RESEARCH METHODOLOGY IN STRATEGY AND MANAGEMENT VOLUME 5

RESEARCH METHODOLOGY IN STRATEGY AND MANAGEMENT

EDITED BY

DONALD D. BERGH
The University of Denver, USA

DAVID J. KETCHEN, Jr.
Auburn University, USA

United Kingdom – North America – Japan
India – Malaysia – China

JAI Press is an imprint of Emerald Group Publishing Limited
Howard House, Wagon Lane, Bingley BD16 1WA, UK

First edition 2009

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-84855-158-9
ISSN: 1479-8387 (Series)

Awarded in recognition of Emerald's production department's adherence to quality systems and processes when preparing scholarly journals for print

INVESTOR IN PEOPLE

CAUTIONARY NOTE ON CONVENIENTLY DISMISSING χ^2 GOODNESS-OF-FIT TEST RESULTS: IMPLICATIONS FOR STRATEGIC MANAGEMENT RESEARCH

Herman Aguinis and Erika E. Harden

ABSTRACT

This cautionary note provides a critical analysis of a statistical practice that is used pervasively by researchers in strategic management and related fields in conducting covariance structure analyses: The argument that a "large" sample size renders the χ^2 goodness-of-fit test uninformative and a statistically significant result should not be an indication that the model does not fit the data well. Our analysis includes a discussion of the origin of this practice, what the attributed sources really say about it, how much merit this practice really has, and whether we should continue using it or abandon it altogether. We conclude that it is not correct to issue a blanket statement that, when samples are large, using the χ^2 test to evaluate the fit of a model is uninformative and should be simply ignored. Instead, our analysis leads to the conclusion that the χ^2 test is informative and should be reported regardless of sample size. In many cases, researchers ignore a statistically significant χ^2 inappropriately to avoid facing the inconvenient fact that (albeit small)

Research Methodology in Strategy and Management, Volume 5, 111–120

ISSN: 1479-8387/doi:10.1108/S1479-8387(2009)0000005005

differences between the observed and hypothesized (i.e., implied) covariance matrices exist.

This cautionary note provides a critical analysis of a practice that is used pervasively by researchers in strategic management and related fields (e.g., human resource management, organizational behavior, applied psychology, and educational measurement). We hope that this critical analysis will provide information that will be useful to substantive researchers in their own work as well as journal reviewers who evaluate the work of others.

Initially, Herman Aguinis, in his role of Editor-in-Chief of *Organizational Research Methods*, noticed this practice in numerous manuscripts submitted for publication consideration. To confirm the pervasiveness of this practice, we subsequently conducted an in-depth review of the Method, Results, and Discussion sections for each of approximately 1,800 articles published between 2000 and 2006 in the following seven journals in strategic management and related fields (i.e., human resource management, organizational behavior, applied psychology, and educational measurement):

- *Academy of Management Journal*
- *Administrative Science Quarterly*
- *Strategic Management Journal*
- *Journal of Applied Psychology*
- *Personnel Psychology*
- *Applied Psychological Measurement*
- *Educational and Psychological Measurement*

We selected these journals because they arguably publish some of the most methodologically sophisticated and rigorous empirical research in strategic management and related fields (Podsakoff, MacKenzie, Bachrach, & Podsakoff, 2005). If a methodological practice is used frequently by researchers publishing in these journals, it is likely that it is used by researchers publishing in many other journals as well. As we illustrate in the next section using several examples, we identified the following practice that, based on our review, is sufficiently popular to be categorized as a statistical and methodological myth and urban legend: *ignore results based on a χ^2 goodness-of-fit test because sample size is "too large."* This is an important issue for strategic management research because the field has a long

tradition of studies using archival data, and many of these studies include large samples.

Next, we critically analyze this practice by answering the following questions: Where did it come from? What did the attributed sources really say about it? How much merit does it really have? Should we continue using this practice or should we abandon it altogether?

