Notes on the Steiger-Lind (1980) Handout

James H. Steiger

University of British Columbia

The computer program EzPATH, published in 1989, contained details of a confidence interval approach to assessment of model fit in covariance structure analysis. The EzPATH manual, and several published articles by myself and others have made reference to the Steiger and Lind (1980) presentation as the primary source for several influential ideas. As a result, I have received numerous requests for the handout which I distributed at that presentation.

The original handout was produced, under considerable time pressure, with a TRS-80 microcomputer. The computer had only uppercase letters, and the handouts were printed on a mimeograph. Perhaps the poor typographical quality contributed to the lukewarm reception the presentation received. Perhaps I tried to say too much in 25 minutes.

To aid the modern reader, I have corrected 2 typographical errors in the original handout. (John C. Lind was "John M. Lind" in the handout, and the second factor pattern on page 3 was mistakenly labelled "$p = 10$, $m = 2$" in the original handout.) Two other typographical omissions were pencilled in on most copies at the time of the original presentation (a 1/2 was pencilled in in the equation for ν on the first page, an N was placed next to $\mathbf{S}^{-1}$ in the 5th line from the bottom of page 6.)

I have also typeset the paper, using upper and lower case, subscripts and superscripts, and changing a few Arabic letters to their obvious Greek counterparts. The actual wording and mathematics are identical to the original version, as are the tables.

Please: remember that this was a handout to take the place of slides. It was not designed to be a complete manuscript, but it contains enough detail to document clearly the time frame for certain ideas. If you examine the handout, you will discover the following things in this one presentation:

- The largest Monte Carlo study ever performed on statistically-based number of factors rules. This was true then, and I believe it remains true now.

- The only Monte Carlo study to ever produce a coherent statistical model for *predicting the performance of the entire sequential LRT procedure.* The model was supported by two conjectures which I later was able to prove only with extensive help from my mathematically more gifted colleagues Michael Browne and Alexander Shapiro. One conjecture was that the noncentral chi-square approximation would hold. The Monte Carlo evidence was extremely convincing, but I lacked a proof in 1980. Proving this required the "population drift" argument found in later papers by Shapiro and others. The second conjecture involved approximating the covariance between adjacent sequential chi-square tests on the same data.

We were able to produce relevant proofs to tie up these loose ends only 5 years later in the *Psychometrika* paper by Steiger, Shapiro, and Browne (1985).
In this study, I presented models for predicting the performance of number of factors rules. I showed that these models worked, thus rendering future expensive Monte Carlo studies essentially unnecessary.

- Introduction of the notion of estimating the noncentrality parameter as a way to assess *population* fit.

- Introduction of the notion of noncentrality *interval estimation* as a way to resolve the paradoxes inherent in hypothesis testing logic.

- Introduction of the idea of *correcting for model parsimony* via the RMS index. This index is referred to as the RMSEA index by Browne and Cudeck.

The attached handout was distributed to all in attendance at the 1980 talk. Attendees included Rod McDonald (who chaired the session), Peter Bentler, Norm Cliff and Michael Browne (both of whom raised interesting questions during the talk), Joseph Kruskal, and Jim Ramsay (who offered suggestions for refining the data presentation afterward).

An updated presentation on the RMS index was given at the November 1987 meeting of the Society of Multivariate Experimental Psychology in Vancouver. Attendees at that presentation included Jeff Tanaka (who gave a presentation on fit indices), Chip Reichardt (who responded very positively to the interval estimation idea), and Peter Bentler.

The original handout of the 1980 talk is still available if you require a copy, but most people will prefer to have this version.

I repeat -- Except for cosmetic changes documented above, this paper is identical to the original handout.

Statistically-Based Tests for the Number of Common Factors[*]

James H. Steiger
and
John C. Lind

University of British Columbia

The Criteria

a) *Sequential Likelihood Ratio Test*

Let p = the number of variables
m = the number of factors fitted
f = the maximum likelihood "badness of fit" function
N = the number of observations
$k = N - (2p + 11)/6 - 2m/3$

Then $U = kf$ is distributed asymptotically as a chi-square variate with $\nu = 1/2((p - m)^2 - p - m)$ degrees of freedom.

