Press.
- Wold, H., & Lyttkens, E. (Eds.) (1969). Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bulletin of the International Statistical Institute*, 43, 29–51.
- Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3, 19–22.

3. Properties of Covariance Matrices in Multivariate Analysis

CRONBACH, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. (28,307 citations in Google Scholar as of April, 2016)

By far, the most-cited article in *Psychometrika* is Lee Cronbach's famous study of coefficient α . I will address four questions: What makes the article special? Why do people cite the article more than other special articles? Was the article also influential beyond psychology? How does α fit in present-day psychometrics?

What makes the article special? In the first half of the twentieth century, psychometricians distinguished two practically useful types of reliability, which were coefficients of stability and coefficients of equivalence. Coefficients of stability are correlations between two test scores obtained with the same test on two different time points, administered with a time interval that serves the psychologist's purpose. The coefficient provides information about the stability of the attribute across the time interval. Coefficients of equivalence are correlations between two test scores based on different sets of items intended to measure the same attribute, and express the degree to which the different sets are interchangeable. Cronbach mentioned two additional coefficient types. One type is the coefficient of equivalence and stability that refers to the correlation between test scores on different sets of items, one set administered first and the other set administered after a useful time interval. The other type is the coefficient of precision, referring to the correlation between the scores on the same test administered twice in one session to the same group of persons. The latter coefficient is hypothetical and not realizable in practice. Cronbach's article is about coefficients of equivalence.

Coefficients of equivalence were determined in two ways. First, one computed the correlation between equivalent forms of a test. However, constructing equivalent forms was impractical when one only intended to use one form. In clinical practice, many clinicians would prefer to avoid burdening their patients with twice the number of items necessary for diagnosing them and rather use one test form. In research, one would rather use the testing time to collect data on additional tests than spend precious time duplicating ones test battery. Hence, few researchers pursued the equivalent-form approach. Second, as a surrogate of equivalent forms, one computed the correlation between the scores on two halves of the same test and then corrected for having obtained the coefficient of equivalence for only half the test rather than the whole test. This approach was the split-half method, only requiring one test form, and for reasons of efficiency, it became the dominant method for computing reliability. The method had two problems. First, several authors thought that the original method proposed independently by Spearman (1910) and Brown (1910) was based on unrealistically restrictive assumptions, and they proposed several, less restrictive, alternative methods. This raised the issue of which method to use. Second, one could split a test of realistic length into a huge number of paired halves, with each pair producing a different split-half reliability. These arbitrary differences raised the issue of which reliability to accept as final.

In the 1930s and 1940s, starting from different points of departure, several authors suggested coefficients resembling what later became coefficient α (Kuder and Richardson, 1937), or coefficients that were equal to the later coefficient α (Guttman, 1945; Hoyt, 1941), but no one explored in detail the relationship of these coefficients to reliability obtained using equivalent forms and paired halves. Cronbach (1951) did, hence producing his seminal 1951 contribution to psychometrics. He demonstrated two results, providing researchers with a tool he called coefficient α that did not need equivalent test forms, or decisions about which split-half method to use and which pair of halve tests to select. First, he demonstrated that coefficient α equals the mean of all possible split-half coefficients based on a method Guttman (1945) derived using realistic assumptions. Second, he argued that for most tests, even when the items in different test halves had different factor

structures, the distribution of all possible split-half values shows modest spread, thus rendering the distribution's mean, which is coefficient α , an efficient summary of the distribution because it captures most split-half values with only little imprecision. Thus, in an era when computers were unavailable, Cronbach demonstrated that filling out a simple equation provides one with the mean of the distribution of split-half reliabilities, which is a quick and usually safe alternative to having to compute at least a sample of split-half values to check whether their spread really was small, justifying the use of coefficient α .

An auxiliary result of studying different factor structures' effect on split-half values and concluding the effect was small, was the conclusion that different factor structures, provided they were typical of real tests, had little effect on coefficient α . Moreover, Cronbach concluded that the test's general or dominant factor (in his words, the first factor concentration in the test), was the main influence on α , and that group factors only affecting responses to subsets of items in addition to the general factor's influence had a much smaller impact. Thus, Cronbach concluded that α quantifies the dominant factor among the items, which led him to relate coefficient α to what he called a test's internal consistency. Cronbach then suggested to quantify internal consistency by means of the mean inter-item correlation and related it to psychological interpretability. Relating α to internal consistency became the much-appreciated second main contribution the article made to psychometrics, test construction, and test practice. Finally, Cronbach delimited α from concepts such as Loevinger's homogeneity and Guttman's reproducibility.