IGNORE RESULTS BASED ON A χ^2 GOODNESS-OF-FIT TEST BECAUSE SAMPLE SIZE IS TOO LARGE

In a covariance structure analysis, the null hypothesis is H_0: $\sum = \sum(\theta)$ (cf. Cheung & Rensvold, 2001). In other words, this null hypothesis tests whether the covariance matrix implied in the hypothesized model and the observed covariance matrix fit identically in the population. The statistic used for testing this null hypothesis is χ^2. In our review, we found that many authors argue that the χ^2 test is uninformative and should simply be ignored. This is because their sample size was "too large" and, therefore, their χ^2 test had "too much statistical power," which made it too easy to reject the null hypothesis that the sample-based data provide evidence of good fit. In other words, a common practice is to simply dismiss a statistically significant χ^2, which would suggest that the data do not fit a hypothesized model well. Note that if it is true that if the null hypothesis is false, χ^2 will be more likely to be statistically significant as sample size increases. This is an undisputed mathematical fact (Marsh, Hau, & Wen, 2004). In contrast, the methodological myth and urban legend is the practice to routinely dismiss a statistically significant χ^2 because it is "uninformative."

There are at least two important problems regarding this practice. First, it has been used for a wide range of samples sizes, in some cases as low as the mid-100s. Second, the argument that N is too large, therefore rendering the χ^2 test uninformative, is used when χ^2 is statistically significant (i.e., signaling poor fit). However, our review did not reveal *any* statements about N and statistical power (i.e., "statistical power may be insufficient to reject the null hypothesis") when the χ^2 is not statistically significant (i.e., signaling adequate fit). Thus, it seems that authors may use a self-serving double standard regarding the interpretation of χ^2 test results: Sample size is too large and the test should be ignored if results are statistically significant suggesting that the data do not fit the hypothesized model well; whereas

sample is just fine (and not "too small") if results are not statistically significant indicating that the data do fit the hypothesized model well.

Our review of the literature found that the belief that the χ^2 test is uninformative and should be simply discarded in the presence of "large" samples seems to be so pervasive that it reaches the category of myth and urban legend. For example, consider the following illustrations from strategic management and related fields. Goerzen and Beamish (2003) examined the performance of multinational enterprises and proposed splitting the concept of geographic scope into international asset dispersion and country environment diversity. After testing their hypothesized model, they found a statistically significant χ^2, but nevertheless concluded that "the research model fits the data well" because "this measure is excessively conservative and is biased against large samples (Bollen, 1989; Joreskog & Sorbom, 1981)" (p. 1300). Hessen, Dolan, and Wicherts (2006) noted that "... chi-square values are inflated by large total sample sizes. Therefore, in the case of the present sample sizes, chi-square difference results are of little use" (p. 239). Similarly, Allen, Van Scotter, and Otondo (2004) justified ignoring a significant χ^2 statistic by stating that "[t]he χ^2 statistic was significant ($\chi^2[30df] = 121.21$, $p<0.05$), but the χ^2 is sensitive to large sample sizes..." (p. 159). Likewise, as yet another example, Scullen, Mount, and Goff (2000) reported that "even excellent models typically yield statistically significant chi-square values when the sample size is large (Hu & Bentler, 1995)" (p. 963). Our literature review revealed numerous additional illustrations of similarly worded statements used to justify the dismissal of a statistically significant χ^2 (e.g., Ang & Huan, 2006; Davis & Finney, 2006; Kim, Cramond, & Bandalos, 2006; Schaufeli, Bakker, & Salanova, 2006, for examples of articles published in 2006 only). In short, researchers find a statistically significant χ^2 and, regardless of the specific sample size, issue a statement that this result will simply be ignored because N is "too large."

The examples above indicate that there is quite a bit of agreement among substantive scholars in strategic management and related fields about this practice. But, is this argument justifiable? What do the cited sources invoked in these articles say about the legitimacy of ignoring a statistically significant χ^2? What is a "large" sample size in the context of the χ^2 test? What are the negative consequences if the χ^2 is too sensitive and has "too much statistical power"? Let us examine what the cited sources really say about each of these issues.

