Reviews, critiques, and simulation studies are available for the chi-square and other measures of goodness of fit for structural equation models. The main exception is Hoelter's critical N (CN) statistic. In this article we present some of the properties of CN, explain why a fixed cutoff value for CN often favors large samples over small ones, and illustrate some of these characteristics with simulated and empirical data. We also describe the ambiguities involved in treating CN as a gauge of the power of the chi-square test.

Some Properties of Hoelter's CN

KENNETH A. BOLLEN
University of North Carolina at Chapel Hill
JERSEY LIANG
University of Michigan

A major controversial issue in applying covariance structure models is deciding whether a model adequately matches the data. A chi-square (Likelihood Ratio) test of statistical significance for a model is available. It is (N – 1)F where N is sample size and F is the value of the fitting functions at the final estimates for maximum likelihood and generalized least squares estimators (see Jöreskog and Sörbom, 1986; Bentler, 1985). The null hypothesis is that the covariance matrix of the observed variables, Σ, equals the covariance matrix predicted by the model, $\Sigma(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector that contains the unknown and unrestricted parameters of the model. Jöreskog (1969), Bentler and Bonett (1980), and many others have noted that even trivial

AUTHORS' NOTE: *We wish to thank Frank Andrews, Dick Campbell, Jon Hoelter, Scott Long, Willard Rodgers, Blair Wheaton, and the referees for their valuable comments. We alone are responsible for any errors.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 16 No. 4, May 1988 492-503

departures of Σ from Σ ($\boldsymbol{\theta}$) can lead to the rejection of the null hypothesis ($\Sigma = \Sigma$ ($\boldsymbol{\theta}$) if N is sufficiently large. The difficulty is that a model is typically meant as an approximation while the chi-square tests whether Σ exactly matches Σ ($\boldsymbol{\theta}$). The power of the chi-square test to detect false models increases with N, so that even minor departures of Σ from Σ ($\boldsymbol{\theta}$) often are detectable with a large sample. Alternatively with smaller samples, the power is reduced and a false null hypothesis is less likely to be rejected.

Given this situation, researchers have proposed other goodness-of-fit measures to supplement the chi-square test. These include the normed and nonnormed fit indices (Tucker and Lewis, 1973; Bentler and Bonett, 1980; Bollen, 1986), the Goodness of Fit Index (GFI) and Root Mean Square Residual (RMR) (Jöreskog and Sörbom, 1986), and the chi-square estimate divided by its degrees of freedom (Jöreskog, 1969). Recently, in *SMR* Hoelter (1983) has suggested another fit measure, Critical N (CN).

Reviews, critiques, and simulation studies are available for the chi-square test, the nonnormed index, GFI, and the RMR (e.g., Anderson and Gerbing, 1984; Boomsma, 1983). Saris and Stronkhorst (1984), Satorra and Saris (1985), and Matsueda and Bielby (1986) have proposed ways of estimating the power of the chi-square test against specific alternative models. But, Hoelter's CN has not been included in simulation studies and other than Hoelter's (1983) original paper little has been written on its characteristics.[1] The purposes of this note are to (1) present some of the properties of CN, (2) to explain why a fixed cutoff value for CN (e.g. CN $\geq$ 200) often favors large samples over small ones, and (3) to illustrate some of these characteristics with simulated and empirical data. We do not review and compare the other goodness-of-fit measures since this is readily available in the works cited above.

HOELTER'S CN

Hoelter (1983: 330) proposes CN as a means to "estimate the size that a sample must reach in order to accept the fit of a given model on a statistical basis." The CN value is[2]

$$\mathrm{CN} = \frac{\text{critical } \chi^2}{(\mathrm{T}/(\mathrm{N}-1))} + 1 \qquad [1]$$

where "critical χ^2" is the critical chi-square value at a selected α value and for degrees of freedom (df) equal to the model's df; and T is the chi-square estimate for the hypothesized model. Note that (T/(N − 1)) equals F, the value of the fitting function at the parameter estimates, so that equation 1 is equivalent to

$$\mathrm{CN} = \frac{\text{critical } \chi^2}{\mathrm{F}} + 1 \qquad [2]$$

Rearranging equation 2 shows the justification for the measure:

$$(\mathrm{CN} - 1)\mathrm{F} = \text{critical } \chi^2 \qquad [3]$$

The left-hand side of equation 3 is similar to the usual chi-square estimate of (N − 1)F except that CN replaces N. The right-hand side is the critical χ^2 value. CN is the sample size at which we would reject H_0: $\Sigma = \Sigma\,(\boldsymbol{\theta})$ at significance level α and the model's df, given the estimated value of F for the ML or GLS estimator. With multigroup analyses, Hoelter suggests that equation 2 be modified so that the +1 on the right-hand side be replaced by +G where G is the number of groups.