I speculate that these latter contributions stood in the shadow of α being put in the zenith of reliability theory and relating α to a test's internal consistency. Taken together, these major contributions and the minor contributions lend the article the appearance of an intellectual *tour de force*, replacing a hodge-podge of concepts and methods, many of which badly understood, with a surprisingly simple method that already existed but whose meaning researchers had not realized until Cronbach clarified it to them. His contribution was so convincing that it stood the test of time until the present day.

Why do people cite the article more than other special articles? Cronbach (1951, p. 300) wrote that the essential problem set in his paper was: "How shall α be interpreted?" No doubt the article discussed this topic but its real contribution, one Cronbach probably did not foresee at the time of writing the article, was that, in a research area where many people were active, it provided researchers with a simple method that replaced chaos with order of the simplest form—a coefficient. A great contribution indeed. Only few publications make contributions of this magnitude, solving generations' psychometric problems. Rare examples are Bryk and Raudenbush (1992, second edition in 2002) in the context of multilevel modelling (27,000+ citations) and Hu and Bentler (1999) in the context of structural equation modeling (32,000+ citations). (Both results retrieved from Harzing's Publish or Perish that also consults Google Scholar.)

Not only has the article about coefficient α been cited frequently, but Cronbach's total oeuvre in which the 1951 article is also the front runner, has been cited 84,308 times (until 9 March 2016; retrieved from Harzing's Publish or Perish). His number two most-cited publication, the seminal article with Paul Meehl (Cronbach and Meehl, 1955) on construct validity, received 8079 citations, a dazzling number that most of us can only dream of and beating *Psychometrika*'s number two most-cited article (6043 citations for Kaiser, 1974). Moreover, a total of 14 of Cronbach's publications received 1000+ citations and another 6 received between 500 and 1000 citations. Cronbach truly had and continues to have tremendous impact.

Was the article also influential beyond psychology? Table 2 is based on the Web of Science (data retrieved on 21 March 2016, 7058 citations in ISI journals), and shows one-third of the citations came from psychology journals and two-thirds from a large variety of other areas. This demonstrates that the 1951 article had a wide influence within and beyond psychology. In addition,

TABLE 2.
Frequencies and percentages of ISI journal citations of Cronbach (1951) in different research areas.

Research areas	Frequency	Percentage
Psychology	2432	34.5
Business Economics	871	12.3
Public Environmental Occupational Health	619	8.8
Psychiatry	589	8.3
Health Care Sciences Services	528	7.5
Education Educational Research	365	5.2
Neurosciences Neurology	342	4.8
Social Sciences Other Topics	311	4.4
Sport Sciences	247	3.5
Computer Science	235	3.3

the spread across many research areas renders it highly unlikely that only psychometricians cited the article, the vast majority citing must be researchers reporting α for their test scores.¹¹

How does α fit into present-day psychometrics? In 1951, the mathematics of classical test theory was not yet fully developed and psychometricians found it difficult to agree about the relation between α and reliability. For example, Cronbach (1951, p. 299) wrote “It has generally been stated that α gives a lower bound to ‘the true reliability’ – whatever that means to that particular writer.” Defining reliability as the correlation between two mathematically parallel tests, Novick and Lewis (1967) proved that coefficient α is a lower bound to the reliability (also, see Lord and Novick, 1968, p. 88). This definition comes close to what Cronbach called a coefficient of precision, expressing the sheer influence of random error on measurement. Later authors (Bollen, 1989; McDonald, 1998) introduced systematic error represented by group and specific factors, suggesting a factor analysis approach to reliability. Cronbach, Gleser, Nanda, and Rajaratnam (1971) introduced generalizability theory, suggesting coefficients expressing reliability of person measurement free of unwanted influences such as caused by different test forms and raters. This development led Cronbach (2004) to abandon coefficient α in favor of generalizability, but probably due to its relative complexity, the latter approach never became as popular as the former. Unlike Cronbach, however, many psychometricians, test constructors and researchers have remained faithful to coefficient α .

TILBURG UNIVERSITY

Klaas Sijtsma

References

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods (First Edition)*. Newbury Park, CA: Sage Publications.
- Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64, 391–418.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

¹¹The author wishes to thank Roger Schmitz for providing the data in Table 2.

- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6, 153–160.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–36.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, 2, 151–160.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1–13.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.

GREENHOUSE, S. W., & GEISSER, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24, 95–112. (3622 citations in Google Scholar as of April, 2016)

To obtain a good sense of the importance of this highly cited Greenhouse and Geisser paper from the 1950s, it is well to remember the state of statistical computation at that time. It was done mostly with mechanical calculators and not by electronic computers with stored programs as is common today. Thus, it was a major contribution to the conduct of applied statistics when some particularly difficult analysis task could be reframed so that it might be done computationally with existing mechanical calculators.¹²

Profile data as discussed in the Greenhouse and Geisser paper can be characterized by subjects within groups observed over a battery of tests (or one test, say, that is repeatedly measured over a number of occasions). Three questions are typically of interest:

- (a) Are the group profiles “parallel”?—Here, “parallel” refers to the group means being equidistant at each measurement occasion;
- (b) Assuming that profiles are parallel, are they also at the same level (that is, are they “coincident”)?
- (c) Assuming coincident profiles, are they also “horizontal”?