The sequential LRT procedure selects as m^*, the estimated number of factors, the minimum m for which U is "not significant" at the .05 level.

b) *Akaike's "AIC" criterion.*

Set m^* equal to that m for which

$$U - 2\nu \text{ is a minimum.}$$

c) *Schwarz's "BIC" criterion.*

Set m^* equal to that m for which

$$U - (\text{Ln } N)\nu \text{ is a minimum.}$$

[*] Handout for a talk presented at the Annual Spring Meeting of the Psychometric Society, Iowa City, May 30, 1980.

The Design

The Monte Carlo experiment was essentially a 4 way factorial design with the following 4 factors:

a) Type of factor pattern (2 levels)

“Strong”
“Weak”

b) Size of correlation matrix (*p*) (2 levels)

$p = 10$
$p = 15$

c) Number of common factors (*m*) (3 levels)

“small” ($m = .2p$)
“medium” ($m = .4p$)
“maximum” (m = the value for which $\nu = 0$, i.e., 6 when $p =10$,
10 when $p = 15$)

d) Sample size (*N*) (5 levels)

$N = 5p, 10p, 15p, 20p, 25p$

Results are based on $r = 50$ replications when $p =10$, and $r = 20$ replications when $p = 15$.

Examples of Population Factor Patterns

"Strong" (*p* =*10*, *m* = 2)

.83	.08
.74	.15
.76	−.12
.70	.18
.88	−.14
−.12	.65
−.09	.70
.06	.81
.13	.75
.08	.82

"Weak" (*p* = 10, *m* = 4, *N* = 250)

.0325	.4171	.6353	.5209
.0504	.4122	.7201	.3302
.8451	.0694	.1508	.0910
.3512	.4661	.5794	.1539
.3084	.2843	.6256	.2875
.5768	.3317	.0272	.3957
.4390	.3783	.3676	.2812
.4495	.3471	.0584	.4172
.0323	.1619	.3323	.5589
.0585	.3373	.5224	.0994

Predicting the Performance of the Decision Criteria

a) The Basic Assumption

Let f^* be the value of the maximum likelihood “badness of fit” function calculated on the population correlation matrix for a given *m*. (Hence we might call f^* our “population badness of fit index.”) Let *k*, *U* be defined as on page 1.

Then we assume that $U = kf$ is distributed approximately as a non-central chi-square variate with v degrees of freedom and non-centrality parameter equal to $\lambda = kf^*$.

b) Predicting Results for the Sequential LRT Procedure

1) Factor analyze the population correlation matrix to obtain f^* values for each *m.*

2) Using assumption (a) above, estimate the power of the likelihood ratio test at $m = 1$, using, for example, the normal approximation to the non-central chi-square distribution given in Abramowitz and Stegun's *Handbook of Mathematical Functions,* Eq. 26.4.28.

3) Then $P(m^*= 1)$, the estimated probability that the nmber of factors will be selected to be 1, is given by (1 – Power at $m = 1$), while $P(m^*>1)$, i.e., the probability that a number of factors greater than 1 will be selected, is given by $(1- P(m^*= 1)$, or, the power value itself.

4) Next, estimate power at $m = 2$, as above.
Then $P(m^*= 2) = (1 -$ Power at $m = 2)(P(m^*>1))$.

5) Repeat the procedure at $m =3$, etc.

c) Predicting Results for AIC and BIC.

AIC and BIC criteria are based on relative values of *U*, and so dependencies between values of *U* at adjacent levels of *m* are much more important for these criteria than for the LRT procedure.

A consistent finding in the Monte Carlo research is that adjacent Chi-square statistics correlate very highly, i.e., about .70 - .95. We incorporate a value of .80 into our predictive model. (It should be noted that the overall predictions of the models are fairly robust to minor variations of this value.