Hoelter's (1983) initial presentation as well as most applications of CN treat it as a goodness-of-fit measure that assesses the closeness of the sample covariance matrix (S) to the model predicted covariance matrix ($\hat{\Sigma}$). A few researchers (e.g., Matsueda and Bielbey, 1986: 130-131) have suggested that CN is a crude gauge of the statistical power of the chi-square test. We consider both viewpoints starting with CN as a goodness-of-fit measure.

A key question in using CN as a fit measure is what should be the cutoff value for an acceptable model? Hoelter (1983) tentatively suggests a criterion of CN ≥ 200G and all the applications that we are aware of follow his lead (e.g., Abbey and Andrews, 1985; Krause, 1987). Also, in practice, researchers seem to view models with CNs much higher than the cutoff as better than ones that are only slightly above it (see Hoelter, 1984: 258). The appropriateness of 200G or any other fixed criterion depends on the mean of the sampling distribution of the fit measure across different sample sizes. For instance, Anderson and Gerbing (1984) found the mean of the sampling distributions of GFI and AGFI (see Jöreskog and Sörbom, 1986) to increase with sample size. This implies that any fixed cutoff for GFI or AGFI would tend to favor large samples over small ones even if the same model is valid for all sample sizes. Hoelter (1983) does not describe the behavior of CN across sample sizes. However, his applications to a small (N = 23) and a large (N = 17,205) sample use the same 200G criterion.

If the mean of the sampling distribution of CN is roughly the same for a given model across sample sizes, then an invariant cutoff might be reasonable. To analyze the mean of CNs sampling distribution, we distinguish two cases. The first is when the structural equation model is valid and the second is when it is not. Of course, we recognize that perfectly valid models are rare, but we believe that part of understanding a fit measure is knowing its properties under ideal conditions as well as under less than ideal ones.

When H_0: $\Sigma = \Sigma(\theta)$ is correct, a criterion of CN $\geq$ 200 or any fixed cutoff value tends to favor large samples over small ones. To understand this consider equation 2. The critical χ^2 value in the numerator stays the same once the df and α are specified. The value of F is constant for a given model estimated for a given sample. However, the mean of the sampling distribution of F decreases with N. When $\Sigma = \Sigma(\theta)$ is true, the value of F goes to zero as N gets larger, that is, $F \rightarrow 0$ as $N \rightarrow \infty$ (Browne, 1982). If the values of the chi-square estimator approximate a χ^2, their mean equals the degree of freedom (df) for the model. This follows since the expected value of χ^2 is its df. Thus for sufficiently large samples the mean of F approximates df/(N – 1) (Bollen, 1987). As this relationship shows, the mean value of F for smaller samples is bigger than the mean value of F for larger samples for a given degrees of freedom. Smaller F values lead to larger CN values so the implication is that CN tends to be higher for large samples than for small ones. *Using a fixed cutoff of 200 leads to the* *frequent rejection of true models for small samples.* In fact, the identical valid model can be rejected if estimated with a covariance matrix from a small sample and accepted when the covariance matrix is from a large sample. A related characteristic is that for a valid model, CN goes to infinity as N goes to infinity. Thus, unlike some model fit indices (e.g., GFI, Δ), CN has no finite upper limit.

We constructed a simple example to illustrate the properties of CN for a valid model. Figure 1 is the path diagram of the model. Following Jöreskog and Sörbom's (1986) LISREL notation, the latent variable model is

$$\eta = \mathbf{B}\eta + \Gamma\xi + \zeta \qquad [4]$$

where ζ is uncorrelated with ξ, $(\mathbf{I} - \mathbf{B})$ is nonsingular, $E(\zeta) = 0$, and η and ξ are deviated from their means. The measurement model is

$$\mathbf{y} = \Lambda_y\eta + \epsilon \qquad [5]$$

$$\mathbf{x} = \Lambda_x\xi + \delta \qquad [6]$$

where ϵ, δ, and ζ are uncorrelated, ϵ is uncorrelated with η, δ is uncorrelated with ξ, $E(\epsilon) = \mathbf{O}$, $E(\delta) = \mathbf{O}$, and $\mathbf{y}$ and $\mathbf{x}$ are deviated from their means.

