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 4/1/2016).

Charles Lewis

Ledyard R Tucker's most frequently cited scientific article, and one of the most frequently cited *Psychometrika* articles, is the paper he and I wrote in 1973 that proposed a goodness of fit index for factor analysis. This index (and important generalizations introduced by Bentler & 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, includes the index in their published results, and wants 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 Fig. 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.

_____

Insert Figure 1 about here.

_____

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, 1958), "An individual differences model for multidimensional scaling" (Tucker & Messick, 1963), "Determination of parameters of a functional relation by factor analysis" (Tucker, 1958), and "Evaluation of factor analytic research procedures by means of simulated correlation matrices" (Tucker, Koopman & 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" (Tucker & Lewis, 1973, p. 9). In other words, only a limited role is proposed for the index, with the emphasis placed on the meaningfulness of the results.

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 propose a "reliability" measure. Essentially, they define 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 chi-square, the value obtained for the estimated reliability is only .73. A second example, based on a larger data set, yields a significant chi-square value $\left(\chi^2_{28} = 126.76,\ p < 0.0001\right)$ 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" (Tucker & Lewis, 1973, 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 chi-square 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 $N-1$[1], it is analogous to a mean square, and is denoted by M_m:

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

Its expectation is written as the sum of two components

$$E(M_m) = \delta_m + \varepsilon_m. \tag{2}$$

[1] We 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 noncentrality parameter) and ε_m is a component associated with sampling variation. For the null case when $\delta_m = 0$, we have

$$E\left(X_m^2 \middle| \delta_m = 0\right) = df_m, \tag{3}$$

so

$$E\left(M_m \middle| \delta_m = 0\right) = E\left(\frac{X_m^2}{(N-1)df_m}\middle| \delta_m = 0\right) = \frac{1}{N-1}. \tag{4}$$

This allows us to write

$$\varepsilon_m = \frac{1}{N-1}. \tag{5}$$

We may also obtain an unbiased estimator for δ_m:

$$\hat{\delta}_m = M_m - \frac{1}{N-1}. \tag{6}$$

Rewriting this estimator gives

$$\hat{\delta}_m = \frac{X_m^2}{(N-1)df_m} - \frac{1}{N-1} = \frac{X_m^2 - df_m}{(N-1)df_m} \tag{7}$$

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

$$\rho_m = \frac{\delta_0 - \delta_m}{\delta_0}. \tag{8}$$

We may write the estimator proposed in our paper as

$$\hat{\rho}_m = \frac{\hat{\delta}_0 - \hat{\delta}_m}{\hat{\delta}_0} = \frac{M_0 - M_m}{M_0 - 1/(N-1)} = \frac{X_0^2/df_0 - X_m^2/df_m}{X_0^2/df_0 - 1}. \tag{9}$$

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" (Tucker & Lewis, 1973, 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 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!

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. Paper presented at Psychometric Society Meeting, Iowa City, IA.

Tucker, L. R (1958). An inter-battery method of factor analysis. *Psychometrika, 23*, 111-136.

Tucker, L. R (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23*, 19-23.

Tucker, L. R, & Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika, 28*, 333-367.

Tucker, L. R (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*, 279-311.

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.

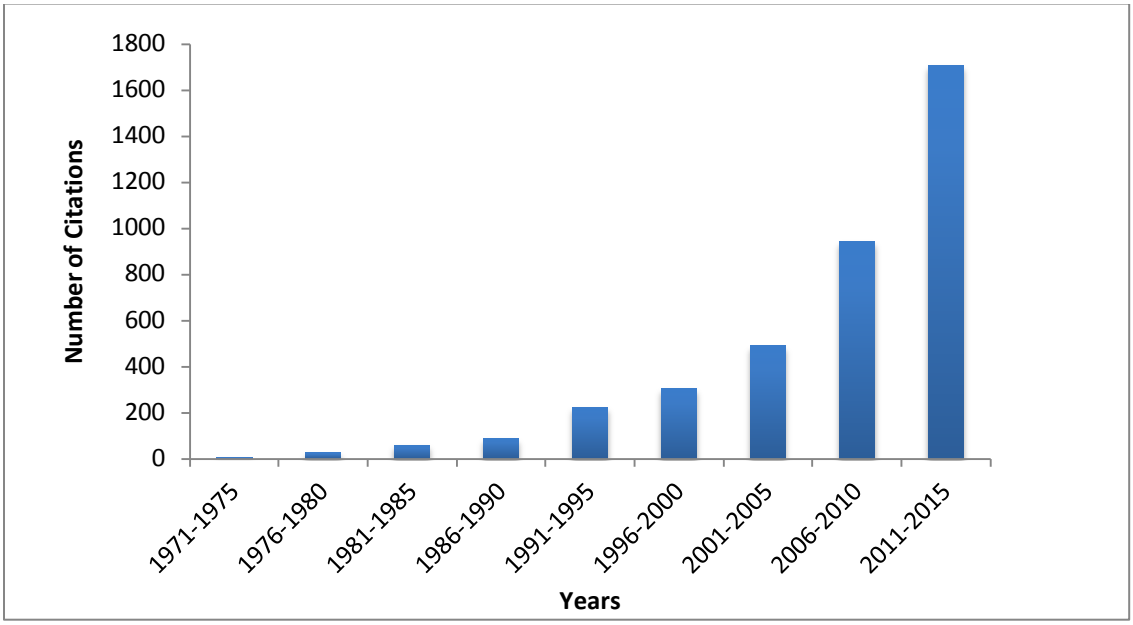

Figure 1. Numbers of citations for Tucker & Lewis (1973) article in five-year intervals.