Bollen (1989) is usually cited as one of the sources to support disregarding a statistically significant χ^2 because sample size is too large. In sharp

contrast to the way Bollen is cited, consider the following statement included in this book:

- "A third condition for $(N-1)F_{ML}$ to approximate a chi-square variate is that the sample be sufficiently large" (p. 267).

Bentler and Bonett (1980) is an article also frequently cited as a source regarding the effects of large sample sizes on the χ^2 test results. Consider the following statements from this frequently cited source:

- "In large samples virtually any model tends to be rejected as inadequate, and in small samples various competing models, if evaluated, might be equally acceptable" (p. 588).
- "While the chi-square test provides valuable information about a statistically false model, problems associated with sample size mitigate the value of the information that is obtained. The increase in ability to detect a false model with increasing sample size represents an important aspect of statistical power, but in the context of most applications in which the exactly correct model is almost certainly unknowable, this effect of sample size is a mixed blessing. Since the chi-square variate is a direct function of sample size, the probability of rejecting any model increases as N increases, even when the model is minimally false ..." (p. 591).
- "There is another problem. In many circumstances one would like to establish that the model provides a plausible representation of the data. In effect, a *non*significant chi-square value is desired.... This procedure cannot generally be justified, since the chi-square variate v can be made small by simply reducing sample size" (p. 591).
- "These difficulties [referring to the effects of N on the chi-square test] can be illustrated by two examples. McGaw and Joreskog (1971) reported an eight-factor exploratory factor analysis of 21 variables based on $N = 11{,}743$... the probability of the associated solution based on the tabled values of the chi-square distribution was less than 0.01.... However, in view of the large sample size, it is likely that no factor model with positive degrees of freedom could be found that would fit the data with probability greater than 0.05.... The converse problem is illustrated in a study by Bentler and Lee (1979). They studied the intercorrelations of four personality variables... in a sample of 68 children.... This solution yielded $v(35) = 43.88$. This value does not exceed critical cutoff values in the chi-square distribution.... However, in view of the small sample size, numerous competing models, if evaluated, might similarly be accepted" (pp. 591–592).

Hu and Bentler (1995), which is also one of the cited sources often invoked to support the notion that when N is too large the χ^2 test is uninformative, offered the following view:

- "...with the increased statistical power of the test afforded by a large sample, a trivial difference between the sample covariance matrix S and the fitted model $\sum$ may result in the rejection of the specified model" (p. 78).

In sum, comparing the statements found in the cited sources with how these sources are cited leads to several conclusions. First, although the cited sources refer to sample size being an issue to consider in interpreting χ^2 test results, some sources actually warn about having a sample that is *too small*. Large samples are needed not only for an accurate estimation of the fit of the data to the model, but also for the accurate estimation of the model's parameters (especially for maximum likelihood estimation). Second, although some of the cited sources conclude that a "large" sample size may lead to a statistically significant χ^2 even when the difference between the sample-based and the implied covariance matrices is trivial, we have been unable to locate conclusive information regarding what a "large" sample is. In their illustrations, Bentler and Bonett (1980) refer to $N = 11{,}743$ as large and $N = 68$ as small. The vast majority of published studies in strategic management and related fields include sample sizes closer to 68 than 11,748, so the large sample size problem (i.e., rejecting good models) may actually not be as pervasive as the small sample size problem (i.e., accepting poor models). Finally, although some authors refer to having "too much statistical power," this is not really a problem of the χ^2 test and, instead, the problem is in how to interpret the meaning of statistical significance (Aguinis, 2004; Cascio & Aguinis, 2005). The p value associated with the χ^2 statistic is the probability of observing the sample data, or data more deviant, given the condition that the null hypothesis $\sum = \sum(\boldsymbol{\theta})$ is true. Thus, a statistically significant χ^2 does not tell us whether the difference (if any) between $\sum$ and $\sum(\boldsymbol{\theta})$ is practically significant, and only tells us that it is unlikely (usually at a probability of $p<0.05$) that the null hypothesis is true.