The population parameters are

$$\Lambda_x = \begin{bmatrix} .3 \\ .3 \\ .3 \end{bmatrix} \qquad \Theta_\delta = \text{diag}[1.8 \quad 1.8 \quad 1.8]$$

$$\Lambda_y = \begin{bmatrix} .3 & 0 \\ .3 & 0 \\ .3 & 0 \\ 0 & .3 \\ 0 & .3 \\ 0 & .3 \end{bmatrix} \qquad \Theta_\epsilon = \text{diag}[.81 \quad .81 \quad .81 \quad 1.66 \quad 1.66 \quad 1.66] \qquad [7]$$

$$\Phi = [5.0] \qquad \Psi = \text{diag}[1.8 \quad 3.69]$$

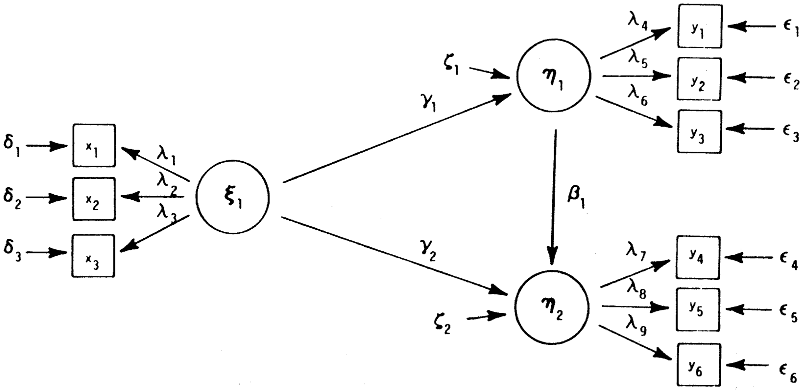

Figure 1: Model for Simulation

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ .3 & 0 \end{bmatrix} \quad \boldsymbol{\Gamma} = \begin{bmatrix} .3 \\ .3 \end{bmatrix}$$

These parameter values lead to a population implied covariance matrix, $\Sigma(\boldsymbol{\theta})$ (see Jöreskog and Sörbom, 1986: I.8). With this as the population covariance matrix, we simulated 50 sample covariance matrices for each sample size of 50, 75, 100, 200, 500, and 1,000. We used the IMSL GGNSM subroutine that is a multivariate normal random variable generator with a given covariance matrix. We exclude all nonconvergent and improper solutions. Table 1 lists the mean, standard deviation, number of CNs less than 200, and the minimum and maximum CN values for each sample size. The df = 24 and the critical χ^2 is 36.415 at the .05 level. It is evident that CN tends to increase with sample size. For instance, when N is 50 the mean CN is 77.5 while an N of 500 leads to a mean CN of 776.0. The range of CN values is wider and the minimum and maximum are much higher values for the big samples compared to the small ones. Using the cutoff of 200 would lead us to reject nearly all models estimated for sample sizes of 100, 75, and 50, and not to reject those with Ns of 200 or more. This is true even though the same model is valid for all samples. The dramatic change in the number of CNs less than 200 between sample sizes of 100 and 200 is not surprising since we would expect roughly 95% of the sample CNs to exceed 200 when N is 200 and α is .05. A much lower percentage of CNs should exceed 200 when N is only 100. Also, the standard deviation of CN greatly increases with N. An implication of this is that confidence intervals around CN should be wider for large Ns than for small ones. Finally, we noted some propensity for CN to produce outliers. However, the general characteristics we describe here (e.g., positive relation of mean CN and N) were not changed by their presence.

What happens when an incorrect model is estimated? Does the mean of the sampling distribution of CN still have a positive association with N? The answer to this question is not only unresolved for CN, but equally unresolved for other measures of goodness of fit such as the GFI, AGFI, and so on. Most Monte Carlo simulation works assume correct specifications when examining fit measures (e.g., Anderson and Gerbing, 1984). The behavior of CN under misspecification is likely to depend on the seriousness and nature of the errors, so that generalizations are difficult. In some situations CN in a poorly specified model may stay considerably below 200 for the typical sample sizes encountered in practice. To illustrate this we estimate a seriously flawed model for the same simulation data used above. The model forces all covariances between the nine observed variables to zero, with only their variances as free parameters (df = 36). This corresponds to the null model used in some fit measures (Bentler and Bonett, 1980; Bollen, 1986). The mean CNs for the sample sizes of 50, 75, 100, 200, 500, and 1,000 are 42.6, 58.9, 64.7, 88.3, 104.2, and 114.9, respectively. For this implausible model we still find a positive relation between the mean of CN and N, but all mean values are less than 200.

