Communication Research Reports

Routledge
Taylor & Francis Group

ISSN: 0882-4096 (Print) 1746-4099 (Online) Journal homepage: http://www.tandfonline.com/loi/rcrr20

Statistical and Practical Concerns With Published Communication Research Featuring Structural Equation Modeling

Alan K. Goodboy & Rex B. Kline

To cite this article: Alan K. Goodboy & Rex B. Kline (2017) Statistical and Practical Concerns With Published Communication Research Featuring Structural Equation Modeling, Communication Research Reports, 34:1, 68-77, DOI: 10.1080/08824096.2016.1214121

To link to this article: http://dx.doi.org/10.1080/08824096.2016.1214121

Published online: 16 Dec 2016.

Communication Research Reports
Vol. 34, No. 1, 2017, pp. 68–77

Routledge
Taylor & Francis Group

ANALYSIS SPOTLIGHT

Statistical and Practical Concerns With Published Communication Research Featuring Structural Equation Modeling

Alan K. Goodboy & Rex B. Kline

Structural equation modeling (SEM) is becoming an increasingly popular data analytic technique in communication studies. Reports of SEM analyses are published in communication journals (including Communication Research Reports) allowing for hypothesis testing with latent variables, estimation of direct and indirect causal effects, and validity testing for measurement instruments. Too often, though, serious mistakes are made by authors of SEM studies that cancel out the potential benefits of SEM. Highlighted in this work are five of the most common mistakes made by communication researchers in analyzing and reporting about structural equation models. These problems concern descriptions of model specification, model identification, and methods to evaluate the degree of model-data correspondence, or fit, and lack of replication that continue to plague the empirical SEM literature. This current work is intended as a primer, one that outlines best practices in contrast to widespread but poor practices in each of the areas just mentioned. The hope is that communication researchers yield benefits from the correct application of SEM while avoiding common pitfalls.

Keywords: *Confirmatory Factor Analysis; Path Analysis; Structural Equation Modeling; Structural Regression Model*

Alan K. Goodboy (PhD, West Virginia University, 2007) is an associate professor in the Department of Communication Studies at West Virginia University. Rex B. Kline (PhD, Wayne State University, 1985) is a professor of Psychology at Concordia University, Montreal. The authors would like to thank R. Lance Holbert for his input and guidance with this manuscript. *Correspondence*: Alan K. Goodboy, West Virginia University, 108 Armstrong Hall, P.O. Box 6293, Morgantown, WV 26506; E-mail: agoodboy@mail.wvu.edu

ISSN 0882-4096 (print)/ISSN 1746-4099 (online)
DOI: 10.1080/08824096.2016.1214121

Quantitative communication researchers increasingly use structural equation modeling (SEM), which consists of a family of techniques, such as path analysis and confirmatory factor analysis (CFA), that estimate causal relations between variables of substantive interest, given variances, covariances, or means from observed variables. There are now several commercial or widely available free SEM computer tools that make it convenient to analyze structural equation models (Byrne, 2012) and require little or no knowledge of syntax in order to specify the data, model, and output, making it much easier to conduct complex multivariate analyses. Communication researchers have welcomed the statistical advantages of SEM, including the capability to test hypotheses about latent variables, estimate direct or indirect causal effects, analyze measurement models in validation studies, and compare alternative models that reflect competing sets of hypotheses (Holbert & Stephenson, 2008), among other possibilities.

The developments and features just described have encouraged communication scholars to apply SEM. Unfortunately, the enthusiasm with which SEM has been applied has generally outstripped better judgment because many published SEM studies have flaws so severe that there can be no meaningful interpretation of the results (Holbert & Stephenson, 2002, 2008; MacCallum & Austin, 2000). These problems persist even though there are journal article reporting standards for SEM (Hoyle & Isherwood, 2013) and numerous works about best practices in SEM studies (e.g., Kline, 2016, ch. 18; Schumacker & Lomax, 2016, ch. 16). Thus, it comes as no surprise that it is easy to find SEM studies with serious defects in communication journals. Our intention in this work is not to "name names" by calling out individual communication scholars who have published flawed SEM studies—readers can easily find examples on their own—so this work is not a kind of "witch hunt." Our purpose is instead more constructive; it is to inform communication researchers about some misuses of SEM in order to discourage them from making similar mistakes in the future. A related hope is that this tutorial promotes more informed use of SEM that will benefit reviewers and readers of communication scholarship.

