Qual Life Res (2012) 21:1623–1624
DOI 10.1007/s11136-012-0112-z

RESPONSE

# "It ain't over till the fat lady sings": a response to Cameron N. McIntosh, improving the evaluation of model fit in confirmatory factor analysis

**Peter M. Fayers · Neil K. Aaronson**

Accepted: 8 January 2012 / Published online: 19 January 2012


We are grateful for the opportunity to respond to Cameron McIntosh, and we do so on behalf of all of the authors. Chad Gundy, the lead author on our manuscript, "Comparing higher order models for the EORTC QLQ-C30," died this past August after a sustained battle with cancer. Chad continued to work on the paper throughout the period of his illness, responding to reviewers' comments and revising the manuscript to its near final form. Chad thoroughly enjoyed the challenge of the relatively complex work reported in our paper, as well as the on-going debate that surrounds the methods that we used. He would have relished the opportunity to respond McIntosh's commentary. We hope that we have captured both the spirit and the intent of Chad's perspective in this response.

McIntosh raises a number of very interesting and pertinent points, most of which we agree with. At the same time, as we argued in our exchanges with the reviewers and we would still argue here, there is a great deal of controversy surrounding the proper use of the chi-square statistic and AGFI. As we did not foresee that consensus on this matter would be achieved any time soon, we decided to report both parameters in our paper. Interestingly, the controversy about the appropriate goodness-of-fit measures was reflected clearly in the diverging viewpoints expressed by the reviewers of our manuscript. Whereas McIntosh argued in favour of using chi-square as an appropriate indicator of model fit, regardless of sample size, another reviewer wrote: '"I'm glad you report the df and Chi-square in Table 2, but please stop talking about it as a measure of fit. It is useless as such with the N that you have." In our multidisciplinary group, from the very start, we had had some particularly heated debates between statisticians and psychometricians. Statisticians, in particular, have long recognized that the chi-square test is fundamentally different from many other statistical significance tests (e.g., Berkson [1]).

Statisticians invariably start by defining the null hypothesis: It is all but impossible to explain or discuss statistical significance tests without reference to the concept of the null hypothesis. Many non-statisticians fail to appreciate the need to take such a basic approach. A statistical significance test aims to estimate the probability that such extreme data as have been observed could have arisen purely by chance, if the null hypothesis is true. In the case of the chi-square test, the null hypothesis is that the specified model will fully explain the patterns in the observed data. The problem is that in many situations, including the present one, it is futile to expect that any of the relatively simple conceptual structural models that we and others have proposed will provide a complete representation of complex psychological and biological mechanisms. We can only hope to obtain approximate fit to the data, and we know, in advance, that any claim for perfect fit is implausible. This is quite different than the null hypothesis of many other statistical significance tests, such as in a clinical trial comparing two or more treatments (where the null hypothesis is commonly "no difference" in treatment effect), or in a regression model in which we might test whether a coefficient differs from the null hypothesis of zero. In the case of

P. M. Fayers
Section of Population Health, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK

P. M. Fayers
Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

N. K. Aaronson (✉)
Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
e-mail: n.aaronson@nki.nl

the chi-square test, there is no coefficient or effect to estimate and test; we simply test the rather absurd hypothesis of perfect and exact fit [2].

Does it spell disaster if we need to acknowledge that all of our (relatively simple) structural models will violate the null hypothesis of perfect fit? Not at all. A model may fit well enough for practical and clinical purposes, and it may then suffice to provide a useful, albeit simplified, conceptual model for the principal structural relationships.

So why does the use of chi-square present a problem? First, as argued above, it is pointless to test a null hypothesis that we know to be false. In such a situation, lack of evidence of misfit as indicated by a non-significant *p*-value simply means that the sample size was inadequate. As Nunnally commented, these hypotheses, called "point null hypotheses," are almost invariably known to be false before any data are collected; if such hypotheses are not rejected, it is because the sample size is too small [3]. By increasing the sample size, we can increase the chi-square statistic and make the *p*-value as highly significant as we wish. The magnitudes of chi-square and the *p*-value are thus completely uninformative. As Berkson summarized in 1938, what is the point of applying a chi-square test to a moderate or small sample if we already know that a large sample would show *p* highly significant? [1].

The second problem, as Berkson also noted, is that the name "goodness-of-fit" is misleading, again because the power of the chi-square test to detect an underlying disagreement between theory and data is controlled largely by the size of the sample. A model may show statistically significant evidence of misfit, yet still be a useful and practical simplification of reality. The term goodness-of-fit implies a measure of adequacy of fit. If a model provides good (or poor) fit, the same measured level of fit should be found irrespective of the size of the sample. Thus, the relationship of chi-square to sample size means that it does not satisfy the basic requirement for a goodness-of-fit index. Instead, a number of other indexes have been proposed that are less sensitive to sample size. Unfortunately, as McIntosh rightly observes, the performance of these indexes is also being called increasingly into question. However, the absence of consensus on alternative indexes does not make chi-square any the more acceptable.

McIntosh queries why we are more willing to rely on the chi-square statistic when comparing two models. Here, the null hypothesis is that the models fit equally well. As McCullagh and Nelder write, "Data will often point with almost equal emphasis on several possible models, and it is important that the statistician recognize and accept this" [4]. In other words, the null hypothesis of no difference is no longer implausible and is now one worthy of testing. In SEM, it is sometimes impossible to discriminate between two or more models. As before, however, chi-square does not inform whether there is necessarily a large enough difference between the two models to be of practical or clinical importance. It merely informs us whether there is any evidence that the data support one model as providing better fit than the other.

In summary, we would argue that use of chi-square is not valid for evaluating goodness-of-fit. We agree with McIntosh that there are a number pitfalls surrounding model fit assessment and that there is a need for clearer guidelines when using confirmatory factor analysis/structural equation modelling. In the meantime, the debate continues, or as Dan Cook, the sports journalist, originally put it: "It ain't over till the fat lady sings."



## References

1. Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the Chi-square test. *Journal of the American Statistical Association, 33*, 526–536.
2. Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association, 82*, 112–122.
3. Nunnally, J. C. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20*, 641–650.
4. McCullagh, P., & Nelder, J. (1990). *Generalized linear models*. London: Chapman and Hall.