Between the extremes of a perfectly valid model and a totally unrealistic one lie most of the empirical applications of structural equations. Without much effort we have found published examples that illustrate a positive association between CN and N. It is reasonable to assume that these models contain some specification errors and fall between the extremes of a valid and null

TABLE 1
Mean and Standard Deviation of CN for Simulated Data with 50 Replications for Each Sample Size

Sample Size	Mean	Standard Deviation	Minimum	Maximum	Number of CN's < 200
50	77.5	21.3	35.6	155.6	50
75	128.9	31.6	74.2	217.5	48
100	153.3	35.6	100.9	266.1	47
200	307.3	115.8	174.4	796.4	3
500	776.0	191.2	378.7	1359.5	0
1000	1484.7	387.1	519.9	2546.0	0

model. The first example is an 11-item model that involves three first-order and one second-order factors (Liang, 1984). We divided the original large sample into six independent, random subsamples of 50, 75, 100, 200, 500, and 1,000. The CNs for these samples are 60.9, 78.0, 119.5, 210.9, 265.4, and 701.4, respectively. We can see that our assessment of fit would depend on whether we had a large or small sample.

Wheaton's (1987) recent work provides additional empirical illustrations of CN's relation to sample size for approximate models. For a confirmatory factor analysis in a sample of 132, he finds that CN never reaches 200 even though several versions of the model have high p-values for the chi-square estimates and appear to fit well by other criteria. Wheaton also estimates a covariance structure model in a sample of 2,568 and a random subsample of 355. He found that the CNs were generally higher in the large sample than in the small one even though the identical models were used. Furthermore, he argues that some of the models with "acceptable" (i.e., ≥ 200) CN values were clearly inadequate.

We do not expect that all empirical examples will show a moderate to strong positive association between CN and N. These examples, however, illustrate that this can occur for misspecified models as well as for correct ones. Thus using CN as a measure of goodness of fit may lead to a systematic bias in favor of models estimated with large samples. *One consequence may be that researchers seeking a CN ≥ 200 will tend to overfit models in small samples and underfit those estimated in large samples.* Furthermore, comparing the CNs may be misleading if the models are fit to different sample sizes.

Rather than measuring goodness of fit like the other fit measures, we could argue that CN is a crude measure of the *power* of the chi-square test. A big CN in a large sample may indicate that a model is adequate. A statistically significant chi-square estimate in this case is due to the excessive power of the chi-square test for large samples, so that even minor misspecifications are detectable. A small CN in a small sample is a warning to the researcher that even though the chi-square test leads to a high p-value, the test lacks the power to reveal even substantial specification errors because of the small sample size.

We see several problems in using CN as an indicator of statistical power. First, the power of a statistical test is the probability of rejecting a false null hypothesis. We know of no way to transform the CN value into a probability. Second, the power of a statistical test should be determined with respect to a particular alternative model. The power of the chi-square test usually varies depending upon the true alternatives for which the power is being assessed and it is highly unlikely that a single number can summarize the power of the chi-square test against all alternative models. Third, sample size often is an important determinant of the statistical power but it is not the only influence. For example, the number of indicators and the reliability of measures can affect statistical power (see Matsueda and Bielby, 1986). Exclusive attention to sample size can be misleading.

A large CN in a large sample does not always mean that a significant chi-square is due to excessive power rather than due to a substantively inadequate specification. The Wheaton (1987) examples show that CNs greater than 200 can occur with seriously flawed models. Hoelter (1984: 259) also rejects a model with a CN of 1,014 (CN cutoff = 400 since G = 2), judging it to be severely misspecified. In small samples, a large CN may be more an indicator of overfitting the data than of having a chi-square test with sufficient power.

CONCLUSIONS

Hoelter's (1983) CN is used as a measure of goodness of fit and sometimes as a crude gauge of statistical power. Our results show that the mean of the sampling distribution of CN is positively related to N for valid models and for some misspecified models. For these cases, any fixed cutoff value (e.g., 200G) of CN tends to favor models estimated in large samples. It is likely that for some misspecified models the positive association of CN and N is much weaker or near zero and that CN is less than 200G for the most typical sample sizes. But we do not know when this is true. The possible positive association of CN and N is important to remember when using CN as a goodness-of-fit measure. One might desire this property for an indicator of statistical power since large samples generally have greater power than small ones. But as we explained earlier, CN's relation to the statistical power of the chi-square test is ambiguous.

One response to these properties of CN is to develop a set of cutoff values applicable under different sample sizes or types of models. However, we think it would be difficult to form a typology of model types and to define unambiguously an adequate fit across various sample sizes.