There are exemplary reports of empirical studies in communication journals that meet standards for good science (Feeley, 2015), but other reports feature dubious methods of data analysis. For instance, Vermeulen and Hartmann (2015) described various poor practices used by some communication researchers that "optimize" hypothesis testing in order to find statistically significant results. Two examples include HARKing (hypothesizing after the results are known) and data peeking (collecting data until statistical significance is attained). Similar kinds of practices are found in SEM studies. As Seaman and Weber (2015) noted, "despite its widespread use, there are persistent misapplications of SEM in published empirical studies in the communication sciences [including] the exploitation of flexibilities in data collection and analysis" (p. 208). These authors described major problems of SEM in communication studies, including (a) ignoring distributional assumptions, such as multivariate normality, of certain estimation methods for continuous outcome variables; (b) testing large models in small samples; (c) failing to compare alternative models; and (d) using inadequate methods for evaluating model fit.

Summarized next in this brief report are what we believe are among the five most frequent misuses of SEM in communication studies. Some of these problems have been flagged to this point, and these and other deficiencies are elaborated next. All of these shortcomings can be avoided with better analysis and reporting techniques. Doing so would also improve the scientific merit of SEM applications in communication studies.

Issue 1: Just-Identified Models, Models With Few Degrees of Freedom, and Perfect Fit

This happens in published SEM research: A communication researcher reports that her or his model has what is described as "exceptional" global fit, but further inspection indicates that the model is just-identified, or saturated, which means that it has zero degrees of freedom (df = 0) and thus has as many freely estimated parameters as observations (variances, covariances, or means) in the data set. Such models are as complex as the data they are supposed to explain and thus must perfectly fit that data, and this is true even for just-identified models that are grossly incorrect. Any respecification of a just-identified model—including ones with contradictory causal hypotheses—will also perfectly fit the same data. Just-identified models test no particular hypothesis because they have no testable implications that can be refuted by the data. Such models thus have little, if any, scientific value.

In order for a model to have less than perfect correspondence with the data, the model must have positive degrees of freedom, or $df > 0$. Models with this property are called overidentified. Such models have testable implications that allow for the possibility that model-data disagreement is so great that the model could be rejected. Thus, researchers should analyze models where $df > 0$. Now, it can happen that models with very few degrees of freedom, such as $df = 1$, can have near-perfect fit—e.g., $\chi^2(1) = .52$, $p = .48$. Such models have so many free parameters relative to the number of observations that they can hardly fail to explain the data to a very close extent. This explains the preference in SEM for models with greater degrees of freedom. Retaining such models is more impressive because they have withstood more opportunities in the form of greater df to disagree with the data.

Some communication researchers place too much emphasis on obtaining excellent fit by including too many paths in a model, thereby analyzing just-identified models (df = 0) or slightly overidentified models (e.g., df = 1 or 2). Doing so favors retention of the model even though it could be very misspecified. As noted by Holbert and Stephenson (2008),

> … claiming support for theory based on good fit statistics for a model that approaches saturation is not very credible, but unfortunately, is an all-too-common practice among some users of SEM. Communication scholars want to hypothesize, test, and interpret overidentified models. In addition, the more degrees of freedom, the better. (p. 192)

There is no magic minimum value for model degrees of freedom, but more is better than less. Holbert and Stephenson (2002) reported an average of 33.38 degrees of freedom in their review of SEM studies published in communication journals. This is an ample amount, but it is just an average, and models in many studies have far fewer degrees of freedom. Unfortunately, Holbert and Stephenson (2002) also found that the value of *df* was incorrectly calculated in about 10% of reviewed studies, which is a serious flaw. Even worse is when the value of *df* is not stated in the article. Researchers should always report for their analysis the number of observations, give a tally of the number of freely estimated parameters, and state *df* as the positive difference between these two quantities. If perfect or near-perfect model fit is reported, this outcome should be a red flag, and the value of *df* should be considered.