IMPLICATIONS FOR STRATEGIC MANAGEMENT RESEARCH

Is it appropriate to dismiss a statistically significant χ^2 test and label the test as uninformative on the grounds that sample size is "too large"? This is a

very important question for strategic management research because the field has a long tradition of studies using archival data, and many of these studies usually include large samples. As is the case with any test of statistical significance, the probability of obtaining a statistically significant result increases as sample size increases (assuming that the population parameters are not exactly identical) (Aguinis, Beaty, Boik, & Pierce, 2005). This is a mathematical fact and it is a desirable property for a test statistic and therefore it is expected that the χ^2 will vary directly with N for incorrectly specified models (Marsh et al., 2004). Note, however, that if a study's sample is inordinately large, one may conclude that the sample-based covariance matrix is dissimilar to the covariance matrix implied by the hypothesized model in the population even if the difference between the matrices is miniscule and practically or scientifically insignificant. This characteristic is by no means unique to the χ^2 test. For example, a correlation coefficient $r = 0.05$ (i.e., variable X explains only 0.25% of the variance in the criterion variable Y) is statistically significant at the 0.05 level if $N = 1{,}600$. In the case of this correlation coefficient, we most likely would not conclude that, in spite of being statistically significant, $r = 0.05$ is scientifically or practically significant given that there is only 0.25% of variance explained in the criterion variable (Aguinis & Harden, 2009). Similarly, a statistically significant χ^2 based on an inordinately large N may not necessarily mean that the difference between the covariance matrices in the population is scientifically or practically significant.

Unfortunately, researchers seem to focus on the effects of what is categorized as a "large" sample size on the interpretation of the χ^2 test and often use this issue as justification for ignoring a result that indicates the data in hand do not fit their hypothesized model well. On the other hand, researchers seldom mention that a small N may be a problem and, in spite of a χ^2 that is not statistically significant, alternative untested models may fit equally as well (Bollen, 1989). This is unfortunate because it is not clear what a "large" sample size is. For example, authors have argued that a χ^2 test should not be used based on the large sample size argument with N's in the mid-100s. Can N of about 150 really be considered such a large sample that it leads the χ^2 test to detect scientifically and practically insignificant differences between the implied and sample-based covariance matrices in the population? Or, are some authors conveniently and self-servingly using this argument to avoid addressing the possibility that the data may not fit the hypothesized model well?

A reasonable question that could be asked is why is this practice so pervasive? We can only speculate on the reasons, but we suspect that some

authors may invoke this statistical myth and urban legend as a pre-emptive strike to counter a potential criticism from a journal reviewer when results do not turn out as predicted (e.g., a statistically significant χ^2 signals poor model fit). Another reason may be lack of proper statistical training of substantive researchers, as has been documented by several studies (Aiken, West, & Millsap, 2008). Regardless of the reason for invoking it, we emphasize that our focus is on a critical analysis of the practice, and not on specific authors who have used it. It is not our intention to point fingers and blame specific authors. Similar to other cautionary notes published elsewhere (e.g., Pierce, Block, & Aguinis, 2004), our goal is to raise researchers' awareness about the relative appropriateness of methodological and statistical procedures that will hopefully serve to foster more accurate practices in the future. This is particularly important in the case of substantive researchers and journal reviewers (i.e., those not specializing in measurement and statistics).

CONCLUDING COMMENTS

The question we addressed is: Is it true that when samples are "large," using the χ^2 test to evaluate the fit of a model is uninformative and can be simply ignored? The answer to this question is no. Our analysis leads to the conclusion that the χ^2 test is informative and should be reported regardless of sample size. We base this conclusion on the following points. First, it is the only test to assess the statistical significance of the difference between the implied and the observed correlation matrices. Other goodness-of-fit indexes exist (e.g., comparative fit index, normed fit index, and root mean square error of approximation), but they are not tests of statistical significance. Second, we do not really know what a "large" sample is. It seems disingenuous to use the same "large N" argument regardless of a study's sample size. Third, the presence of samples that are too small (leading to incorrectly accepting a model) seems to be more common in strategic management and related fields than the presence of samples that are too large (leading to incorrectly rejecting an incorrect model). Thus, it is likely that in published research in strategic management and related fields, some inappropriate models have been retained as adequate (due to small sample size). Fourth, even if a sample is inordinately large, which is not the most typical scenario in strategic management and related fields, the χ^2 test is informative because, if interpreted correctly, it provides information regarding the fit between the observed covariance matrix in relation to the covariance matrix in the