(...ctd. on next page)

(...ctd from page 4)

In predicting these criteria, we assume that key decisions are made at adjacent levels, i.e., that a decision to reject $m = 2$ in favor of $m = 1$ will not be reversed at $m = 3$. (In fact, empirical data show that this is clearly not true, but that reasonably good predictions can still be made under this highly simplifying assumption.)

Procedure will be illustrated for AIC: generalization to BIC should be immediately obvious.

1) Recall first that, in deciding between $m = 1$ and $m = 2$, the AIC criterion chooses $m^* = 2$ if $U_1 - 2v_1$ is greater than $U_2 - 2v_2$. (where U_1 refers to the value of U at $m = 1$, and v_1 is its associated degrees of freedom, etc.) Hence, we estimate P(m^*= 1) as $\mathrm{P}\left[\left(U_1 - U_2\right) < 2\left(v_1 - v_2\right)\right]$. We estimate P($m^*$=1) as follows:

a) We estimate the means and variances of U_1 and U_2 from the non-central chi-square assumption.

b) We then assume, for simplicity, that $\left(U_1 - U_2\right)$ is approximately normal, with mean equal to the difference in the expected values of U_1 and U_2, and variance equal to the sum of the variances of U_1 and U_2, minus twice the estimated covariance (which is calculated assuming a correlation of .80 between U_1 and U_2).

c) Estimates of P(m^* = 1) and P(m^*>1) are then easily calculated from normal curve probability values.

d) At the next level, we estimate $\mathrm{P}\left[\left(U_2 - U_3\right) < 2\left(v_2 - v_3\right)\right]$, and multiply this value times P(m^*>1) to estimate P(m^* = 2), etc.

e) Estimates for the BIC procedure proceed in the same manner, except that the value Ln(N) is substituted for 2 as the multiplier.

Confidence Interval Procedure for Assessing Fit

The sequential LRT procedure is essentially flawed from a logical standpoint, primarily because it is an "accept-support" strategy, i.e., accepting the null hypothesis that population fit is perfect (i.e., that $f^* = 0$ for a given *m*) leads to the "desirable" decision to retain the current (hopefully small) number of factors.

This strategy is undesirable for several reasons. First, and perhaps foremost, the null hypothesis of perfect fit (i.e., that $f^* = 0$) for *m* less than the "maximum" as defined on page 2, is extremely unlikely to be true. On *a priori* grounds, we would expect the "true" number of factors to be, inevitably, this maximum value.

Consequently, the performance of the sequential LRT procedure is highly dependent on the power function of the LRT. "Too much" power leads, inevitably, to a large number of common factors, because the hypothesis of perfect fit will be rejected despite the fact that the actual f^* is quite close to zero. This seems undesirable, and no more logical than testing the null hyothesis that two means are equal when $N =$ 1,000,000.

On the other hand, if power is too low, we may (all too conveniently) decide that the number of factors is small when the true number is actually large.

It seems to us that the solution to the above problems is not to dispense with the hard-won advantages of having a statistical basis for our judgements, and simply state, rather vaguely, that "one should not retain significant but meaningless factors." Rather, we should clarify what we are trying to accomplish with our technology, and shift to a more appropriate rationale. To us, this means shifting from what is clearly an outmoded and suboptimal hypothesis-testing approach toward a confidence-interval based procedure.

By obtaining confidence interval estimates for f^*, we obtain all the information inherent in the hypothesis testing procedure, (i.e., the traditional test fails to reject when an appropriately defined confidence interval includes zero), and we also receive valuable information about the precision of estimate of f^* afforded by our data.

f^* is itself a rather natural index of "badness of fit" of a multivariate model, since it is approximately equal to the quadratic form

$$\mathbf{d'S^{-1}d}$$

where **d** is a vector of the elements of the residual correlation matrix, and $N\mathbf{S}^{-1}$ is the inverse variance-covariance matrix of the parameter estimates. Hence we can think of f^* as a "sum of squared, standardized residuals." Like the non-centrality parameter we computed in our analysis of variance classes, it is a rather natural, and easily comparable index of fit.