Issue 2: Overly Complex Models

Ever-increasingly complex models are being published in communication journals. Some of these models have 20–30 or more paths represented with arrows that connect variables in ways that resemble the schematic for a Rube Goldberg machine, or a deliberately overengineered contrivance that performs a simple task in a very complicated way. Arrows in such models fly right, up, or down, connecting many pairs of variables sometimes with little theoretical justification offered in the text. Paths in some of these overly complicated models resemble *spider webs* that connect most exogenous variables to basically all endogenous variables. Such models may also feature a *cascading waterfall* of arrows from one variable to another and another in a long sequence of alleged indirect causation.

The theoretical rationale for structural equation model should *always* be stated. This is especially true for directionality specifications. For example, the hypothesis that X causes Y ($X \rightarrow Y$) instead of the reverse ($Y \rightarrow X$) must be explained, especially in cross-sectional designs where all variables are concurrently measured. Such designs feature no time precedence, which means that putative causes are measured before their presumed outcomes, such as in longitudinal designs. Sometimes the nature of variables rule out certain directionality specifications even in cross-sectional designs. For example, it would make no sense to specify that a demographic variable, such as age, is caused by a psychological variable, such as self-esteem. But most other times it may not be obvious in a cross-sectional design how to distinguish between two psychological variables, such as anxiety and depression, as cause versus outcome without an ironclad rationale (Tate, 2015). Nothing other than argument in cross-sectional designs supports directionality specifications, and these arguments should be clearly stated.

Overly complicated models take advantage of the fact that if a researcher draws enough paths to estimate, the model will eventually fit, perhaps not due to its correctness but due to the paucity of degrees of freedom—see Issue 1. Another reason for concern is that highly complex models are not typically based on theory, so they are unlikely to replicate. After all, it is highly unlikely that any body of theory in communication is so specific that it would directly translate to 30+ paths in a spider web pattern or in a cascading waterfall pattern. As Floyd (2014) noted,

> … our statistical methods grow in complexity, often to the detriment of our message. To me, the proliferation of increasingly complex statistical methods is the most troubling of these trends. I often see reviewers encouraging authors to use more complicated analyses than they need. Many of our papers are now filled with convoluted structural equation models that are simply unnecessary to test the claims of the study. Complexity, it seems, has become an end unto itself. (p. 3)

Box (1976) put it like this: Overelaboration and overparameterization—that is, unnecessary complexity—is often the hallmark of scientific mediocrity. Box (1976) and Floyd (2014) are both correct—too many convoluted models are analyzed in communication journals. Researchers should start with basic models and stick to parsimonious model building from a core set of theoretically guided predictions. Complex model building without theoretical justification leads to the issue considered next.

Issue 3: Data-Driven (Empirical) Respecification

The overrated and much-hyped *p* value generated in statistical significance testing is glorified by too many communication researchers, who overemphasize *p* values despite the growing controversy in the behavioral and other sciences about significance testing (see Kline, 2013; Levine, Weber, Hullett, Park, & Lindsey, 2008). Significance testing is banned in at least one psychology journal, *Basic and Applied Social Psychology* (Trafimow & Marks, 2015), and the American Statistical Association (2016) recently issued a warning about the limitations of *p* values in significance testing.

Overreliance on significance testing is also prevalent in SEM. For example, the conventional criterion value of statistical significance, $\alpha = .05$, is specified by most researchers in SEM. If a model is found to be consistent with the data—perhaps due to too few degrees of freedom in an overly complex model—see Issues 1 and 2—next the model is often "trimmed" by deleting paths with coefficients where $p > .05$. Nonsignificant paths may be eliminated one at a time or in sets of multiple paths, but the final model is one where all paths are statistically significant. For example, when testing measurement models in the technique of CFA, some researchers delete indicators from factors with pattern coefficients (loadings) that are not significant. The final model features pattern coefficients that are all statistically significant.