It might be argued that researchers should only apply CN to large samples. This is not a fully satisfactory response for several reasons. One is that we still do not know when the sample is large enough. We have some simulation work on the behavior of the chi-square estimate for different sample sizes (Boomsma, 1983), but it is not clear whether these results hold for CN. Second, even for large samples, CN shows the same tendency to favor bigger samples over smaller ones for some models, so that the identical model will have a better CN value on average if the N is 1,000 compared to an N of 500. Again, any fixed cutoff point favors big Ns over small ones. Third, CN's variance can increase with N as we found with the first simulation model. In this situation confidence regions around CN will grow with sample size. Indeed the bounds can be quite large for big samples.

From an applied point of view we believe that researchers should not rely on CN or any other *single* measure of model fit, a position that Hoelter's (1983) paper also would support. If a researcher is concerned about the excessive power of the chi-square test in large samples, the procedures described in Saris and Stronkhorst (1984), Satorra and Saris (1985), and Matsueda and Bielby (1986) provide more direct means of estimating the power of the chi-square test with respect to specific alternative hypotheses than does the CN value. We recommend multiple-fit indices and when possible the analysis of the power of the chi-square test as aids in evaluating model fit.

NOTES

1 Matsueda and Bielby (1986: 130-131) have a brief treatment of CN.

2. As Matsueda and Bielby (1986) note, Hoelter's (1983) formula for CN uses an approximation for estimating the critical χ^2, so his formula differs from ours. Equation 1 or 2 is more accurate, particularly for models with low df.

REFERENCES

ABBEY, A. and F. ANDREWS (1985) "Modeling psychological determinants of life quality." Social Indicators Research 16: 1-34.

ANDERSON, J. and D. W. GERBING (1984) "The effects of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis." Psychometrika 49: 155-173.

BENTLER, P. M. (1985) Theory and Implementation of EQS: A Structural Program. Los Angeles: BMDP Statistical Software.

BENTLER, P. M. and D. G. BONETT (1980) "Significance tests and goodness-of-fit in the analysis of covariance structures." Psych. Bull. 88: 588-600.

BOLLEN, K. A. (1986) "Sample size and Bentler and Bonett's nonnormed fit index." Psychometrika 51: 375-377.

BOLLEN, K. A. (1987) "Structural equations with latent variables." (Unpublished manuscript)

BOOMSMA, A. (1983) On the Robustness of LISREL (Maximum Likelihood Estimation) Against Small Sample Size and Nonnormality. Amsterdam: Sociometric Research Foundation.

BROWNE, M. (1982) "Covariance structures," pp. 72-141 in D. M. Hawkins (ed.) Topics in Multivariate Analysis. New York: Cambridge.

HOELTER, J. (1983) "The analysis of covariance structures: goodness-of-fit indices." Soc. Methods & Research 11: 325-344.

HOELTER, J. (1984) "Relative effects of significant others on self-education." Social Psychology Q. 47: 255-262.

JÖRESKOG, K. G. (1969) "A general approach to the confirmatory maximum likelihood factor analysis." Psychometrika 34: 183-202.

JÖRESKOG, K. G. and D. SÖRBOM (1986) LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods. Moorseville, IN: Scientific Software.

KRAUSE, M. (1987) "Chronic financial strain, social support, and depressive symptoms among older adults." Psychology and Aging 2: 185-191.

LIANG, J. (1984) "Dimensions of the Life Satisfaction Index A: a structural formulation." J. of Gerontology 39: 613-622.

MATSUEDA, R. and W. BIELBY (1986) "Statistical power in covariance structure models," pp. 120-158 in N. B. Tuma (ed.) Sociological Methodology 1986. Washington, DC: American Sociological Association.

SARIS, W. E. and L. H. STRONKHORST (1984) Casual Modeling in Nonexperimental Research. Amsterdam: Sociometric Research Foundation.

SATORRA, A. and W. E. SARIS (1985) "Power of the likelihood ratio test in covariance structure analysis." Psychometrika 50: 83-90.

TUCKER, L. R. and C. LEWIS (1973) "A reliability coefficient for maximum likelihood factor analysis." Psychometrika 38: 1-10.

WHEATON, B. (1987) "Assessment of fit in overidentified models with latent variables." Soc. Methods & Research 16: 118-154.

Kenneth A. Bollen is an Associate Professor of Sociology at the University of North Carolina at Chapel Hill. His major research areas are in international development and statistical methodology. He is preparing a book titled Structural Equations with Latent Variables, *which will be published in John Wiley's Applied Probability and Statistics series.*

Jersey Liang is an Associate Research Scientist and Associate Professor at the Institute of Gerontology and the School of Public Health, respectively, at the University of Michigan. His research activities center on (1) quality of life among the aged, (2) health and aging, and (3) comparative aging.