(....ctd)

We recall our assumption (a) on page 4. To the extent that this assumption is correct, we can obtain an interval estimate on f^* by first obtaining an interval estimate for λ, and then dividing its endpoints by k (or whatever multiplier was used).

Several procedures are available for obtaining confidence interval estimates on the non-centrality parameter of a chi-square distribution with known degrees of freedom. One very simple procedure uses the general interval construction method given in Cox and Hinkley's "Theoretical Statistics," page 212. Using the aforementioned normal approximation to the non-central chi-square distribution, we solve (iteratively, using a quasi-Newton approach) for those values of the noncentrality parameter λ which would yield a distribution in which the observed value of U falls at the .05 and .95 percentile points. Such values yield a 90% confidence interval. An alternative, more elegant procedure is given by Winterbottom in a recent article in the Journal of the Royal Statistical Society.

Hence, using a very simple procedure, programmable in a few lines, one may obtain a confidence interval estimate for λ and f^*. However, in assessing the comparative fit of multivariate models, such as the factor model at various levels of m, we must keep in mind that, as more factors are retained, some improvement in fit is inevitable, simply as a function of reduced degrees of freedom. Ideally, we should assess "badness of fit" by a method which distinguishes that improvement which is inevitable from that which results from choosing a model which is actually a better match to the structure of the data.

Hence, we propose that a more useful confidence interval is one calculated on the "root mean square standardized residual," i.e., a confidence interval on $\left(\frac{f^*}{\nu}\right)^{\frac{1}{2}}$. As we show in the numerical examples on the next page, such confidence intervals can alter or reinforce our perspective on some classical data sets.

Some Numerical Examples

24 Psychological Variables (*N* = 145)

m	DF	U	Low	High
1	252	617.5	.093	.114
2	229	418.9	.066	.090
3	207	292.8	.040	.069
4	186	220.8	.008	.055
5	166	183.7	.000	.050

Guttman's 1954 Circumplex Data (*N* = 710)

DF	U	Low	High	
12	27.1	.021	.063	Circumplex
4	16.1	.035	.098	Quasi-Circumplex

Table 1
Mean Number of Factors Obtained
(Predicted Values in Parentheses)

"Weak Factor" Condition

p = 10

		Chi Square	AIC	BIC
m = 2	*N* = 50	2.06 (2.05)	2.20 (2.03)	2.00 (1.99)
	100	2.02 (2.05)	2.12 (2.03)	1.94 (2.00)
	150	2.06 (2.05)	2.08 (2.03)	1.90 (2.00)
	200	2.02 (2.05)	2.10 (2.03)	1.98 (2.00)
	250	2.04 (2.05)	2.16 (2.03)	2.00 (1.98)
m = 4	*N* = 50	1.50 (1.46)	1.64 (1.52)	1.00 (1.00)
	100	1.98 (1.88)	2.24 (2.04)	1.12 (1.02)
	150	2.28 (2.32)	2.68 (2.40)	1.60 (1.71)
	200	3.44 (3.44)	3.68 (3.71)	2.90 (2.97)
	250	3.72 (3.84)	3.98 (3.97)	2.74 (2.65)
m = 6	*N* = 50	1.66 (1.81)	1.86 (1.49)	1.04 (1.00)
	100	2.88 (3.05)	3.38 (3.22)	1.68 (1.57)
	150	2.72 (3.20)	3.30 (3.06)	1.22 (1.11)
	200	3.38 (3.53)	3.74 (3.72)	1.80 (1.49)
	250	3.34 (3.43)	3.64 (3.38)	2.26 (2.42)