Adopting some of the notation of Greenhouse and Geisser, it is assumed that p tests, $x_1, \dots, x_p$, are given to each of n_k individuals ($k = 1, 2, \dots, g$) in each of g established groups. An individual i in group k is said to have the profile, $(x_{i1k}, \dots, x_{ijk}, \dots, x_{ipk})$, for $1 \leq i \leq n_k$, $1 \leq j \leq p$, and $1 \leq k \leq g$. The group k profile is represented by $(\bar{x}_{.1k}, \bar{x}_{.2k}, \dots, \bar{x}_{.pk})$. It is assumed that each individual profile is a random vector sampled from a p -variate normal distribution with arbitrary variance–covariance matrix. It is also assumed implicitly that the p (test) variables are commensurable (that is, the variables have the same metric). This last assumption allows meaning to be given to the question of whether the profiles have the same “shape” (here, profiles are said to have the same “shape” when they are parallel and the tests are commensurable).

The question of profiles having the same shape may be best explained using just two group profiles, $(\bar{x}_{.11}, \bar{x}_{.21}, \dots, \bar{x}_{.p1})$ and $(\bar{x}_{.12}, \bar{x}_{.22}, \dots, \bar{x}_{.p2})$. Letting $d_1, d_2, \dots, d_p$ represent the differences between group means for each of the tests, a necessary and sufficient condition for the two profiles to be parallel is $d_1 = d_2 = \dots = d_p$. But this equality of the differences actually means that there is no interaction between groups and tests. In other words, an evaluation of whether the group profiles have the same shape is really an evaluation of the group–test interaction.

By assuming the particular form for the variance–covariance matrix called compound symmetry (where the tests have equal variances and are equally correlated in pairs), the classical mixed model for g samples is generated. The resulting analysis-of-variance table would appear as follows:

Sources	df	Sum squares	Mean squares
Tests	$p - 1$	SS_{Tests}	MS_{Tests}
Groups	$g - 1$	SS_{Groups}	MS_{Groups}
Individuals (w. groups)	$N - g$	$SS_{\text{Ind (w. groups)}}$	$MS_{\text{Ind (w. groups)}}$
Group $\times$ Test	$(p - 1)(g - 1)$	$SS(G \times T)$	$MS(G \times T)$
Ind $\times$ Test (w. groups)	$(p - 1)(N - g)$	$SS(I \times T) \text{ (w. groups)}$	$MS(I \times T) \text{ (w. groups)}$

Based on this analysis-of-variance table, there are three F -tests for the three conjectures of interest:

¹²For another *Psychometrika* paper on approximate methods for the analysis of repeated measures (but written some 20 years after Greenhouse and Geisser), see the informative review by Huynh Huynh (1978).

- (a) Parallel (same shape) group profiles:
 $MS(G \times T)/MS(I \times T)$ (w. groups) $\sim F_{(p-1)(g-1), (p-1)(N-g)}$.

Here, the total number of observations, $N = (\sum_{k=1}^g n_k)$.

- (b) Coincident group profiles:
 MS_{Groups}/MS_{Ind} (w. groups) $\sim F_{(g-1), (N-g)}$.

This test is equivalent to first obtaining a single score for each individual by summing over tests and then performing a one-way analysis-of-variance on the resulting scores.

- (c) Horizontal group profiles:
 $MS_{Tests}/MS(I \times T)$ (w. groups) $\sim F_{(p-1), (p-1)(N-g)}$.

When the assumption of compound symmetry for the covariance matrix does not hold, the F -tests for parallel group profiles and for horizontal group profiles are no longer valid. In fact, both tests are too liberal and would reject the corresponding (null) hypotheses too often; or, in other words, the obtained p -values are too small. The F -test for coincident group profiles is not affected and remains appropriate because it is based on summed scores for individuals. The genius of the Greenhouse–Geisser paper is that it showed how to correct the degrees-of-freedom for the tests for parallel and horizontal profiles based on an estimated function, ϵ , obtained from the (sample) variances and covariances among the tests (which is no longer assumed to have a compound symmetry form). For the parallel profiles test, the F -distribution used would be $F_{\epsilon(p-1)(g-1), \epsilon(p-1)(N-g)}$; for the horizontal group profiles test, the F -distribution used would be $F_{\epsilon(p-1), \epsilon(p-1)(N-g)}$. Because $1 \geq \epsilon$, the numerator and denominator degrees-of-freedom are both discounted, making the tests more conservative (and, therefore, more appropriate when compound symmetry does not hold); when compound symmetry does hold, $\epsilon = 1$, and no discounting of the degrees-of-freedom occurs.