Deleting nonsignificant paths from a structural equation model is a *terrible* way to trim the model (Kline, 2016). This is because doing so strongly capitalizes on sample-specific variation (chance). This means that the coefficient might be statistically significant in a replication sample. If the power of the significance test is low, which can happen due to sample sizes that are too small to detect an effect of a particular magnitude in the population, then a theoretically meaningful path may be deleted with no good reason. Sample sizes in SEM are generally too small for adequate power in significance testing (e.g., Westland, 2010). If the sample size is very large, though, then paths with coefficients that are practically zero can be significant; that is, effects of trivial magnitude can be flagged as statistically significant in very large samples.

Loehlin (2004) gives this good advice: There is no reason to ritualistically drop every path that is nonsignificant. Doing so can appreciably affect the solution in unanticipated ways. If a path was theoretically justified in the first place, then it should be retained regardless of whether its coefficient is significant or not significant. Thus, outcomes of significance testing are irrelevant for such paths.

It is just as bad to add paths to an extant model with unacceptable fit to the data based solely on *p* values. A common example occurs when the value of a modification index (MI), which estimates the amount by which the model χ^2 would drop if the path that corresponds to a particular effect were added to the model (i.e., a previously fixed-to-zero parameter is respecified as a free parameter), is statistically significant. A modification index is interpreted as a χ^2 (1) statistic, and if significant at, say, the .05 level, then the overall model χ^2 will decrease by an amount that may also be significant at the .05 level, given no other change to the model other than adding the path associated with a particular MI statistic. The temptation is to add the path to the model with the greatest MI value that is also significant, but such a respecification may have little, if any, theoretical rationale (e.g., adding many error correlations because of MI values). This is because MI statistics exploit sample-specific variation that may not be found in replication samples; that is, the MI for the same path but estimated in a different sample may not be significant.

Either dropping paths solely because their coefficients are not significant or adding paths solely because their MI is significant amounts to what is called a blind specification search. Such respecifications are likely to reflect mere sampling error instead of the truth or the population model. Indeed, blind model respecification is likely to lead the researcher away from the true model, not toward it. There is a similar phenomenon in regression analysis: Empirical selection of predictors, such as in the stepwise method, generates results that will not typically replicate due to extreme capitalization on chance. Stepwise and related empirical methods for selecting predictors have been banned in some journals, and with good reason, too (e.g., Whittingham, Stephens, Bradbury, & Freckleton, 2006). There should be no place in SEM for analogous practices. Thus, researchers are urged to begin SEM analyses by specifying the simplest model that is supported by theory and then add or drop paths based on a rationale consistent with that theory. Results of significance testing should play little role except for effects predicted in advance to be zero or not zero; otherwise, significance testing is merely a potential distraction in SEM, not a tool for discovery.

Issue 4: Evaluation of Model Fit

Poor practice about model fit evaluation is rampant in SEM studies, so the issues considered next are important. Model fit can and should be evaluated from two different perspectives, global and local. Global fit concerns the overall or average fit of the whole model to the data; specifically, it concerns whether the model can predict the sample variances, covariances, or means (i.e., the observations). The model χ^2 with its degrees of freedom and *p* value are reported in nearly all SEM studies. This statistic tests the null hypothesis of exact or perfect fit, which would be rejected if the p value is less than the level of α specified by the researcher. Other global fit statistics, called approximate fit indexes, are intended to measure the degree of model-data correspondence along a continuum (i.e., they are not significance tests). Examples of commonly reported approximate fit indexes include the Steiger-Lind root mean square error of approximation (RMSEA) with its 90% confidence interval, the Bentler Comparative Fit Index (CFI), and the standardized root mean square residual (SRMR). There are dozens of other approximate fit indexes, which presents the potential problem that different researchers report the values of different subsets of such statistics. There is also the possibility for "cherry-picking," or the reporting of values for only those fit statistics that favor the model. The minimal set of global fit statistics just described (χ^2, RMSEA with 90% CI, CFI, SRMR), though, should be sufficient (Kline, 2016).