p = 15

		Chi Square	AIC	BIC
m = 3	*N* = 75	2.65 (2.55)	3.00 (2.89)	2.00 (2.00)
	150	3.05 (3.04)	3.10 (3.04)	2.90 (2.88)
	225	2.85 (3.00)	3.10 (3.04)	2.35 (2.27)
	300	3.05 (2.72)	3.25 (3.04)	3.00 (3.00)
	375	3.15 (3.05)	3.05 (3.04)	3.00 (3.00)
m = 6	*N* = 75	3.30 (2.55)	3.50 (2.95)	1.00 (1.04)
	150	4.35 (4.62)	4.75 (4.63)	2.60 (2.79)
	225	4.55 (5.02)	5.40 (5.02)	2.25 (2.22)
	300	4.45 (4.41)	4.90 (4.70)	3.90 (3.96)
	375	5.10 (5.30)	5.75 (5.59)	3.75 (3.73)
m = 10	*N* = 75	2.85 (3.52)	3.05 (2.26)	1.00 (1.01)
	150	3.70 (4.21)	4.30 (3.66)	1.95 (2.02)
	225	4.80 (5.62)	5.30 (4.73)	2.10 (2.04)
	300	5.50 (5.91)	6.00 (5.51)	1.80 (1.50)
	375	5.75 (6.39)	6.40 (5.82)	2.80 (2.76)

Table 2
Mean Number of Factors Obtained
(Predicted Values in Parentheses)

"Strong Factor" Condition

p = 10

			Chi Square	AIC	BIC
m = 2	*N* =	50	2.10 (2.05)	2.36 (2.03)	2.00 (2.00)
		100	2.04 (2.05)	2.20 (2.03)	2.00 (2.00)
		150	2.10 (2.05)	2.30 (2.03)	2.00 (2.00)
		200	2.08 (2.05)	2.30 (2.03)	2.00 (2.00)
		250	2.04 (2.05)	2.30 (2.03)	2.00 (2.00)
m = 4	*N* =	50	2.94 (3.06)	3.32 (3.08)	2.66 (2.77)
		100	3.12 (3.15)	3.48 (3.21)	3.00 (3.00)
		150	3.12 (3.23)	3.62 (3.37)	3.00 (3.00)
		200	3.32 (3.31)	3.67 (3.53)	3.00 (3.00)
		250	3.40 (3.40)	3.78 (3.66)	3.00 (3.00)
m = 6	*N* =	50	2.08 (2.62)	2.74 (2.11)	1.12 (1.03)
		100	3.20 (3.66)	3.80 (3.52)	1.56 (1.22)
		150	3.72 (4.00)	4.18 (3.99)	2.44 (1.73)
		200	3.94 (4.13)	4.30 (4.11)	3.04 (2.51)
		250	4.08 (4.19)	4.30 (4.17)	3.74 (3.24)

p = 15

			Chi Square	AIC	BIC
m = 3	*N* =	75	3.05 (3.05)	3.20 (3.04)	3.00 (3.00)
		150	3.10 (3.05)	3.10 (3.04)	3.00 (3.00)
		225	3.00 (3.05)	3.40 (3.04)	3.00 (3.00)
		300	3.15 (3.05)	3.40 (3.04)	3.00 (3.00)
		375	3.15 (3.05)	3.20 (3.04)	3.00 (3.00)
m = 6	*N* =	75	5.85 (5.72)	6.35 (5.75)	3.65 (2.54)
		150	6.00 (6.04)	6.10 (6.03)	5.70 (5.57)
		225	6.00 (6.05)	6.10 (6.03)	5.90 (5.98)
		300	6.05 (6.05)	6.15 (6.03)	6.00 (6.00)
		375	6.00 (6.05)	6.10 (6.03)	6.00 (6.00)
m = 10	*N* =	75	4.00 (4.60)	4.85 (3.35)	1.35 (1.29)
		150	5.60 (5.80)	6.15 (5.38)	2.85 (2.07)
		225	5.80 (6.39)	6.45 (6.21)	3.35 (2.82)
		300	5.95 (6.81)	6.55 (6.68)	4.60 (3.87)
		375	6.50 (7.11)	7.10 (7.00)	5.15 (4.61)