The Greenhouse–Geisser paper also includes a lower bound for ϵ (that is, $\epsilon > \frac{1}{p-1}$). Thus, it is possible to discount the degrees-of-freedom maximally, giving the most conservative tests we would need: use $F_{g-1, N-g}$ to test for parallel profiles; use $F_{1, N-g}$ to test for horizontal profiles. The argument then continues as follows: if one rejects with the maximally discounted degrees-of-freedom, rejection would also occur with any value of ϵ greater than its lower bound.¹³

As hinted at in the beginning of this note, the Greenhouse–Geisser paper was important for the analysis of profile data in an era when computation was rather primitive and done without electronic computers. The mixed model for multiple groups assuming compound symmetry leads to an analysis-of-variance table that was well within the numerical capabilities of the mechanical calculators of the time. More importantly, the various tests could be modified to mitigate the effects of having to make the compound symmetry assumption. As is now well known to anyone taking an applied multivariate analysis class (based, for example, on a text like Johnson and Wichern, 2002), profile analysis can be done in an alternative manner with various multivariate analysis-of-variance techniques on difference scores. This approach requires significant numerical effort involving sample variance–covariance matrices that soon becomes prohibitive without computers and currently available statistical routines (in SYSTAT, SPSS, SAS, R, and Matlab, for example).

There is still one very important use for the Greenhouse–Geisser approach even in the face of all the computational power now commonly available. It occurs in the situation where the number of individuals is less than the number of tests; in these instances, the multivariate procedures are impossible to carry out because the necessary degrees-of-freedom are zero or negative for some

¹³It should be noted that the Greenhouse–Geisser *Psychometrika* paper does not include proofs for the explicit form that ϵ should take nor for its lower bound. Formal demonstrations are given in a companion 1958 paper in *The Annals of Mathematical Statistics*, with the authors listed in the traditional alphabetical order (Geisser and Greenhouse, 1958). As is common for an *Annals* paper, the proofs are extremely cryptic and given with little or no interpolated verbal explanation or discussion.

of the F -approximations. Examples of profile analysis data where subjects are fewer than the number of observation occasions, abound in neuroimaging analyses done using fMRI data.

In closing we give two comments on the Greenhouse–Geisser paper from the authors themselves. The first is from Greenhouse on the paper being named a Citation Classic (Greenhouse, 1982):

Seymour Geisser joins me in expressing our pleasure in learning that our 1959 paper is now a Citation Classic. Each of us was aware that our procedures were being applied because of the many letters and calls we received over the years. In preparing for this statement, I reread the paper. I must say it reads very well and is quite lucid in its exposition. I believe we have the then-editors of *Psychometrika* to thank for this in that, contrary to editorial strictures currently imposed, they did not demand that we shorten the paper. The pace of the exposition and the examples presented made it possible for any interested reader to apply the methods we were describing.

It is interesting to note that Paul Horst and Dorothy Adkins were the two co-editors of *Psychometrika* when the Greenhouse–Geisser paper was reviewed and published.

The second, not-so-nice comment is by Geisser taken from an interview published in *Statistical Science* (Christensen and Johnson, 2007, pp. 627–628):

Wes: What kind of problems did you work on at NIH?

Seymour: Well, at NIH, with Sam Greenhouse, I wrote my most infamous paper. “Infamous,” I say because it wasn’t a very important or very hard paper. It was just a paper that seemed to have caught on with social scientists and some medical people. It was just this profile analysis paper which ended up being a citation classic, which means that it had a lot of citations. It still has more citations than all of my [other] papers altogether. [Laughs all around.]

Wes: This was the Greenhouse-Geisser paper?

Seymour: There are two papers. The first paper was in the *Annals*, which actually worked out all of the quadratic forms, their expectations, and the mathematics. And the second was in *Psychometrika* and that was the citation classic. That was just to show the methodology—how to use this. It wasn’t a very big deal. I worked much harder on other papers and I think I produced much better work. But *de gustibus non disputandum est* [there’s no accounting for taste].

UNIVERSITY OF ILLINOIS

Hans-Friedrich Köhn

References

- Christensen, R., & Johnson, W. (2007). A conversation with Seymour Geisser. *Statistical Science*, 22, 621–636.
- Geisser, S., & Greenhouse, S. W. (1958). An extension of Box’s results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 29, 885–891.
- Greenhouse, S. W. (1982). Citation classic commentary. *Current Contents/Social and Behavioral Sciences*, number 28.
- Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika*, 43, 161–175.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice-Hall.

Manuscript Received: 28 SEP 2016

Published Online Date: 29 NOV 2016