Brief mention is warranted of the so-called normed chi-square (NC), which is calculated as χ^2/df. For models where $df > 1$, NC < χ^2. Some researchers report the NC statistic in an attempt to reduce the sensitivity of the model χ^2 to sample size (N). One problem is that the value of df has nothing to do with N, and another is that there never was any clear-cut rule about maximum values of the NC that are "acceptable" (e.g., < 2.0?–< 3.0?). Because there is little statistical or logical foundation for NC, it should have no role in global fit assessment (i.e., do not bother to report it; Kline, 2016).

All global fit statistics share the limitation that they measure overall or average fit. But it can and does happen that values of global fit statistics seem to favor the model, but the model clearly fails to explain the association between certain pairs of observed variables (Tomarken & Waller, 2003). In this way, values of global fit statistics can hide appreciable poor fit at the variable level. This is where local fit statistics are invaluable. Residuals in SEM analyses are local fit statistics, and a residual is calculated as the difference between the observed (sample) association and that predicted by the model for every pair of observed measures. Residuals are printed in computer outputs in four metrics: correlation, covariance, standardized, or normalized residuals. Correlation residuals for pairs of continuous variables are the easiest to interpret, and absolute correlations residuals that exceed .10 may indicate poor prediction for that variable pair (Kline, 2016). The real details about model fit can be found by inspecting the residuals. Unfortunately, too many researchers report values of global fit statistics only and say nothing about the residuals. Such reports are incomplete and fail to reassure readers that the decision to retain a model is actually justified at the level of local fit, not just global fit.

Best practice is to either report a matrix of residuals or, if doing so is impractical in a published report (e.g., the matrix is very large), then at least describe for readers the pattern of the residuals. For example, if a standardized residual is statistically significant for a pair of continuous variables, then also inspect the corresponding correlation residual. If the value of the absolute correlation residual is < .10, then the degree of model-data disagreement for that pair of variables may not be serious; otherwise, specification error is indicated. There is no "magic number" of absolute correlation residuals > .10 that would invalidate a model, but the more of such results, the greater the possibility for specification error. Normalized residuals provide a more conservative significance test than standardized residuals, and sometimes the computer may be unable to calculate a standardized residual for a particular pair of variables, especially in a complex latent variable model. In any event, correlation residuals are basically effect size statistics that measure the magnitude of model-data discrepancy in a standard metric and thus are invaluable in local fit assessment.

Issue 5: Replication

The last of (what we see as) the "big five" problems in many SEM studies is less of a statistical issue than one of research strategy. In theory, SEM is a set of confirmatory techniques, but many communication researchers use SEM for exploratory analyses without replication. As Ullman and Bentler (2009) warned,

> … SEM is a confirmatory technique, therefore when model modification is done to improve fit that analysis changes from confirmatory to exploratory. Any conclusions drawn from a model that has undergone substantial modification should be viewed extremely cautiously. Cross-validation should be performed on modified models whenever possible. (p. 448)

Given the aforementioned issues in this article, it is of the utmost importance that communication researchers *replicate* their findings beyond one-shot studies. It is very rare to see the analysis of a path model, CFA model, or any other type of structural equation model replicated in a second study with an independent sample. This means that most of the large number of SEM studies in the literature are based on one-shot analyses with no evidence of replication. It is especially important to replicate model analyses given the issues previously mentioned (Issue 2, overly complex models are unlikely to replicate; Issue 3, model respecification driven mainly by p values is problematic). Accordingly, the replication of SEM analyses allows for more confidence in the testing of true models, especially models that are relatively complex or have required substantial modification. Likewise, replicating a well-fitting model is good science and is also important for theory, scale construction, and future model testing. The great Achilles heel of SEM in communication and other disciplines is the lack of systematic replication. In this sense, the "replication crisis" has been underway in the SEM literature for some time. Routine replication of SEM analyses would be one of the single greatest reforms.