population underlying the hypothesized model. A statistically significant χ^2 tells us that the hypothesized and sample-based covariance matrices are not likely to be identical in the population, but does not tell us whether this difference is practically or scientifically important. Unfortunately, the argument that sample size is "too large" and, therefore, a statistically significant χ^2 test should be ignored, seems to be used sometimes as a rationalization for ignoring the result that the null hypothesis $\sum = \sum(\boldsymbol{\theta})$ (i.e., the hypothesized model is correct) has been rejected. In many cases, this argument is used inappropriately to avoid facing the inconvenient fact that (albeit small) differences between the observed and implied covariance matrices exist and a researcher's proposed model may actually be incorrect.

ACKNOWLEDGMENTS

We thank Gordon Cheung, Mark Gavin, Chuck Lance, and Bob Vandenberg for constructive comments on earlier drafts.

This research was conducted, in part, while Herman Aguinis held the Mehalchin Term Professorship in Management at the University of Colorado Denver and visiting appointments at the University of Salamanca (Spain) and University of Puerto Rico.

REFERENCES

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York, NY: Guilford.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94–107.

Aguinis, H., & Harden, E. E. (2009). Sample size rules of thumb: Evaluating three common practices. In: C. E. Lance & R. J. Vandenberg (Eds), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences* (pp. 269–288). New York: Routledge.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*, 32–50.

Allen, D. G., Van Scotter, J. R., & Otondo, R. F. (2004). Recruitment communication media: Impact on prehire outcomes. *Personnel Psychology, 57*, 143–171.

Ang, R. P., & Huan, V. S. (2006). Academic expectations stress inventory: Development, factor analysis, reliability, and validity. *Educational and Psychological Measurement, 66*, 522–539.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Bentler, P. M., & Lee, S. Y. (1979). A statistical development of three-model factor analysis. *British Journal of Mathematical and Statistical Psychology, 32*, 87–104.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Cascio, W. F., & Aguinis, H. (2005). Test development and use: New twists on old questions. *Human Resource Management, 44*, 219–235.

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4*, 236–264.

Davis, S. L., & Finney, S. J. (2006). A factor-analytic study of the cross-cultural adaptability inventory. *Educational and Psychological Measurement, 66*, 318–330.

Goerzen, A., & Beamish, P. W. (2003). Geographic scope and multinational enterprise performance. *Strategic Management Journal, 24*, 1289–1306.

Hessen, D. J., Dolan, C. V., & Wicherts, J. M. (2006). The multigroup common factor model with minimal uniqueness constraints and the power to detect uniform bias. *Applied Psychological Measurement, 30*, 233–246.

Hu, L. T., & Bentler, P. (1995). Evaluating model fit. In: R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). London: Sage.

Joreskog, K. G., & Sorbom, D. (1981). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least squares method*. Chicago, IL: National Educational Resources.

Kim, K. H., Cramond, B., & Bandalos, D. L. (2006). The latent structure and measurement invariance of scores on the Torrance Tests of Creative Thinking – Figural. *Educational and Psychological Measurement, 66*, 459–477.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320–341.

McGaw, B., & Joreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socio-economic status. *British Journal of Mathematical and Statistical Psychology, 24*, 154–168.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64*, 916–924.

Podsakoff, P. M., MacKenzie, S. B., Bachrach, D. G., & Podsakoff, N. P. (2005). The influence of management journals in the 1980s and 1990s. *Strategic Management Journal, 26*, 473–488.

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The measurement of work engagement with a short questionnaire: A cross-national study. *Educational and Psychological Measurement, 66*, 701–716.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956–970.