Conclusion

A total of five problems with SEM analyses by communication researchers were highlighted in this article. There are many other issues we could discuss, but we see these five as among the most common and serious flaws in published SEM communication research. Following are some solutions that researchers should take into account and reviewers should consider. Communication researchers should:

- Specify their models with positive and ample degrees of freedom according to a clear theoretical rationale.
- Abandon the practice of deleting paths that are not statistically significant or adding paths with significant modification indexes as the main basis for respecification.
- Report a standard core set of global fit statistics.
- Specify initial models that are as parsimonious as possible while still respecting the major predictions of theory; that is, prioritize the hypotheses so that only the most important are represented in the initial model.
- Pay close attention to local fit testing by examining the residuals. Describe the residuals in written reports of SEM analyses.
- Take replication of SEM analyses seriously and as a necessary next step in the analysis.

Anyone who has taken a course in SEM or who has a basic working knowledge of it will not find these suggestions to be radical or even novel. But authors and reviewers should know that SEM studies are too often based on poor practices or mistaken emphasis on methods of data analysis that are unhelpful in model testing, such as statistical significance testing. The family of SEM techniques is powerful and flexible, but without wisdom in its use, the potential of these techniques cannot be realized. We hope to see communication researchers publish fewer examples of the all-too-common but completely avoidable problems just reviewed.

References

American Statistical Association. (2016). *Statement on statistical significance and p-values* [Press release]. Retrieved from https://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*, 791–799. doi:10.1080/01621459.1976.10480949

Byrne, B. M. (2012). Choosing structural equation modeling computer software: Snapshots of LISREL, EQS, Amos, and Mplus. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 307–324). New York, NY: Guilford.

Feeley, T. H. (2015). *Research from the inside-out: Lessons from exemplary studies in communication*. New York, NY: Routledge.

Floyd, K. (2014). Taking stock of research practices: A call for self-reflection. *Communication Monographs*, *81*, 1–3. doi:10.1080/03637751.2014.892670

Holbert, R. L., & Stephenson, M. T. (2002). Structural equation modeling in the communication sciences, 1995–2000. *Human Communication Research*, *28*, 531–551. doi:10.1111/j.1468-2958.2002.tb00822.x

Holbert, R. L., & Stephenson, M. T. (2008). Commentary on the uses and misuses of structural equation modeling in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The Sage sourcebook of advanced data analysis methods for communication research* (pp. 185–218). Thousand Oaks, CA: Sage.

Hoyle, R. H., & Isherwood, J. C. (2013). Reporting results from structural equation modeling analyses in *Archives of Scientific Psychology*. *Archives of Scientific Psychology*, *1*, 14–22. doi:10.1037/arc0000004

Kline, R. B. (2013). *Beyond significance testing: Statistical reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, *34*, 171–187. doi:10.1111/j.1468-2958.2008.00317.x

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226. doi:10.1146/annurev.psych.51.1.201

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York, NY: Routledge.

Seaman, C. S., & Weber, R. (2015). Undisclosed flexibility in computing and reporting structural equation models in communication science. *Communication Methods and Measures*, *9*, 208–232. doi:10.1080/19312458.2015.1096329

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology*, *37*, 235–246. doi:10.1080/01973533.2015.1062380

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well-fitting" models. *Journal of Abnormal Psychology*, *112*, 578–598. doi:10.1037/0021-843X.112.4.578

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2. doi:10.1080/01973533.2015.1012991

Ullman, J. B., & Bentler, P. M. (2009). Structural equation modeling. In M. Hardy & A. Bryman (Eds.), *Handbook of data analysis* (pp. 431–458). Thousand Oaks, CA: Sage.

Vermeulen, I., & Hartmann, T. (2015). Questionable research and publication practices in communication science. *Communication Methods and Measures*, *9*, 189–192. doi:10.1080/19312458.2015.1096331

Westland, C. J. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, *9*, 476–487. doi:10.1016/j.elerap.2010.07.003

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*, 1182–1189. doi:10.1111/j.1365-2656.2006.01141.x